Australian National University

Crawford School of Public Policy

CAMA

Centre for Applied Macroeconomic Analysis

An Automated Prior Robustness Analysis in Bayesian Model Comparison

CAMA Working Paper 45/2019
June 2019

Joshua C. C. Chan
Purdue University
Centre for Applied Macroeconomic Analysis, ANU

Liana Jacobi
University of Melbourne

Dan Zhu
Monash University

Abstract

The marginal likelihood is the gold standard for Bayesian model comparison although it is well-known that the value of marginal likelihood could be sensitive to the choice of prior hyperparameters. 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. In this paper we develop efficient and feasible methods to compute the sensitivities of 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 setting of Markov chain Monte Carlo simulation. 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 Classification

C11, C53, E37

Address for correspondence:

(E) cama.admin@anu.edu.au

ISSN 2206-0332

An Automated Prior Robustness Analysis in Bayesian Model Comparison

Joshua C.C. Chan
Purdue University and CAMA

Liana Jacobi
University of Melbourne

Dan Zhu[*]
Monash University

April 2019

Abstract

The marginal likelihood is the gold standard for Bayesian model comparison although it is well-known that the value of marginal likelihood could be sensitive to the choice of prior hyperparameters. 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. In this paper we develop efficient and feasible methods to compute the sensitivities of 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 setting of Markov chain Monte Carlo simulation. 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 and intensity of marginal likelihood estimations.

In practice, even when a prior sensitivity analysis is conducted, therefore often only a narrow aspect is investigated. For example, researchers might assess a specific aspect of marginal likelihood sensitivities by recomputing its value using a different set of hyperparameters. However, this approach is ad hoc and requires a substantial amount of computational overhead. Computational problems are especially severe in this context because estimating the marginal likelihood under one set of priors would typically take a substantial amount of time. In this paper we introduce a computationally feasible and systematic approach to assess marginal likelihood sensitivities with respect to a variety of hyperparameters that does not require the re-running of the MCMC chain. In particular, we develop methods based on Automatic Differentiation (AD) to analyze the complete set of prior hyper-parameter sensitivities of the marginal likelihood alongside the model estimation.

In a nutshell, the AD approach provides an efficient way to compute derivatives of an algorithm—i.e., local sensitivity of the outputs with respect to the inputs. 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 sen-

[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).

sitivities. Importantly for our purpose, the AD would only require running the original algorithm once. While AD methods are now commonly used in Financial Mathematics and Machine Learning, the approach is yet to be widely adopted in Econometrics or Statistics. The computational intensity of AD together with the focus of standard AD methods (and packages) on continuous mappings, pose extra challenges for an application to common Bayesian MCMC computations and algorithms.

Jacobi, Joshi, and Zhu (2018) have addressed many of these challenges and develop the first AD-based approach for input sensitivity analysis of Markov chain Monte Carlo (MCMC) output from continuous and discontinuous high-dimensional mappings. Chan, Jacobi, and Zhu (2018) 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 contributes to this line of research by further extending the AD-based approach to the more computationally intensive setting of computing the marginal likelihood using MCMC output and adaptive importance sampling. There is by now a large 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).

A key innovation of this paper is to study the derivative of the cross-entropy parameter with respect to the prior hyper-parameters, such that the cross-entropy parameters are obtained via a numerical search of optimum. In the case where an analytical expression for the gradient and Hessian are provided, we can readily apply 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. Instead of working with the algorithm like the standard AD, we use its implicit derivative and estimates using the simulated samples. The derivation of associated derivative is based on the fact that the cross-entropy parameter is the optimal in minimizing the Kullback–Leibler divergence measure. Hence, the result is easily extendable to a wider range of estimation context, such as Variational Bayes, which allows analysts to assess the impact of the prior assumptions on the parameter estimates.

A further challenge to address is the memory intensity of AD methods. With our goal of applying AD for simulation-based marginal likelihood measures, we emphasize that AD does not, despite its name, fully automate differentiation and can yield inefficient code if naively implemented. The first difficulty of applying AD here lies in the memory constraints. The most natural form of applying AD in most statistical inference is to apply the "Operator Overloading", i.e., a new class of objects which contain the value of a variable and its associated differential component. The memory requirement is then simply twice the requirement for a function evaluation. What is more, an object needs be stored in memory until all the nodes connected to that object have been evaluated. Unlike the original paper of Jacobi, Joshi, and Zhu (2018), the marginal likelihood consists of a post MCMC estimation procedure using intermediary values generated in the original MCMC run. Hence, the dependence of the final estimate on the prior inputs contains components that implicitly through those MCMC intermediary values. A naive storage of all intermediary values and its associated derivatives will quickly exhaust the memory of a standard computer. Hence, we need to formulate the marginal likelihood estimation procedure such that the minimal amount of storage is needed. We specify the exact quantities stored in the application section.

We illustrate our new methodology with two empirical applications in the context of multivariate series analysis using vector autorogressive and factor models. In each case we use AD to compute the gradient of each estimator with respect to a variety of hyperparameters and assess various aspects of the marginal likelihood sensitivity. The first application compares two vector autoregressions (VARs) for modeling a US macroeconomic dataset that involves GDP inflation, real output growth and Federal funds rate. In the second application, we fit daily returns on nine foreign exchange rates using factor models with different number of latent factors. While the conclusion in the first application—that the VAR with t errors are 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 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 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. Lastly, Section 5 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 $p(\mathbf{y} \,|\, \boldsymbol{\psi}_k, M_k)$ and a prior on the model-specific parameter vector $\boldsymbol{\psi}_k$ denoted by $p(\boldsymbol{\psi}_k \,|\, M_k)$. 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

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

where

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

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:

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

where $p(\mathbf{y}_{t+1} \,|\, \mathbf{y}_{1:t}, M_k)$ 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 $p(\mathbf{y}_1 \,|\, M_k)$ depends entirely on the prior distribution and not on the data. More generally, the component $p(\mathbf{y}_{t+1} \,|\, \mathbf{y}_{1:t}, M_k)$ 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. 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 $p(\boldsymbol{\psi};\boldsymbol{\theta}_0)$. Furthermore, from here onwards we suppress the model indicator for clarity. For example, we denote the likelihood function simply by $p(\mathbf{y}\,|\,\boldsymbol{\psi})$.

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

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

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

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

The posterior ordinate $\boldsymbol{\psi}^*$ can in principle be any point in the support of the posterior distribution, 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 analytically. 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 can estimate a model using a 3-block Gibbs sampler, and we have

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

The first quantity $p(\boldsymbol{\psi}_1^*\,|\,\mathbf{y};\boldsymbol{\theta}_0)$ can be estimated using posterior draws and the last quantity $p(\boldsymbol{\psi}_3^* \,|\, \mathbf{y}, \boldsymbol{\psi}_1^*, \boldsymbol{\psi}_2^*)$ can be evaluated exactly. The middle term $p(\boldsymbol{\psi}_2^* \,|\, \mathbf{y}, \boldsymbol{\psi}_1^*; \boldsymbol{\theta}_0)$ can be estimated using draws from a reduced run that cycles through $p(\boldsymbol{\psi}_2 \,|\, \mathbf{y}, \boldsymbol{\psi}_1^*, \boldsymbol{\psi}_3; \boldsymbol{\theta}_0)$ 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

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

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

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

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 $f(\cdot;\widehat{\mathbf{v}}^*_{\mathrm{ce}})$ is obtained, it is used to construct the importance sampling estimator:

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

where $\boldsymbol{\psi}^1,\ldots,\boldsymbol{\psi}^N$ are independent draws from the optimal importance sampling density $f(\boldsymbol{\psi};\widehat{\mathbf{v}}^*_{\mathrm{ce}})$.[3] 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 hyperparameters, $\boldsymbol{\theta}_0 \in \mathbb{R}^p$. This builds on recent work by Jacobi, Joshi, and Zhu (2018) that has introduced prior robustness for MCMC output based on Automatic Differentiation (AD), which is designed to compute sensitivities with respect to the full set of input parameters.

3.1 AD Implementation

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

$$AD : g \rightarrow \frac{\partial}{\partial\theta}g.$$

Like the symbolic differentiation implemented in many widely used softwares, AD computes exact partial derivatives of the original mapping up to floating point errors. Yet, unlike the symbolic differentiation that focuses on obtaining the exact expression of $\frac{\partial}{\partial\theta}g$,

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

AD evaluates the derivatives *alongside* the original evaluation of g, which in turn alleviates the issue of expression overloading and hence typically maintains a relative fast computation. AD is the same as symbolic differentiation once the derivative expression obtained from the latter is simplified to minimize the computational complexity, but the essence is that the simplification is automatic once the evaluation of derivatives is embedded *alongside* the original algorithm. More importantly, AD completely avoids the infeasible derivation of symbolic expressions, and focuses on the actual evaluation of derivative values. This differentiates AD from symbolic differentiation.

Bayesian MCMC algorithms are complicated high-dimensional mappings that take inputs such as hyperparameters of the prior distribution, the starting values of the chain and the data. For many applications, we are typically interested in the effect of a subset of these inputs, say, $\boldsymbol{\theta}_0$, 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 in combination with $\boldsymbol{\theta}_0$ that are mapped via some MCMC algorithm $\mathbf{G}$ into posterior quantities, albeit the analyst is not interested in its relative sensitivities. 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 $\boldsymbol{\theta}_0$.

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

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

where

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

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 compute 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 matrix because each operation $\mathbf{G}_i$ typically only updates one or two variables.

In the context of MCMC, 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. AD accomplishes this by differentiating the evolution of the underlying state variables along each path. In comparison to the widely used numerical finite difference methods, AD requires additional model analysis and programming, but this additional effort is often justified by the improvement in the quality and comprehensiveness of calculated local sensitivities. Due to the computation burden of numerical finite difference methods, typically only a very limited prior robustness 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). In particular, the paper develops an AD approach and AD based methods for a comprehensive prior robustness and convergence 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 in various standard models. Since both Chib's and the cross-entropy methods require posterior draws of the model parameters, we apply the AD approach developed in Jacobi, Joshi, and Zhu (2018) to obtain the first-order partial derivatives of the model parameters with respect to $\boldsymbol{\theta}_0$. What is new here is the 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 have also provided Matlab and R-code to implement the AD-based prior sensitivity analysis. While most AD package act as a black box supporting a statistical inference, we emphasize the extensibility of these package when it is taken to a new application, in our case, computing the marginal likelihood as a post MCMC procedure. It should allow advanced users to incorporate a custom derivative method for a function. In this section 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.

Given 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 first two components can be obtained via:

$$\frac{\partial \log p(\mathbf{y} \,|\, \boldsymbol{\psi}^*)}{\partial \boldsymbol{\theta}_0} = \frac{1}{p(\mathbf{y} \,|\, \boldsymbol{\psi}^*)} \left.\frac{\partial p(\mathbf{y} \,|\, \boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\right|_{\boldsymbol{\psi}=\boldsymbol{\psi}^*} \frac{\partial \boldsymbol{\psi}^*}{\partial \boldsymbol{\theta}_0}$$

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

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

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

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

We can estimate the derivative of the first term via the original MCMC

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

where $\frac{\partial p(\boldsymbol{\psi}_1^* \,|\, \boldsymbol{\psi}_2^r, \boldsymbol{\psi}_3^r, \mathbf{y}; \boldsymbol{\theta}_0)}{\partial \boldsymbol{\psi}}$ denotes the partial derivative of $p(\boldsymbol{\psi}_1 \,|\, \boldsymbol{\psi}_2, \boldsymbol{\psi}_3, \mathbf{y}; \boldsymbol{\theta}_0)$ 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 $\boldsymbol{\psi}_1^*$

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

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.

3.3 Gradient of the Cross-Entropy Estimator

To calculate the gradient of the cross-entropy marginal likelihood estimator, we need to first obtain $\frac{\partial \mathbf{v}_{\mathrm{ce}}^*}{\partial \boldsymbol{\theta}_0}$. For cases where analytical expressions of $\mathbf{v}_{\mathrm{ce}}^*$ is available, e.g., for 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. $\mathbf{v}_{\mathrm{ce}}^*$ 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 $\mathbf{v}_{\mathrm{ce}}^*$ requires numerical search such the Newton-Raphson method, we can compute the derivative via the implicit function theorem.

Proposition 1. *Assuming that the importance sampling density $f(\boldsymbol{\psi}; \mathbf{v})$ is twice contin-*

uously differentiable in both $\mathbf{v}$ *and* $\boldsymbol{\psi}$ *with*

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

and

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

is positive definite, then

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

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 side

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

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

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

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 in a way 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 $f(\boldsymbol{\psi}; \widehat{\mathbf{v}}^*_{\mathrm{ce}})$, the derivative of the CE estimator is given by:

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

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

4 Empirical Applications

In this section we present 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 number 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.

In the first application we perform a formal Bayesian model comparison exercise to compare two popular VARs for fitting a US macroeconomic dataset. 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, real GDP and Federal funds rate from 1954:Q3 to 2017:Q4. These three variables are commonly used in structural analysis and forecasting (e.g., Banbura, Giannone, and Reichlin, 2010; Koop, 2013). Both GDP deflator and real GDP series are transformed to annualized growth rates, whereas the Federal funds rate is not transformed. All data are sourced from the Federal Reserve Bank of St. Louis economic database, and they are plotted in Figure 1.

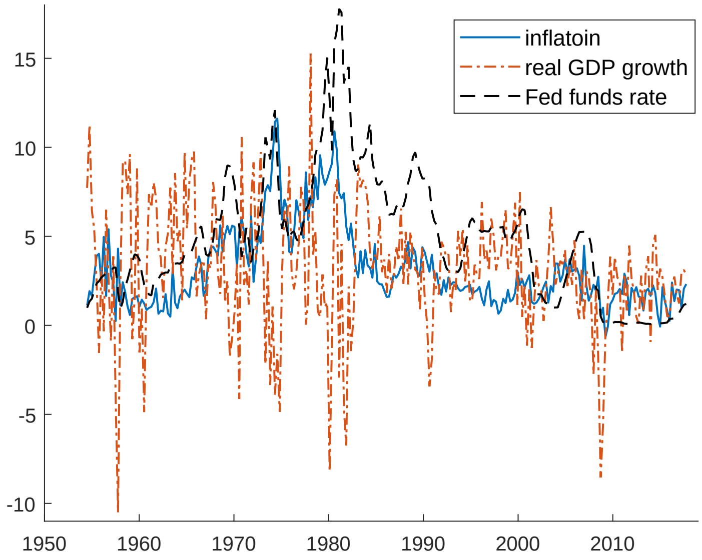

Figure 1: Plots of GDP deflator growth, real GDP growth and Federal funds rate.

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. We set $\kappa_1 = 0.2^2$ 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 $\kappa_3 = 1$. These hyperparameters are fairly standard 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 are fairly standard. Estimation of the Gaussian

VAR can be found in standard textbooks such as Koop and Korobilis (2010) or Chan (2019). Estimation of a regression with t innovations can be found in Koop, Poirier, and Tobias (2007) or Chan (2019).

4.1.2 Empirical Results

We fit the US quarterly dataset using VARs with Gaussian and t errors. For each VAR, we compute the marginal likelihood value using both Chib's method and the cross-entropy method. The results are reported in Table 4.1.2.

Both Chib's and the cross-entropy methods give essentially 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 work 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, 2018). In addition, the t VAR with the heaviest tail ($\nu = 5$) receives the most support.

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	−1416.7	−1322.2	−1344.7	−1381.5
Chib	−1416.7	−1322.2	−1344.7	−1381.5

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 $\mathbf{\Sigma}$.

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	424.3	−0.01	10.3	471.7	−0.01	5.6
Chib	424.3	−0.01	10.3	471.8	−0.01	5.6

Our results show that the marginal likelihood estimates are relatively sensitive to κ_1 and κ_3, but not to κ_2. For example, increasing κ_1 from the baseline value of 0.04 to 0.05 would increase the log marginal likelihood value of the Gaussian VAR by about 4.2,[4] but increasing κ_2 by the same proportion—from the baseline value of 100 to 125—has little impact on the marginal likelihood value.

Interestingly, even though the three hyperparameters are common across the two VARs, their impacts on the marginal likelihood values differs across the two VARs. For example, increasing κ_1, i.e., decreasing the strength of shrinkage, helps the t VAR fit the data better relative to the Gaussian 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 -1425 and -1332, respectively. Since the difference of the two values remains large, the conclusion that the data strongly prefer the t VAR is reasonably robust.

To assess how the marginal likelihood estimates vary over a wider range of hyperparameter values, one can plot the estimates together with the corresponding derivatives against some hyperparameters of interest. For example, Figure 2 plots these values against κ_1.

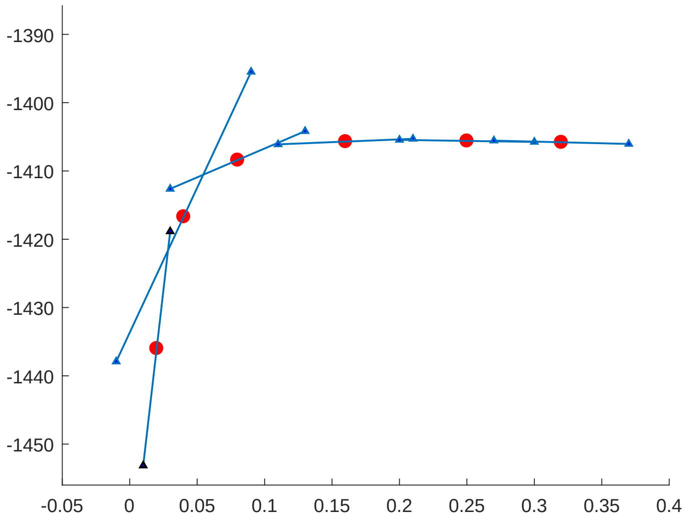

Figure 2: Marginal likelihood estimates (red dots) and the corresponding derivatives (blue tangents) of the Gaussian VAR against κ_1.

[4] To check this estimate, we redid the marginal likelihood estimation with $\kappa_1 = 0.05$, while keeping other hyperparameters exactly the same. The new marginal likelihood value increases by 3.6, which is similar to the original estimate.

As the figure shows, for values of κ_1 less than, say, 0.1, 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 0.15, the derivative values are small in magnitude and even negative, suggesting that the maximizer is less than 0.15. Overall, the marginal likelihood values are all less than -1400, confirming that the data favor the VAR with t errors.

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:

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

where $\mathbf{X}_t$ is an $n \times m$ matrix of regressors, $\boldsymbol{\beta}$ is the associated $m \times 1$ vector of coefficients and $\mathbf{A}$ 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 unity (see, for example, the discussion in Geweke and Zhou, 1996).

In our empirical work below we would only include intercepts but no other regressors. Hence, $\mathbf{X}_t = \mathbf{I}_n$ and $\boldsymbol{\beta}$ is an $n \times 1$ column of intercepts.

4.2.1 Data, Priors and Estimation

In the second application we analyze daily returns on nine international currency exchange rates relative to US dollar beginning in January 2007 and ending in December 2010. 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. The nine currencies are the

Australian Dollar (AUD), Canadian Dollar (CAD), Euro (EUR), Japanese Yen (JPY), Swiss Franc (CHF), British Pound (GBP), 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.

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 $\boldsymbol{\beta}, \mathbf{a}, \boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, and $\boldsymbol{\Omega} = \text{diag}(\omega_1^2, \ldots, \omega_k^2)$. We consider the following independent priors:

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

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 $\boldsymbol{\beta}_0 = \mathbf{0}$, $\mathbf{V}_{\boldsymbol{\beta}} = \kappa_4 \mathbf{I}_n$, $\mathbf{a}_0 = \mathbf{0}$, $\mathbf{V}_{\mathbf{a}} = \kappa_5 \mathbf{I}_{k_a}$, $\nu_{\sigma_i^2} = 3$, $S_{\sigma_i^2} = \kappa_6$, $\nu_{\omega_j^2} = 3$ and $S_{\omega_j^2} = \kappa_7$, where $\kappa_4 = \kappa_5 = \kappa_6 = \kappa_7 = 1$.

Estimation of the factor model with fixed number of factors k is standard. Estimation details can be found in Geweke and Zhou (1996) and Lopes and West (2004), and we do not repeat them here. For marginal likelihood computation, we also need the evaluation of the (integrated) likelihood; the analytical expression is given in the Appendix.

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 compute the marginal likelihood value using both Chib's and the cross-entropy methods. We report the results in Table 4.2.2.

In contrast to the previous application, here Chib's method and the cross-entropy method give slightly different marginal likelihood estimates. However, both methods are consistent in terms of the ranking of the models. In particular, both methods indicate that the 4-factor model is most preferred by the data. But the weight of evidence is relatively weak—the log Bayes factor in favor of the 4-factor model against the 3-factor model is less than 3.5 for both methods. Furthermore, both methods indicate a substantial initial increase in log marginal likelihood values—e.g. log marginal likelihood increases by about 270 from $k = 1$ to $k = 2$ factors—but the increase plateaus when $k = 3$.

Table 3: 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	-10039	-9768.1	-9687.8	-9685.6
Chib	-10025	-9753.9	-9673.4	-9670.0

Next, Table 4 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 respectively control the overall shrinkage strength of the intercepts $\boldsymbol{\beta}$ and factor loadings $\mathbf{a}$; κ_6 and κ_7 control the prior means of σ_i^2 and ω_i^2, respectively.

Table 4: 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	-4.5	-3.7	-24.7	-8.0	-4.5	-6.2	-24.9	-9.1
Chib	-4.5	-3.8	-25.1	-7.7	-4.5	-6.1	-27.3	-11.3

The results suggest that the marginal likelihood values are relatively insensitive to the four key hyperparameters. And with the possible exception of κ_5, they seem to have similar impact on the two factor models. Since κ_5 controls the overall shrinkage strength of the factor loadings $\mathbf{a}$, and the dimension of $\mathbf{a}$ grows with k, this result might not be surprising. Due to the differential impact of κ_5, when we increase κ_5, the 3-factor model is less penalized and performs better relative to the 4-factor model. Given that these two models have similar marginal likelihood values at the baseline setting, we conclude that they essentially receive the same support from the data. This highlights the value of performing a prior sensitivity analysis when comparing models via the marginal likelihood.

5 Concluding Remarks and Future Research

We have developed a general method based on Automatic Differentiation to compute the sensitivities of marginal likelihood with respect to a set of prior hyperparameters. We have illustrated the methodology using two empirical applications. While the conclusion in the VAR application 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 routinely conduct a prior sensitivity analysis in Bayesian model comparison.

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 are particularly useful when strong prior information is used, such as in estimating dynamic stochastic general equilibrium models.

Appendix: 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 $(\boldsymbol{\Omega}^{-1} + \mathbf{A}'\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}$.[5] 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

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

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

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.

Chan, J. C. C. (2018): "Large Bayesian VARs: A Flexible Kronecker Error Covariance Structure," *Journal of Business and Economic Statistics*, Forthcoming.

——— (2019): "Large Bayesian Vector Autoregressions," *CAMA Working Paper 19/2019.*

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 (2018): "How Sensitive Are VAR Forecasts to Prior Hyperparameters? An Automated Sensitivity Analysis," *CAMA Working Paper 25/2018.*

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.

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.

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.

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.

Koop, G., D. J. Poirier, and J. L. Tobias (2007): *Bayesian Econometric Methods.* Cambridge University Press.

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.

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.

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

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.

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