Fast Computation of the Deviance Information Criterion for Latent Variable Models

Joshua C.C. Chan*
Research School of Economics,
Australian National University

Angelia L. Grant†
Centre for Applied Macroeconomic Analysis,
Australian National University

July 2014

Abstract

The deviance information criterion (DIC) has been widely used for Bayesian model comparison. However, recent studies have cautioned against the use of certain variants of the DIC for comparing latent variable models. For example, it has been argued that the conditional DIC—based on the conditional likelihood obtained by conditioning on the latent variables—is sensitive to transformations of latent variables and distributions. Further, in a Monte Carlo study that compares various Poisson models, the conditional DIC almost always prefers an incorrect model. In contrast, the observed-data DIC—calculated using the observed-data likelihood obtained by integrating out the latent variables—seems to perform well. It is also the case that the conditional DIC based on the maximum a posteriori (MAP) estimate might not even exist, whereas the observed-data DIC does not suffer from this problem. In view of these considerations, fast algorithms for computing the observed-data DIC for a variety of high-dimensional latent variable models are developed. Through three empirical applications it is demonstrated that the observed-data DICs have much smaller numerical standard errors compared to the conditional DICs. The corresponding MATLAB code is available upon request.

Keywords: Bayesian model comparison, state space, factor model, vector autoregression, semiparametric model.

*Correspondence to: Research School of Economics, ANU College of Business and Economics, LF Crisp Building 26, The Australian National University, Canberra ACT 0200, Australia. Email: joshua.chan@anu.edu.au. Tel.: +61 2 612 57358; fax: +61 2 612 50182.

†Angelia Grant would like to acknowledge the Sir Roland Wilson Foundation for supporting her PhD studies.

1 Introduction

Hypothesis testing, and more generally model comparison, has long been an important problem in statistics and econometrics. Bayesian model comparison has traditionally been performed using the Bayes factor, which is defined to be the ratio of the marginal likelihoods of the two competing models. This model comparison criterion has a natural interpretation and is often easy to compute for a wide range of simple models (see, e.g., Kroese and Chan, 2014, pp. 251-254). However, the development of Markov chain Monte Carlo (MCMC) methods has made it possible to fit increasingly flexible and complex models, and estimating the marginal likelihoods of these typically high-dimensional models is often difficult. In fact, there is a vast and growing literature on marginal likelihood estimation using MCMC methods (see, e.g., Gelfand and Dey, 1994; Chib and Jeliazkov, 2001; Friel and Pettitt, 2008; Bauwens and Rombouts, 2012; Chan and Eisenstat, 2014, among many others). Despite these recent advances, computing the marginal likelihood remains a difficult problem in practice, which often involves nontrivial programming efforts and heavy computation. In addition, the values of the Bayes factor are often found to be sensitive to the choice of prior distributions.

These considerations have motivated the search for alternative model selection criteria. In particular, since Spiegelhalter et al. (2002) introduced the concept in their seminal paper, the deviance information criterion (DIC) has been widely used for Bayesian model comparison. Its popularity is further enhanced by the introduction of a number of alternative definitions of the DIC—many of them easy to compute—for latent variable models in Celeux et al. (2006).In addition, DIC computation is implemented in standard software packages, including WinBUGS. The DIC has been successfully applied to a wide variety of applications, such as comparing various stochastic volatility models in finance (see, e.g., Berg et al., 2004; Abanto-Valle et al., 2010; Wang et al., 2013), testing functional forms in energy modeling (see, e.g., Xiao et al., 2007), and discriminating between competing models for inflation as well as other macroeconomic time series (see, e.g., Lopes and Salazar, 2006; Chen et al., 2012; Mumtaz and Surico, 2012). A Monte Carlo study comparing the DIC with other Bayesian model selection criteria can be found in Ward (2008).

Nevertheless, some recent studies have cautioned against the use of the DIC for comparing latent variable models. For instance, Li et al. (2012) argue that the DIC should not be used with data augmentation, as the complete-data likelihood of the augmented data is nonregular and hence invalidates the standard asymptotic arguments that are needed to justify the DIC. Moreover, the DIC based on the complete-data likelihood is sensitive to transformations of latent variables and distributional representations. In the context of comparing Poisson models, Millar (2009) provides a Monte Carlo study which shows that the DIC based on the conditional likelihood—obtained by conditioning on the latent variables—almost always prefers the Poisson-gamma model instead of the Poisson-lognormal model, even when data are simulated from the latter. The author concludes that "the DIC is a potentially dangerous tool in the present context." In contrast, he shows that the DIC calculated using the integrated likelihood—obtained by

integrating out the latent variables—seems to perform well. This result is not surprising since standard asymptotic arguments for justifying the DIC apply to the DIC based on the integrated likelihood. However, evaluation of the integrated likelihood is typically time-consuming, which is the main reason why it is rarely used in applied work. We take a first step to address these issues by proposing fast methods for computing the DIC based on the integrated likelihood for a variety of high-dimensional latent variable models.

More specifically, the contribution of this paper is twofold. Firstly, we provide analytical expressions for the integrated likelihoods under three popular families of latent variable models: factor models, linear Gaussian state space models and semiparametric models. To evaluate these integrated likelihoods, we draw on recent advances in sparse matrix algorithms, and the computational details are carefully discussed. Secondly, we document the differences in variability of the DICs computed using the complete-data likelihood, the conditional likelihood and the integrated likelihood in three empirical examples. We show that the DICs based on the complete-data and conditional likelihoods generally have large numerical standard errors. On the other hand, the DICs based on the integrated likelihoods are more accurately estimated. This result is intuitive since integrating out the high-dimensional latent variables is expected to reduce the variance in Monte Carlo simulation. Our results provide another practical reason for why DICs based on conditional and complete-data likelihoods should not be used.

The rest of this paper is organized as follows. In Section 2 we introduce the concept of deviance and several definitions of the DIC. Section 3 discusses fast algorithms for computing the DIC based on the integrated likelihood for three classes of latent variable models. In Section 4, the proposed methods are illustrated via three empirical applications, involving returns on stock portfolios, US macroeconomic time series and female body mass index and wages.

2 Deviance Information Criterion

In complex hierarchical models, basic concepts like parameters and their dimension are not always clear and they may take several equally acceptable definitions. In their seminal paper, Spiegelhalter et al. (2002) introduce the concept of *effective number of parameters* and develop the theory of *deviance information criterion* (DIC) for model comparison. The model selection criterion is based on the *deviance*, which is defined as

$$D(\boldsymbol{\theta}) = -2 \log f(\mathbf{y} \,|\, \boldsymbol{\theta}) + 2 \log h(\mathbf{y}),$$

where $f(\mathbf{y} \,|\, \boldsymbol{\theta})$ is the likelihood function of the parametric model and $h(\mathbf{y})$ is some fully specified standardizing term that is a function of the data alone. Then the effective number of parameters p_D is defined as

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\widetilde{\boldsymbol{\theta}}),$$

where

$$\overline{D(\boldsymbol{\theta})} = -2\mathbb{E}_{\boldsymbol{\theta}}[\log f(\mathbf{y} \,|\, \boldsymbol{\theta}) \,|\, \mathbf{y}] + 2 \log h(\mathbf{y})$$

is the posterior mean deviance and $\widetilde{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}$, which is typically taken as the posterior mean or mode. Then, the deviance information criterion is defined as

$$\text{DIC} = \overline{D(\boldsymbol{\theta})} + p_D.$$

The posterior mean deviance can be used as a Bayesian measure of model fit or adequacy. Hence, the deviance information criterion, which is the sum of the posterior mean deviance and the effective number of parameters, can be viewed as a trade-off between model adequacy and complexity. For model comparison, we set $h(\mathbf{y}) = 1$ for all models. Therefore, the DIC becomes

$$\begin{aligned} \text{DIC} &= D(\widetilde{\boldsymbol{\theta}}) + 2p_D \\ &= -4\mathbb{E}_{\boldsymbol{\theta}}[\log f(\mathbf{y} \,|\, \boldsymbol{\theta}) \,|\, \mathbf{y}] + 2 \log f(\mathbf{y} \,|\, \widetilde{\boldsymbol{\theta}}). \end{aligned}$$

Given a set of competing models for the data, the preferred model is the one with the minimum DIC value.

In a subsequent paper, Celeux et al. (2006) point out that there are a number of alternative, yet natural, definitions of the DIC in latent variable models. To set the stage, suppose we augment the model $f(\mathbf{y} \,|\, \boldsymbol{\theta})$ with a vector of latent variables $\mathbf{z}$ such that

$$f(\mathbf{y} \,|\, \boldsymbol{\theta}) = \int f(\mathbf{y}, \mathbf{z} \,|\, \boldsymbol{\theta}) \mathrm{d}\mathbf{z},$$

where $f(\mathbf{y}, \mathbf{z} \,|\, \boldsymbol{\theta})$ is the *complete-data likelihood*. To avoid ambiguity, we refer to the likelihood $f(\mathbf{y} \,|\, \boldsymbol{\theta})$ as the *observed-data likelihood* or the *integrated likelihood*. In what follows, we discuss three distinct definitions of the DIC, the naming of which follows Celeux et al. (2006). One of the definitions is based on the integrated likelihood, one is based on the complete-data likelihood and one is based on the conditional likelihood. These definitions are chosen given that one goal of this paper is to show that the DICs based on the complete-data and conditional likelihoods have larger numerical standard errors relative to the DIC based on the integrated likelihood. The readers are referred to Celeux et al. (2006) for discussion of other variants.

When the integrated likelihood can be evaluated quickly, one can use the original definition of the DIC. In particular, consider

$$\text{DIC}_2 = -4\mathbb{E}_{\boldsymbol{\theta}}[\log f(\mathbf{y} \,|\, \boldsymbol{\theta}) \,|\, \mathbf{y}] + 2 \log f(\mathbf{y} \,|\, \widehat{\boldsymbol{\theta}}), \tag{1}$$

where the estimate $\widetilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is set as the posterior mode $\widehat{\boldsymbol{\theta}}$. The first term on the right-hand side of (1), i.e., the expectation $\mathbb{E}_{\boldsymbol{\theta}}[\log f(\mathbf{y} \,|\, \boldsymbol{\theta}) \,|\, \mathbf{y}]$, can be estimated by averaging the log-integrated likelihoods $\log f(\mathbf{y} \,|\, \boldsymbol{\theta})$ over the posterior draws of $\boldsymbol{\theta}$. Since we have analytically "integrated out" the typically high-dimensional vector of latent variables $\mathbf{z}$, this expectation can often be estimated precisely. In addition, to avoid the potentially

difficult optimization problem involved in locating the posterior mode $\widehat{\boldsymbol{\theta}}$, one often approximates it by the draw that has the highest value of $f(\mathbf{y} \,|\, \boldsymbol{\theta}) f(\boldsymbol{\theta})$ among the posterior draws, where $f(\boldsymbol{\theta})$ is the prior density. Once we have obtained an approximation of $\widehat{\boldsymbol{\theta}}$, the second term on the right-hand side of (1) can be readily computed.

However, it is often time-consuming to compute the integrated likelihood in a wide variety of latent variable models. In those cases, one often considers alternative definitions of the DIC that are based on the complete-data likelihood, $f(\mathbf{y}, \mathbf{z} \,|\, \boldsymbol{\theta})$, or the *conditional likelihood*, $f(\mathbf{y} \,|\, \boldsymbol{\theta}, \mathbf{z})$. For example, consider

$$\mathrm{DIC}_5 = -4\mathbb{E}_{\boldsymbol{\theta},\mathbf{Z}}[\log f(\mathbf{y}, \mathbf{Z} \,|\, \boldsymbol{\theta}) \,|\, \mathbf{y}] + 2 \log f(\mathbf{y}, \widehat{\mathbf{z}} \,|\, \widehat{\boldsymbol{\theta}}), \tag{2}$$

where $(\widehat{\mathbf{z}}, \widehat{\boldsymbol{\theta}})$ is the joint maximum a posteriori (MAP) estimate of the pair $(\mathbf{z}, \boldsymbol{\theta})$ given the data $\mathbf{y}$. The latent variable structure is usually chosen so that the joint distribution $f(\mathbf{y}, \mathbf{z} \,|\, \boldsymbol{\theta})$ is available in closed form. Hence, the first term on the right-hand side of (2) can be estimated by averaging the log-complete-data likelihoods $\log f(\mathbf{y}, \mathbf{z} \,|\, \boldsymbol{\theta})$ over the posterior draws of the pair $(\mathbf{z}, \boldsymbol{\theta})$. However, even though this expectation can be consistently estimated via simulation, its variance is likely to be large since we need to average over the typically high-dimensional vector of latent variables $\mathbf{z}$. To compute the second term, we can again approximate the MAP estimate by the best pair among the posterior draws, i.e., the pair that has the highest value of $f(\mathbf{y}, \mathbf{z} \,|\, \boldsymbol{\theta}) f(\boldsymbol{\theta})$.

The last alternative definition of the DIC that we consider is based on the conditional likelihood, which can typically be evaluated quickly:

$$\mathrm{DIC}_7 = -4\mathbb{E}_{\boldsymbol{\theta},\mathbf{Z}}[\log f(\mathbf{y} \,|\, \boldsymbol{\theta}, \mathbf{Z}) \,|\, \mathbf{y}] + 2 \log f(\mathbf{y} \,|\, \widehat{\mathbf{z}}, \widehat{\boldsymbol{\theta}}), \tag{3}$$

where $(\widehat{\mathbf{z}}, \widehat{\boldsymbol{\theta}})$ is the joint MAP estimate. As before, the first term on the right-hand side of (3) can be estimated by averaging the log-conditional likelihoods $\log f(\mathbf{y} \,|\, \boldsymbol{\theta}, \mathbf{z})$ over the posterior draws of $(\mathbf{z}, \boldsymbol{\theta})$. This estimate is also expected to be imprecise due to the high-dimensional latent variable vector $\mathbf{z}$.

For the three definitions of the DIC discussed in this section, the estimate $\widetilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is set as the posterior mode $\widehat{\boldsymbol{\theta}}$. Additional variants can be obtained by choosing different estimates of $\boldsymbol{\theta}$. For example, if we set $\widetilde{\boldsymbol{\theta}}$ to be the posterior mean in (1), we obtain what Celeux et al. (2006) call DIC_1. We refer the readers to Celeux et al. (2006) for further discussion of other variants. Our conclusion—that the DICs based on the conditional and complete-data likelihoods have large numerical standard errors—remains the same when they are alternatively defined using the posterior mean. While DICs based on the MAP are widely used in applied work, care must be taken when one uses DIC_5 and DIC_7: the joint MAP estimator might not exist even for models with proper priors. That is, a proper prior ensures a proper posterior, but not the existence of a posterior mode. This is a potentially serious issue as using only a posterior sample to estimate the joint MAP might lead to a spurious answer, with no indication that there is a problem. (We are grateful to an anonymous referee for pointing out this issue). The priors in the three applications in this paper are chosen to ensure the existence of the joint MAP estimator.

3 Fast Computation of the Observed-Data DIC

In this section we discuss fast methods—building upon recent advances in sparse matrix algorithms—for computing the observed-data DICs under three families of latent variable models: factor models, linear Gaussian state space models and semiparametric models. In particular, we derive analytical expressions for the integrated likelihoods under these models and discuss the computational details of implementing the integrated likelihood evaluation.

3.1 Factor Model

Factor models have been used in many areas including psychology, genomics, epidemiology, economics and finance. They are particularly useful for modeling the dependence structure of high-dimensional data. One central interest in factor analysis is to determine the number of latent factors. In this section we consider the following k-factor model, where the $n \times 1$ vector of observations $\mathbf{y}_t$ depends on a vector of k latent factors $\mathbf{f}_t$:

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta} + \mathbf{A}\mathbf{f}_t + \boldsymbol{\varepsilon}_t, \tag{4}$$

where $\mathbf{f}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$, $\mathbf{A}$ is the associated $n \times k$ loading matrix, $\mathbf{X}_t$ is an $n \times m$ design matrix, $\boldsymbol{\beta}$ is the $m \times 1$ vector of coefficients and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ assumed to be diagonal, i.e., $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. For the purpose of identification, we assume $\boldsymbol{\Omega} = \text{diag}(\omega_1^2, \ldots, \omega_k^2)$ is diagonal and $\mathbf{A}$ is lower triangular where the diagonal elements are unity. In addition, we also require $n \geqslant 2k + 1$ (see, for example, the discussion in Geweke and Zhou, 1996).

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 the integrated likelihood for this model 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 ameliorate this computation problem by using the Woodbury matrix identity:

$$(\mathbf{A}\boldsymbol{\Omega}\mathbf{A}' + \boldsymbol{\Sigma})^{-1} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{A}(\boldsymbol{\Omega}^{-1} + \mathbf{A}'\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}'\boldsymbol{\Sigma}^{-1}, \tag{5}$$

which only requires computing the $k \times k$ inverse $(\boldsymbol{\Omega}^{-1} + \mathbf{A}'\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}$. 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 (4) 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:

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

Finally, the log-integrated likelihood of this model is given by

$$\begin{aligned} \log f(\mathbf{y} \,|\, \boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\Omega}, \boldsymbol{\Sigma}) = &- \frac{Tn}{2} \log(2\pi) - \frac{T}{2} \log |\mathbf{A}\boldsymbol{\Omega}\mathbf{A}' + \boldsymbol{\Sigma}| \\ &- \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \left(\mathbf{I}_T \otimes (\mathbf{A}\boldsymbol{\Omega}\mathbf{A}' + \boldsymbol{\Sigma})^{-1} \right) (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned} \tag{6}$$

We comment on a few computational details in evaluating the log-integrated likelihood given in (6). First, the inverse $(\mathbf{A}\boldsymbol{\Omega}\mathbf{A}' + \boldsymbol{\Sigma})^{-1}$ can be obtained by the Woodbury matrix identity in (5). In computing this quantity, note that $\boldsymbol{\Sigma}^{-1}$ is an $n \times n$ diagonal matrix and is thus sparse. Using sparse matrix routines to compute (5) can therefore further speed up the calculations. Similarly, the quadratic term in (6) can be obtained by using fast sparse matrix routines since $\mathbf{I}_T \otimes (\mathbf{A}\boldsymbol{\Omega}\mathbf{A}' + \boldsymbol{\Sigma})^{-1}$ is a block-diagonal matrix. Lastly, to compute the log-determinant quickly and accurately, we can first obtain the Cholesky factor $\mathbf{C}_\mathbf{y}$ of $\mathbf{A}\boldsymbol{\Omega}\mathbf{A}' + \boldsymbol{\Sigma}$. Since $\mathbf{C}_\mathbf{y}$ is lower triangular, we can return $\log |\mathbf{A}\boldsymbol{\Omega}\mathbf{A}' + \boldsymbol{\Sigma}| = 2 \log |\mathbf{C}_\mathbf{y}| = 2 \sum_{i=1}^n \log c_{ii}$, where $c_{ii}, i = 1, \ldots, n$ are the diagonal elements of $\mathbf{C}_\mathbf{y}$.

3.2 Linear Gaussian State Space Model

A wide variety of popular macroeconometric models can be written in state space form. These include autoregressive moving average models, time-varying parameter vector autoregressions (VARs) and factor-augmented VARs, among many others. In this section we consider the following linear Gaussian state space model, where the $n \times 1$ vector of observations $\mathbf{y}_t$ depends linearly on the $q \times 1$ latent state vector $\boldsymbol{\beta}_t$ according to the hidden Markov structure

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

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\zeta}_t, \tag{8}$$

where $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\zeta}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$ are independent for all leads and lags, $\boldsymbol{\gamma}$ is a $k \times 1$ vector of time-invariant parameters, $\mathbf{W}_t$ and $\mathbf{X}_t$ are respectively $n \times k$ and $n \times q$ covariate matrices and the state equation (8) is initialized with $\boldsymbol{\beta}_1 \sim \mathcal{N}(\mathbf{b}_0, \mathbf{Q}_0)$ for constant matrices $\mathbf{b}_0$ and $\mathbf{Q}_0$. We note that instead of the random walk transition equation in (8), one can assume a stationary transition equation and estimation follows similarly; see, e.g., Chan and Jeliazkov (2009). For later reference, let $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_T')'$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_T')'$ and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1', \ldots, \boldsymbol{\varepsilon}_T')'$.

The linear Gaussian state space model (7)–(8) can be estimated using forward-filtering and backward-smoothing methods such as those in Carter and Kohn (1994) and Durbin and Koopman (2002). Recently, more efficient algorithms that exploit the band structure of the precision matrix of $p(\boldsymbol{\beta} \,|\, \mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$—the joint density of the states given the data

and other model parameters—are considered in Chan and Jeliazkov (2009) and McCausland et al. (2011); see also McCausland (2012) and Chan et al. (2013) for extensions to nonlinear models. In addition, for linear Gaussian state space models, $p(\boldsymbol{\beta} \,|\, \mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$ is a Gaussian density that can be evaluated quickly using fast band matrix routines. Hence, the integrated likelihood can also be evaluated quickly via the identity

$$f(\mathbf{y} \,|\, \boldsymbol{\gamma}, \boldsymbol{\Sigma}, \boldsymbol{\Omega}) = \frac{f(\mathbf{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}) p(\boldsymbol{\beta} \,|\, \boldsymbol{\gamma}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})}{p(\boldsymbol{\beta} \,|\, \mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})},$$

where $f(\mathbf{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma})$ is the complete-data likelihood function and $p(\boldsymbol{\beta} \,|\, \boldsymbol{\gamma}, \boldsymbol{\Sigma}, \boldsymbol{\Omega}) = p(\boldsymbol{\beta} \,|\, \boldsymbol{\Omega})$ is the conditional prior density for $\boldsymbol{\beta}$. Here we continue the line of research on efficient algorithms by deriving an explicit expression for the integrated likelihood of the model in (7)–(8) by analytically integrating out the states. By eliminating redundant terms, evaluating the integrated likelihood using the new expression will be faster.

To that end, we stack the measurement equation (7) over t:

$$\mathbf{y} = \mathbf{W}\boldsymbol{\gamma} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{9}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_T \otimes \boldsymbol{\Sigma})$,

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \\ \vdots \\ \mathbf{W}_T \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \dots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{X}_T \end{pmatrix}.$$

It follows from (9) that $(\mathbf{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}) \sim \mathcal{N}(\mathbf{W}\boldsymbol{\gamma} + \mathbf{X}\boldsymbol{\beta}, \mathbf{I}_T \otimes \boldsymbol{\Sigma})$ and hence the sampling density $f(\mathbf{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma})$ is given by

$$f(\mathbf{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{Tn}{2}} |\boldsymbol{\Sigma}|^{-\frac{T}{2}} \mathrm{e}^{-\frac{1}{2}(\mathbf{y}-\mathbf{W}\boldsymbol{\gamma}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})(\mathbf{y}-\mathbf{W}\boldsymbol{\gamma}-\mathbf{X}\boldsymbol{\beta})}. \tag{10}$$

Since $\mathbf{X}$ and $\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1}$ are band matrices, the sampling density $f(\mathbf{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma})$ can be evaluated quickly using band matrix routines.

Next, let $\mathbf{H}$ denote the first difference matrix, i.e.,

$$\mathbf{H} = \begin{pmatrix} \mathbf{I}_q & \mathbf{0} & \cdots & \mathbf{0} \\ -\mathbf{I}_q & \mathbf{I}_q & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \cdots & -\mathbf{I}_q & \mathbf{I}_q \end{pmatrix}.$$

Then, we can rewrite (8) as

$$\mathbf{H}\boldsymbol{\beta} = \widetilde{\boldsymbol{\alpha}} + \boldsymbol{\zeta},$$

where $\widetilde{\boldsymbol{\alpha}} = (\mathbf{b}_0', \mathbf{0}, \dots, \mathbf{0})'$, $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1', \dots, \boldsymbol{\zeta}_T')' \sim \mathcal{N}(\mathbf{0}, \mathbf{S})$ and $\mathbf{S} = \mathrm{diag}(\mathbf{Q}_0, \boldsymbol{\Omega}, \dots, \boldsymbol{\Omega})$. Since the determinant of $\mathbf{H}$ is unity, it is invertible. Let $\boldsymbol{\alpha} = \mathbf{H}^{-1}\widetilde{\boldsymbol{\alpha}}$. Then, we have $(\boldsymbol{\beta} \,|\, \boldsymbol{\Omega}) \sim \mathcal{N}(\boldsymbol{\alpha}, (\mathbf{H}'\mathbf{S}^{-1}\mathbf{H})^{-1})$, and therefore the prior density $p(\boldsymbol{\beta} \,|\, \boldsymbol{\Omega})$ is given by

$$p(\boldsymbol{\beta} \,|\, \boldsymbol{\Omega}) = (2\pi)^{-\frac{Tq}{2}} |\mathbf{Q}_0|^{-\frac{1}{2}} |\boldsymbol{\Omega}|^{-\frac{T-1}{2}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\alpha})'\mathbf{H}'\mathbf{S}^{-1}\mathbf{H}(\boldsymbol{\beta}-\boldsymbol{\alpha})}. \tag{11}$$

Since both $\mathbf{H}$ and $\mathbf{S}^{-1}$ are again band matrices, this prior density $f(\boldsymbol{\beta} \,|\, \boldsymbol{\Omega})$ can also be evaluated quickly using fast routines for band matrices.

In Appendix A we show that the log-integrated likelihood $\log f(\mathbf{y} \,|\, \boldsymbol{\gamma}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$ is given by:

$$\begin{aligned} &\log f(\mathbf{y} \,|\, \boldsymbol{\gamma}, \boldsymbol{\Sigma}, \boldsymbol{\Omega}) \\ &\quad = -\frac{Tn}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{Q}_0| - \frac{T-1}{2}\log|\boldsymbol{\Omega}| - \frac{T}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\log|\mathbf{K}_{\boldsymbol{\beta}}| \\ &\qquad - \frac{1}{2}\left((\mathbf{y}-\mathbf{W}\boldsymbol{\gamma})'(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})(\mathbf{y}-\mathbf{W}\boldsymbol{\gamma}) + \boldsymbol{\alpha}'\mathbf{H}'\mathbf{S}^{-1}\mathbf{H}\boldsymbol{\alpha} - \mathbf{d}_{\boldsymbol{\beta}}'\mathbf{K}_{\boldsymbol{\beta}}^{-1}\mathbf{d}_{\boldsymbol{\beta}}\right), \end{aligned} \tag{12}$$

where $\mathbf{K}_{\boldsymbol{\beta}} = \mathbf{X}'(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{X} + \mathbf{H}'\mathbf{S}^{-1}\mathbf{H}$ and $\mathbf{d}_{\boldsymbol{\beta}} = \mathbf{X}'(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})(\mathbf{y} - \mathbf{W}\boldsymbol{\gamma}) + \mathbf{H}'\mathbf{S}^{-1}\mathbf{H}\boldsymbol{\alpha}$.

We comment on a few computational details in evaluating the log-integrated likelihood given in (12). First, recall that $\mathbf{S}$, $\mathbf{H}$ and $\mathbf{X}$ are all band matrices. Consequently, $\mathbf{d}_{\boldsymbol{\beta}}$ can be obtained quickly. Moreover, $\mathbf{K}_{\boldsymbol{\beta}}$ is also a band matrix. Second, to obtain the product $\mathbf{K}_{\boldsymbol{\beta}}^{-1}\mathbf{d}_{\boldsymbol{\beta}}$, one can simply solve the band linear system $\mathbf{K}_{\boldsymbol{\beta}}\mathbf{x} = \mathbf{d}_{\boldsymbol{\beta}}$ for $\mathbf{x}$. This avoids the time-consuming operation of computing the inverse $\mathbf{K}_{\boldsymbol{\beta}}^{-1}$. Third, calculating the log determinant $\log|\mathbf{K}_{\boldsymbol{\beta}}|$ can be done quickly as follows: obtain the Cholesky factor $\mathbf{C}_{\boldsymbol{\beta}}$ of $\mathbf{K}_{\boldsymbol{\beta}}$, which can be completed quickly as $\mathbf{K}_{\boldsymbol{\beta}}$ is a band matrix. Since $\mathbf{C}_{\boldsymbol{\beta}}$ is lower triangular, we can return $\log|\mathbf{K}_{\boldsymbol{\beta}}| = 2\log|\mathbf{C}_{\boldsymbol{\beta}}| = 2\sum_{i=1}^{Tq}\log c_{ii}$, where $c_{ii}, i = 1, \ldots, Tq$ are the diagonal elements of $\mathbf{C}_{\boldsymbol{\beta}}$.

3.3 Semiparametric Regression

Another popular model used in economics is the semiparametric regression. Economic theory rarely dictates a functional form for the regression relationship between the dependent variable and the regressors, while a fully nonparametric regression suffers from the curse of dimensionality. Hence, it is often desirable to consider the partial linear regression or semiparametric regression, which can be written as follows:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + f(z_i) + \varepsilon_i \tag{13}$$

for $i = 1, \ldots, n$, where y_i is a scalar dependent variable, z_i is a scalar explanatory variable, $\mathbf{x}_i$ is a $k \times 1$ vector of explanatory variables, $\boldsymbol{\beta}$ is a $k \times 1$ vector of parameters, $f(\cdot)$ is an unknown function and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. In other words, only z_i is treated nonparametrically and all other variables in $\mathbf{x}_i$ enter the regression linearly. Also note that since the unknown function $f(\cdot)$ plays the role of an intercept, $\mathbf{x}_i$ does not include an intercept.

To estimate the semiparametric regression in (13), we follow Koop and Poirier (2004) and specify a hierarchical prior on the functional values. For a textbook treatment, see, Koop et al. (2007), pp. 187-190; for various extensions, see, e.g., Koop et al. (2005), Kline and Tobias (2008) and Chib et al. (2009). We first sort the data by values of z such that $z_1 < z_2 < \cdots < z_m$. Note that we allow for the possibility that different observations may have the same values of z so that $m \leqslant n$. Next, let $\theta_i = f(z_i)$ for $i = 1, \ldots, m$ and

stack $\mathbf{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$. We can then rewrite the semiparametric regression (13) as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{14}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$, $\mathbf{X}$ is an $n \times k$ matrix of regressors and $\mathbf{D}$ is an $n \times m$ matrix that selects the appropriate element in $\boldsymbol{\theta}$, i.e., each row of $\mathbf{D}$ is a $1 \times m$ vector of zeros except one unity entry that picks the appropriate element in $\boldsymbol{\theta}$. Hence, if we sort the data by values of z, the associated matrix $\mathbf{D}$ is in fact a band matrix. See also the discussion in Chib et al. (2009).

Define $\Delta_i = z_i - z_{i-1} > 0$ for $i = 2, \ldots, m$ and construct the $m \times m$ matrix $\mathbf{G}$ as follows:

$$\mathbf{G} = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \Delta_2^{-1} & -(\Delta_3^{-1} + \Delta_2^{-1}) & \Delta_3^{-1} & 0 & \cdots & 0 & 0 & 0 \\ 0 & \Delta_3^{-1} & -(\Delta_4^{-1} + \Delta_3^{-1}) & \Delta_4^{-1} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \Delta_{m-1}^{-1} & -(\Delta_m^{-1} + \Delta_{m-1}^{-1}) & \Delta_m^{-1} \end{pmatrix}.$$

It follows that

$$\mathbf{G}\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \frac{\theta_3 - \theta_2}{\Delta_3} - \frac{\theta_2 - \theta_1}{\Delta_2} \\ \frac{\theta_4 - \theta_3}{\Delta_4} - \frac{\theta_3 - \theta_2}{\Delta_3} \\ \vdots \\ \frac{\theta_m - \theta_{m-1}}{\Delta_m} - \frac{\theta_{m-1} - \theta_{m-2}}{\Delta_{m-1}} \end{pmatrix},$$

where terms of the form $(\theta_i - \theta_{i-1})/\Delta_i = (f(z_i) - f(z_{i-1}))/(z_i - z_{i-1})$ can be interpreted as pointwise derivatives of $f(\cdot)$ at z_{i-1}.

We then consider the prior $(\mathbf{G}\boldsymbol{\theta} \,|\, \tau) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_\tau)$, where $\boldsymbol{\Omega}_\tau = \operatorname{diag}(V_1, V_2, \tau, \ldots, \tau)$, and V_1 and V_2 are fixed constants. Since $\mathbf{G}$ is a lower triangular $m \times m$ matrix with positive determinant, it is invertible regardless of the data. Therefore, we can equivalently write the prior as

$$(\boldsymbol{\theta} \,|\, \tau) \sim \mathcal{N}(\mathbf{0}, (\mathbf{G}'\boldsymbol{\Omega}_\tau^{-1}\mathbf{G})^{-1}). \tag{15}$$

Now, combining the complete-data likelihood (14) and the prior (15), we can obtain an analytical expression for the integrated likelihood $f(\mathbf{y} \,|\, \boldsymbol{\beta}, \sigma^2, \tau)$. More precisely, the log-integrated likelihood of the semiparametric model is given by (see Appendix A for details):

$$\begin{aligned} \log f(\mathbf{y} \,|\, \boldsymbol{\beta}, \sigma^2, \tau) = &- \frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\log|\boldsymbol{\Omega}_\tau| + \log|\mathbf{G}| - \frac{1}{2}\log|\mathbf{K}_{\boldsymbol{\theta}}| \\ &- \frac{1}{2}\left(\frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \mathbf{d}_{\boldsymbol{\theta}}'\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{d}_{\boldsymbol{\theta}}\right), \end{aligned} \tag{16}$$

where $\mathbf{K}_{\boldsymbol{\theta}} = \mathbf{D}'\mathbf{D}/\sigma^2 + \mathbf{G}'\boldsymbol{\Omega}_\tau^{-1}\mathbf{G}$ and $\mathbf{d}_{\boldsymbol{\theta}} = \mathbf{D}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/\sigma^2$. Since $\mathbf{G}$, $\mathbf{K}_{\boldsymbol{\theta}}$ and $\boldsymbol{\Omega}_\tau$ are band matrices, the log-integrated likelihood in (16) can be quickly evaluated using band and sparse matrices routines; see the discussion on computations in Section 3.2.

4 Empirical Applications

In this section we illustrate the proposed algorithms introduced in Section 3 for computing the observed-data DICs for three classes of latent variable models. The complete-data and conditional DICs are also reported for comparison. The empirical applications involve returns on stock portfolios, US macroeconomic time series data and female body mass index and wages.

4.1 Factor Models for Stock Portfolio Returns

In the first application we analyze monthly excess returns for 10 NYSE/AMEX/NASDAQ market capitalization decile portfolios for the sample period January 1952 to December 2011. More precisely, the data are returns in excess of the one-month US Treasury bill yield, and all data are obtained from the Center for Research in Security Prices (CRSP) at the University of Chicago. A similar dataset is fitted using a variety of factor models in Nardari and Scruggs (2007) (their sample period ends in December 2003) and they find 3-factor models fit the data best using the Bayes factor as the model comparison criterion. Here we perform a similar model comparison exercise with the three alternative versions of the DIC.

More specifically, $\mathbf{y}_t$ is a column vector of size $n = 10$ consisting of the monthly excess returns of the decile portfolios, which we denote as *Cap1* (the smallest size decile portfolio) to *Cap10* (the largest size decile portfolio). Following Nardari and Scruggs (2007), the first two elements in $\mathbf{y}_t$ are *Cap6* and *Cap10*; other elements are the excess returns for the remaining decile portfolios arranged from the smallest to largest. As such, the first factor can be interpreted as a "stock market factor" and the second as a "size factor"; see Nardari and Scruggs (2007) for further discussion. We also include an intercept in the factor model (4), i.e., the design matrix $\mathbf{X}_t$ is the identity matrix of dimension n.

4.1.1 Priors and Results

We now discuss the specification of the priors for the k-factor model. Let $\mathbf{a}$ be the $l \times 1$ vector with $l = kn - k(k+1)/2$ that contains the free elements in the factor loadings $\mathbf{A}$. Then, we consider the following independent priors:

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, \mathbf{V}_{\boldsymbol{\beta}}), \quad \mathbf{a} \sim \mathcal{N}(\mathbf{a}_0, \mathbf{V}_{\mathbf{a}}), \quad \sigma_i^2 \sim \mathcal{IG}(\nu_{\sigma_i^2}, S_{\sigma_i^2}), \quad \omega_j^2 \sim \mathcal{IG}(\nu_{\omega_j^2}, S_{\omega_j^2}) \tag{17}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, k$, where $\mathcal{IG}(\cdot, \cdot)$ denotes the inverse-gamma distribution. The hyperparameters are set as $\boldsymbol{\beta}_0 = \mathbf{0}$, $\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{I}_n$, $\mathbf{a}_0 = \mathbf{0}$, $\mathbf{V}_{\mathbf{a}} = \mathbf{I}_l$, $\nu_{\sigma_i^2} = 3$, $S_{\sigma_i^2} = 2$, $\nu_{\omega_j^2} = 3$ and $S_{\omega_j^2} = 2$. These hyperparameters imply prior means $\mathbb{E}\sigma_i^2 = \mathbb{E}\omega_j^2 = 1$.

We consider four factor models with numbers of factors ranging from 1 to 4. (Recall that for identification we require $n \geqslant 2k+1$. Hence, we can allow for at most four factors.) For each model, we use the Gibbs sampler in Appendix B to construct 10 parallel chains each of which is of length 10000 after a burn-in period of 1000. We then use the algorithms described in Sections 2 and 3.1 to compute the three DICs. The results are reported in Table 1.

Table 1: Estimated DICs, numerical standard errors and computation times (in seconds) for the competing factor models.

	DIC_2	Time (s)	DIC_5	Time (s)	DIC_7	Time (s)
1-factor	26370	290	27059	327	23132	365
	(0.56)		(3.12)		(2.97)	
2-factor	22410	345	**23265**	387	16781	434
	(0.43)		(7.12)		(13.2)	
3-factor	22081	402	25336	452	15550	488
	(0.88)		(17.0)		(34.5)	
4-factor	**22065**	466	26309	527	**15198**	551
	(0.88)		(38.4)		(86.5)	

Among the factor models, both DIC_2 and DIC_7 indicate that the best model is the 4-factor model, whereas DIC_5 prefers the 2-factor model. This highlights the fact that model comparison using different definitions of the DIC might select different models. In addition, the results show that even with a moderate size of posterior draws (a total of 100000), it is still difficult to estimate DIC_5 and DIC_7 accurately as indicated by the relatively large numerical standard errors. These results are perhaps not surprising as the computation of DIC_5 and DIC_7 involves the high-dimensional latent factors, which increases variability of the Monte Carlo simulation. On the other hand, the DIC_2 estimates are more accurately estimated. It is also interesting to note that the computation times for DIC_2 are in fact less than those for DIC_5 and DIC_7.

4.2 Vector Autoregressions for the US Economy

In this empirical application we compare various popular VARs for fitting a US macroeconomic time series dataset, which is obtained from the US Federal Reserve Bank of St. Louis. Specifically, the dataset consists of 260 quarterly observations from 1948Q1 to 2012Q4 on $n = 4$ variables: real GDP growth, 3-Month Treasury Bill rate, unemployment rate and CPI inflation rate. The main goal of this exercise is to investigate which specification best models the evolution and interdependence among these macroeconomic

time series.

4.2.1 Time-Invariant and TVP-VARs

The first model we consider is a standard vector autoregression (VAR) model popularized by Sims (1980). More specifically, we consider the first-order VAR

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Pi}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t \tag{18}$$

for $t = 1, \ldots, T$, where $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, $\mathbf{y}_t$ is a vector containing measurements on the aforementioned $n = 4$ macroeconomic variables, $\boldsymbol{\mu}$ is an $n \times 1$ vector of intercepts and $\boldsymbol{\Pi}$ is an $n \times n$ matrix of VAR coefficients. For estimation (18) is often written in the form of a seemingly unrelated regression (SUR) model:

$$\mathbf{y}_t = \mathbf{W}_t\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_t, \tag{19}$$

where $\mathbf{W}_t = \mathbf{I}_n \otimes (1, \mathbf{y}'_{t-1})$, $\boldsymbol{\gamma}$ is a $q \times 1$ vector with $q = n(n+1)$, obtained by stacking the parameters in $\boldsymbol{\mu}$ and $\boldsymbol{\Pi}$ equation by equation, i.e., $\boldsymbol{\gamma} = \text{vec}\left((\boldsymbol{\mu},\ \boldsymbol{\Pi})'\right)$.

Although the conventional constant coefficients VAR has enjoyed great success, recent literature has highlighted the importance of allowing for potential structural instabilities in time series via time-varying parameters. Consequently, we also consider the following time-varying parameter vector autoregression (TVP-VAR) (e.g., Canova, 1993; Koop and Korobilis, 2010):

$$\mathbf{y}_t = \mathbf{X}_t\boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t, \tag{20}$$

where $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. The vector of VAR coefficients $\boldsymbol{\beta}_t$ evolves according to the random walk process

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\zeta}_t \tag{21}$$

for $t = 2, \ldots, T$, where $\boldsymbol{\beta}_1 \sim \mathcal{N}(\mathbf{b}_0, \mathbf{Q}_0)$, $\boldsymbol{\zeta}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$ and $\boldsymbol{\Omega} = \text{diag}(\omega_1^2, \ldots, \omega_q^2)$ is assumed to be a diagonal matrix.

Empirically, the TVP-VAR in (20) that allows all VAR coefficients to change over time is often found to perform better than a constant coefficients VAR. For example, Chan and Eisenstat (2014) find that the TVP-VAR is preferred to the constant coefficients VAR according to the Bayes factor. However, it is plausible that a TVP-VAR where only some coefficients are time-varying while others are time-invariant will perform better than both alternatives. This possibility has been investigated in, e.g., Belmonte et al. (2014), Koop and Korobilis (2012) and Eisenstat et al. (2014). We follow this line of research and consider four different restricted TVP-VARs. More specifically, we restrict the coefficients in each of the $n = 4$ equations to be time-invariant, whereas coefficients in the other $n-1$ equations are time-varying. Each of these restricted TVP-VARs can be written in the form of (7). For example, if we restrict the coefficients in the first equation to be constant, then

$$\mathbf{W}_t = \begin{pmatrix} (1, \mathbf{y}'_{t-1}) \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{X}_t = \begin{pmatrix} \mathbf{0} \\ \mathbf{I}_{n-1} \otimes (1, \mathbf{y}'_{t-1}) \end{pmatrix},$$

and $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_t$ are appropriately defined.

4.2.2 Priors and Results

We now specify the prior distributions under each of the models considered. For all cases, we assume proper but relatively noninformative prior distributions. Moreover, since the main goal of this exercise is to compare different specifications, we assume similar prior distributions across models where possible.

For the constant coefficients VAR, we assume the following independent prior distributions: $\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\gamma}_0, \mathbf{V}_{\boldsymbol{\gamma}}), \quad \boldsymbol{\Sigma} \sim \mathcal{IW}(\nu_{\boldsymbol{\Sigma}}, \mathbf{S}_{\boldsymbol{\Sigma}})$, where $\mathcal{IW}(\cdot, \cdot)$ denotes the inverse-Wishart distribution. Specifically, we set $\boldsymbol{\gamma}_0 = \mathbf{0}$, $\mathbf{V}_{\boldsymbol{\gamma}} = 5 \times \mathbf{I}_q$, $\nu_{\boldsymbol{\Sigma}} = n + 3$ and $\mathbf{S}_{\boldsymbol{\Sigma}} = \mathbf{I}_n$ so that the prior means are $\mathbb{E}\boldsymbol{\gamma} = \mathbf{0}$ and $\mathbb{E}\boldsymbol{\Sigma} = 0.5 \times \mathbf{I}_n$. Next, for the TVP-VAR where all the VAR coefficients are time-varying, we assume $\omega_i^2 \sim \mathcal{IG}(\nu_i, S_i)$, $i = 1, \ldots, q$, where $\mathcal{IG}(\cdot, \cdot)$ denotes the inverse-gamma distribution. We set the hyperparameters $\nu_{\boldsymbol{\Sigma}}$ and $\mathbf{S}_{\boldsymbol{\Sigma}}$ to be the same as in the constant coefficients VAR, i.e, $\nu_{\boldsymbol{\Sigma}} = n + 3$ and $\mathbf{S}_{\boldsymbol{\Sigma}} = \mathbf{I}_n$. Moreover, we set $\mathbf{b}_0 = \mathbf{0}, \mathbf{Q}_0 = 5 \times \mathbf{I}_q$, $\nu_i = 5$ and $S_i = 0.02$, so that $\mathbb{E}\sigma_i^2 = 0.005$ for $i = 1, \ldots, q$. Lastly, for the four TVP-VARs where the coefficients in the first equation are constant, we assume the same prior distributions for $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ as in the unrestricted TVP-VAR. For the time-invariant VAR coefficients $\boldsymbol{\gamma}$, we assume the prior $\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\gamma}_0, \mathbf{V}_{\boldsymbol{\gamma}})$, where $\boldsymbol{\gamma}_0 = \mathbf{0}$ and $\mathbf{V}_{\boldsymbol{\gamma}} = 5 \times \mathbf{I}_k$.

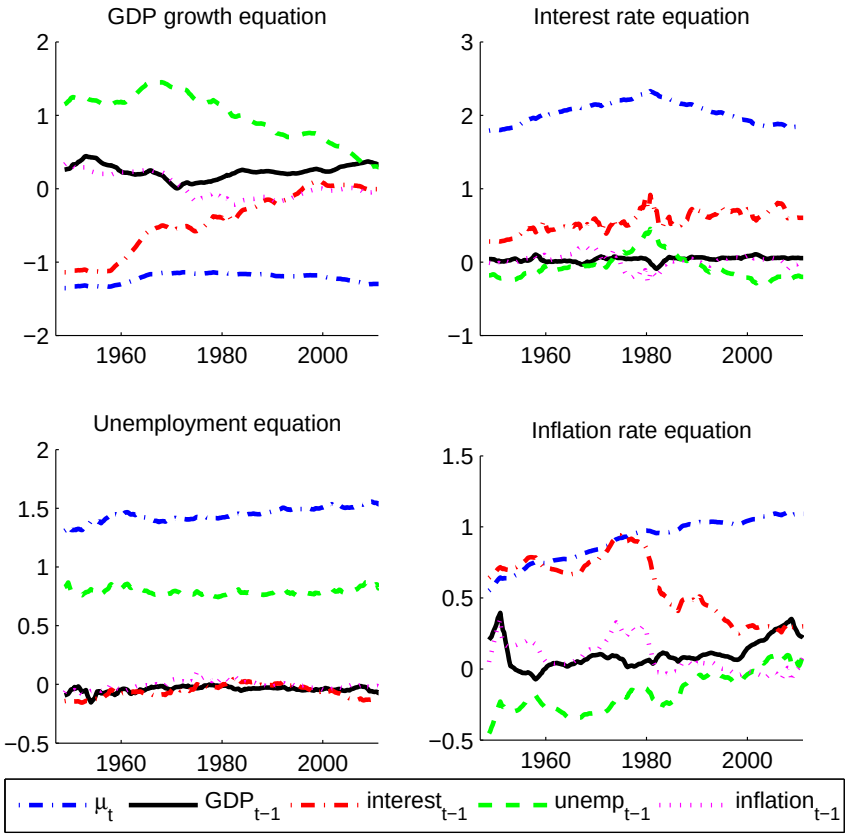

Figure 1: Evolution of the VAR coefficients $\boldsymbol{\beta}_t$ in the time-varying parameter VAR model.

For each of the models, we obtain a total of 100000 posterior draws from 10 parallel chains using the Gibbs sampler in Appendix B. The posterior estimates under the unrestricted TVP-VAR are reported in Figure 1. On the whole, the estimates suggest that there is substantial time-variation in the VAR coefficients. However, the parameters in the unemployment equation seem to vary less than those in other equations.

To investigate the possibility that a TVP-VAR where only some coefficients are time-varying while others are time-invariant is a better alternative, we estimate the three DICs for various TVP-VARs and the results are reported in Table 2. VAR denotes the constant coefficients VAR and TVP-VAR denotes the unrestricted TVP-VAR. We also consider four other variants of the TVP-VAR. For example, Const-GDP denotes a TVP-VAR where the coefficients of the GDP equation are constant, while those of the remaining three equations are time-varying; the others are defined similarly. Since the constant coefficients VAR does not involve any latent variables (states), its DIC is computed using the (observed-data) likelihood and we classify this as DIC_2. According to DIC_2, the best model is the variant of TVP-VAR with coefficients in the unemployment equation fixed, followed by the constant coefficients VAR. However, according to both DIC_5 and DIC_7, the best model is the TVP-VAR, followed by the TVP-VAR variant where the coefficients in the GDP equation are constant. This gives another real data example of where different models are considered "the best" according to different definitions of the DIC.

Table 2: Estimated DICs, numerical standard errors and computation times (in seconds) for the competing VAR models.

	DIC_2	Time (s)	DIC_5	Time (s)	DIC_7	Time (s)
VAR	3009.5	163	–	–	–	–
	(0.34)		–	–	–	–
TVP-VAR	3195.2	3020	**-13007**	1677	**1917.2**	1582
	(4.54)		(35.3)		(25.0)	
Const-GDP	3176.8	1819	-9690.0	1027	1985.5	1179
	(2.99)		(47.4)		(18.3)	
Const-Interest	3244.8	1833	-8439.7	1021	2572.4	1187
	(2.27)		(35.3)		(17.8)	
Const-Unemployment	**2998.1**	1823	-8042.7	1025	2168.9	1165
	(2.42)		(33.1)		(16.4)	
Const-Inflation	3204.1	1814	-9586.7	1022	2083.0	1160
	(2.41)		(29.4)		(15.6)	

Moreover, the results suggest that DIC_2 estimates are the most accurate among the three, but they also take more time to compute. However, DIC_2 is still more accurately estimated when the computation time is taken into account. For example, for computing the DICs of the TVP-VAR, the estimate of DIC_2 is about 58 times ($25^2/1582 \times 3020/4.54^2$) more accurate than that of DIC_7 in terms of variance reduction; it is over 100 times more accurate compared to the estimate of DIC_5. Since the vector of latent variables is of

high dimension in these TVP-VARs, even computing the complete-data likelihood-based DICs takes a long time. To reduce the numerical standard error by half requires quadruple computation time. Hence, discriminating between competing models using complete-data likelihood-based DICs might be impractical. On the other hand, DIC_2 can be computed accurately using the computation methods described in Section 3.2.

4.3 Semiparametric Regression: The Wages of BMI

In this section we revisit an empirical application in Kline and Tobias (2008) that studies the role of body mass index (BMI) in the production of log wages. More specifically, they consider a semiparametric treatment-response model with a skewnormal error distribution. Here we abstract from the non-Gaussian assumption and the endogeneity issue; we consider the semiparametric regression in (13) and investigate if the main conclusions hold in this simplified framework. In particular, we compare different priors for the smoothness parameter τ; we also compare the semiparametric regression with a conventional linear regression.

4.3.1 Data

The data are from the 1970 British Cohort Study that tracks the cohort of all people born in Great Britain between 5–11 April, 1970. We use a subsample that contains $n = 1782$ women with $m = 672$ unique values of BMI. The covariates that are treated linearly (contained in $\mathbf{x}_i$) are individual and family background characteristics. Individual characteristics include tenure on the current job, labor market experience, family income, an indicator denoting the completion of a lower level of secondary education, an indicator denoting the completion of a higher level of secondary education, an indicator for the completion of a college degree program, an indicator denoting whether the individual is married and an indicator denoting whether the individual has a union job.

For family background characteristics, we have an indicator denoting whether the individual's mother or father held a college degree, an indicator denoting whether the individual's mother or father worked in a managerial or professional position, and the BMIs of the individual's mother and father. For a more detailed discussion of the dataset and definitions of various variables, see Kline and Tobias (2008).

4.3.2 Priors and Results

For the semiparametric regression, we consider the following independent priors:

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, \mathbf{V}_{\boldsymbol{\beta}}), \quad \sigma^2 \sim \mathcal{IG}(\nu_{\sigma^2}, S_{\sigma^2}), \quad \tau \sim \mathcal{IG}(\nu_\tau, S_\tau), \tag{22}$$

where the hyperparameters are set as $\boldsymbol{\beta}_0 = \mathbf{0}$, $\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{I}_k$, $\nu_{\sigma^2} = 3$ and $S_{\sigma^2} = 2$. For τ, we consider three different sets of hyperparameters: $\nu_\tau = 3$ and $S_\tau = 10^{-4}, 10^{-5}$ and

10^{-6}. These hyperparameters imply prior means $\mathbb{E}\tau = 5 \times 10^{-5}$, $\mathbb{E}\tau = 5 \times 10^{-6}$ and $\mathbb{E}\tau = 5 \times 10^{-7}$, respectively. For the conventional linear regression, we simply append an intercept and the individual's BMI to the vector of covariates $\mathbf{x}_i$. The priors for $\boldsymbol{\beta}$ and σ^2 are the same as given in (22).

For each model, we use the Gibbs sampler in Appendix B to construct 10 parallel chains each of which is of length 10000 after a burn-in period of 1000. The posterior draws are then used to compute the three DICs, and the results are reported in Table 3. Among the semiparametric regressions, both DIC_2 and DIC_7 indicate that the best model is the one with a smoothness prior that implies $\mathbb{E}\tau = 5 \times 10^{-6}$, which is also the choice in Kline and Tobias (2008). However, DIC_5 prefers the model with prior mean $\mathbb{E}\tau = 5 \times 10^{-7}$. This again highlights the fact that different definitions of the DIC might prefer different models. Compared with the semiparametric regressions, the conventional linear regression has a slightly lower DIC (DIC_2), indicating that the extra flexibility of the semiparametric regressions does not seem to be justified by the better fit. Nevertheless, it may be justified if the main interest is on the possibly nonlinear impact of BMI on log wages (the estimate of $f(\cdot)$ does appear to be nonlinear; see Figure 2).

Table 3: Estimated DICs, numerical standard errors and computation times (in seconds) for the competing semiparametric models.

	DIC_2	Time (s)	DIC_5	Time (s)	DIC_7	Time (s)
Linear regression	**1362.0**	38	–	–	–	–
	(0.20)		–	–	–	–
Semiparametric;	1397.3	171	-8566.8	128	1341.2	134
$\mathbb{E}\tau = 5 \times 10^{-5}$	(0.40)		(23.0)		(1.80)	
Semiparametric;	1392.7	173	-9709.8	130	**1340.5**	132
$\mathbb{E}\tau = 5 \times 10^{-6}$	(0.41)		(32.7)		(1.49)	
Semiparametric;	1426.6	172	**-10894**	132	1341.8	133
$\mathbb{E}\tau = 5 \times 10^{-7}$	(1.90)		(40.0)		(2.27)	

As for the numerical standard errors of the different DICs, the DIC_2 estimates are the most accurate at a slightly higher computational cost as expected. For example, when computing the DICs for the semiparametric regression with prior mean $\mathbb{E}\tau = 5 \times 10^{-6}$, the estimate of DIC_2 is about 17 times ($(1.49)^2/132 \times 173/(0.41)^2$) more accurate than that of DIC_7 after accounting for the computation times (in terms of variance reduction). The estimates for DIC_5 are quite inaccurate even though they are computed using a total of 100000 posterior draws. It is also worth noting that even though the computation times for calculating DIC_5 and DIC_7 are similar, the latter are about 300-500 times more accurate than the former (in terms of variance reduction).

Finally, in Figure 2 we report the posterior means of the function $f(\cdot)$ with the prior $\mathbb{E}\tau = 5 \times 10^{-6}$. The shape of the curve is qualitatively similar to that reported in Kline and Tobias (2008). More specifically, the marginal increases in BMI incur little penalty in terms of wages, if at all, for individuals with low values of BMI (18-22). But when the

individual moves towards higher values of BMI, the penalty increases until it levels off from about 28 (roughly the 85-th percentile of the BMI distribution).

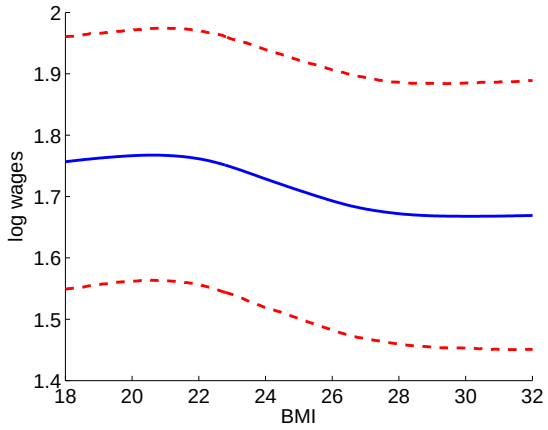

Figure 2: Posterior means of the function $f(\cdot)$ with the prior mean $\mathbb{E}\tau = 5 \times 10^{-6}$ (solid line) and the corresponding 90% credible intervals (dotted lines).

5 Concluding Remarks and Future Research

We have derived analytical expressions for the integrated likelihoods under three classes of latent variable models: factor models, linear Gaussian state space models and semi-parametric models, with the goal of evaluating the observed-data DICs. In the empirical examples, we found that given the same computation time (and with care taken to ensure that the joint MAP estimator exists), complete-data and conditional DICs typically have much larger numerical standard errors compared to the DICs based on the integrated likelihoods. This highlights the need to report numerical standard errors of the DICs, which is often not done in empirical research.

The analytical expressions for the integrated likelihoods derived in this paper can be used to develop more efficient MCMC samplers for estimation and algorithms for computing the marginal likelihood. We leave these possibilities for future research. In addition, we have only considered models where the integrated likelihoods are available analytically. One popular family of models that do not fit into this framework is stochastic volatility models. However, it is still possible to quickly evaluate the integrated likelihood of a certain subset of these models using importance sampling (e.g., McCausland, 2012, is one such example). For models where the integrated likelihood cannot be quickly evaluated, such as general nonlinear and non-Gaussian models, one could consider other model selection criteria such as the marginal likelihood or the robust deviance information criterion of Li et al. (2012), both of which require evaluating the integrated likelihood only once.

Appendix A: Derivation of the Integrated Likelihoods

In this appendix we provide the details of the derivation of the integrated likelihoods of the linear Gaussian state space model in (7)–(8) and the semiparametric model in (14)–(15).

We first consider the integrated likelihood for the linear Gaussian state space model in (7)–(8). Let $c_1 = (2\pi)^{-\frac{T(n+q)}{2}}|\mathbf{Q}_0|^{-\frac{1}{2}}|\boldsymbol{\Omega}|^{-\frac{T-1}{2}}|\boldsymbol{\Sigma}|^{-\frac{T}{2}}$. Then, it follows from (10) and (11) that the integrated likelihood $f(\mathbf{y}\,|\,\boldsymbol{\gamma},\boldsymbol{\Sigma},\boldsymbol{\Omega})$ is given by

$$\begin{aligned} f(\mathbf{y}\,|\,\boldsymbol{\gamma},\boldsymbol{\Sigma},\boldsymbol{\Omega}) &= \int f(\mathbf{y}\,|\,\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\Sigma})p(\boldsymbol{\beta}\,|\,\boldsymbol{\Omega})\mathrm{d}\boldsymbol{\beta} \\ &= c_1 \int \mathrm{e}^{-\frac{1}{2}(\mathbf{y}-\mathbf{W}\boldsymbol{\gamma}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{I}_T\otimes\boldsymbol{\Sigma}^{-1})(\mathbf{y}-\mathbf{W}\boldsymbol{\gamma}-\mathbf{X}\boldsymbol{\beta})}\mathrm{e}^{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\alpha})'\mathbf{H}'\mathbf{S}^{-1}\mathbf{H}(\boldsymbol{\beta}-\boldsymbol{\alpha})}\mathrm{d}\boldsymbol{\beta} \\ &= c_1 \int \mathrm{e}^{-\frac{1}{2}\left(\boldsymbol{\beta}'\mathbf{K}_{\boldsymbol{\beta}}\boldsymbol{\beta}-2\boldsymbol{\beta}'\mathbf{d}_{\boldsymbol{\beta}}+(\mathbf{y}-\mathbf{W}\boldsymbol{\gamma})'(\mathbf{I}_T\otimes\boldsymbol{\Sigma}^{-1})(\mathbf{y}-\mathbf{W}\boldsymbol{\gamma})+\boldsymbol{\alpha}'\mathbf{H}'\mathbf{S}^{-1}\mathbf{H}\boldsymbol{\alpha}\right)}\mathrm{d}\boldsymbol{\beta}, \end{aligned}$$

where $\mathbf{K}_{\boldsymbol{\beta}} = \mathbf{X}'(\mathbf{I}_T\otimes\boldsymbol{\Sigma}^{-1})\mathbf{X} + \mathbf{H}'\mathbf{S}^{-1}\mathbf{H}$ and $\mathbf{d}_{\boldsymbol{\beta}} = \mathbf{X}'(\mathbf{I}_T\otimes\boldsymbol{\Sigma}^{-1})(\mathbf{y}-\mathbf{W}\boldsymbol{\gamma}) + \mathbf{H}'\mathbf{S}^{-1}\mathbf{H}\boldsymbol{\alpha}$. Now, complete the square:

$$\begin{aligned} \boldsymbol{\beta}'\mathbf{K}_{\boldsymbol{\beta}}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{d}_{\boldsymbol{\beta}} &= \boldsymbol{\beta}'\mathbf{K}_{\boldsymbol{\beta}}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{d}_{\boldsymbol{\beta}} + \mathbf{d}_{\boldsymbol{\beta}}'\mathbf{K}_{\boldsymbol{\beta}}^{-1}\mathbf{d}_{\boldsymbol{\beta}} - \mathbf{d}_{\boldsymbol{\beta}}'\mathbf{K}_{\boldsymbol{\beta}}^{-1}\mathbf{d}_{\boldsymbol{\beta}} \\ &= (\boldsymbol{\beta}-\mathbf{K}_{\boldsymbol{\beta}}^{-1}\mathbf{d}_{\boldsymbol{\beta}})'\mathbf{K}_{\boldsymbol{\beta}}(\boldsymbol{\beta}-\mathbf{K}_{\boldsymbol{\beta}}^{-1}\mathbf{d}_{\boldsymbol{\beta}}) - \mathbf{d}_{\boldsymbol{\beta}}'\mathbf{K}_{\boldsymbol{\beta}}^{-1}\mathbf{d}_{\boldsymbol{\beta}}. \end{aligned}$$

Hence,

$$\begin{aligned} &f(\mathbf{y}\,|\,\boldsymbol{\gamma},\boldsymbol{\Sigma},\boldsymbol{\Omega}) \\ &= c_1\,\mathrm{e}^{-\frac{1}{2}\left[(\mathbf{y}-\mathbf{W}\boldsymbol{\gamma})'(\mathbf{I}_T\otimes\boldsymbol{\Sigma}^{-1})(\mathbf{y}-\mathbf{W}\boldsymbol{\gamma})+\boldsymbol{\alpha}'\mathbf{H}'\mathbf{S}^{-1}\mathbf{H}\boldsymbol{\alpha}-\mathbf{d}_{\boldsymbol{\beta}}'\mathbf{K}_{\boldsymbol{\beta}}^{-1}\mathbf{d}_{\boldsymbol{\beta}}\right]}\int \mathrm{e}^{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{K}_{\boldsymbol{\beta}}^{-1}\mathbf{d}_{\boldsymbol{\beta}})'\mathbf{K}_{\boldsymbol{\beta}}(\boldsymbol{\beta}-\mathbf{K}_{\boldsymbol{\beta}}^{-1}\mathbf{d}_{\boldsymbol{\beta}})}\mathrm{d}\boldsymbol{\beta} \\ &= c_1(2\pi)^{\frac{Tq}{2}}|\mathbf{K}_{\boldsymbol{\beta}}|^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}\left[(\mathbf{y}-\mathbf{W}\boldsymbol{\gamma})'(\mathbf{I}_T\otimes\boldsymbol{\Sigma}^{-1})(\mathbf{y}-\mathbf{W}\boldsymbol{\gamma})+\boldsymbol{\alpha}'\mathbf{H}'\mathbf{S}^{-1}\mathbf{H}\boldsymbol{\alpha}-\mathbf{d}_{\boldsymbol{\beta}}'\mathbf{K}_{\boldsymbol{\beta}}^{-1}\mathbf{d}_{\boldsymbol{\beta}}\right]} \\ &= (2\pi)^{-\frac{Tn}{2}}|\mathbf{Q}_0|^{-\frac{1}{2}}|\boldsymbol{\Omega}|^{-\frac{T-1}{2}}|\boldsymbol{\Sigma}|^{-\frac{T}{2}}|\mathbf{K}_{\boldsymbol{\beta}}|^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}\left[(\mathbf{y}-\mathbf{W}\boldsymbol{\gamma})'(\mathbf{I}_T\otimes\boldsymbol{\Sigma}^{-1})(\mathbf{y}-\mathbf{W}\boldsymbol{\gamma})+\boldsymbol{\alpha}'\mathbf{H}'\mathbf{S}^{-1}\mathbf{H}\boldsymbol{\alpha}-\mathbf{d}_{\boldsymbol{\beta}}'\mathbf{K}_{\boldsymbol{\beta}}^{-1}\mathbf{d}_{\boldsymbol{\beta}}\right]}. \end{aligned}$$

Next, we derive the integrated likelihood of the semiparametric model in (14)–(15). Let $c_2 = (2\pi)^{-n}(\sigma^2)^{-\frac{n}{2}}|\boldsymbol{\Omega}_\tau|^{-\frac{1}{2}}|\mathbf{G}|$, $\mathbf{K}_{\boldsymbol{\theta}} = \mathbf{D}'\mathbf{D}/\sigma^2 + \mathbf{G}'\boldsymbol{\Omega}_\tau^{-1}\mathbf{G}$ and $\mathbf{d}_{\boldsymbol{\theta}} = \mathbf{D}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/\sigma^2$. Moreover, by completing the square we have

$$\boldsymbol{\theta}'\mathbf{K}_{\boldsymbol{\theta}}\boldsymbol{\theta} - 2\boldsymbol{\theta}'\mathbf{d}_{\boldsymbol{\theta}} = \left(\boldsymbol{\theta}-\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{d}_{\boldsymbol{\theta}}\right)'\mathbf{K}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}-\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{d}_{\boldsymbol{\theta}}\right) - \mathbf{d}_{\boldsymbol{\theta}}'\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{d}_{\boldsymbol{\theta}}.$$

Then the integrated likelihood is given by:

$$
\begin{aligned}
f(\mathbf{y}\,|\,\boldsymbol{\beta},\sigma^2,\tau) &= \int f(\mathbf{y}\,|\,\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2) f(\boldsymbol{\theta}\,|\,\tau)\mathrm{d}\boldsymbol{\theta} \\
&= c_2 \int \mathrm{e}^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{D}\boldsymbol{\theta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{D}\boldsymbol{\theta})} \mathrm{e}^{-\frac{1}{2}\boldsymbol{\theta}'\mathbf{G}'\boldsymbol{\Omega}_\tau^{-1}\mathbf{G}\boldsymbol{\theta}} \mathrm{d}\boldsymbol{\theta} \\
&= c_2 \int \mathrm{e}^{-\frac{1}{2}\left(\boldsymbol{\theta}'\mathbf{K}_{\boldsymbol{\theta}}\boldsymbol{\theta} - 2\boldsymbol{\theta}'\mathbf{d}_{\boldsymbol{\theta}} + \frac{1}{\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right)} \mathrm{d}\boldsymbol{\theta}, \\
&= c_2 \mathrm{e}^{-\frac{1}{2}\left(\frac{1}{\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \mathbf{d}_{\boldsymbol{\theta}}'\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{d}_{\boldsymbol{\theta}}\right)} \int \mathrm{e}^{-\frac{1}{2}\left(\boldsymbol{\theta}-\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{d}_{\boldsymbol{\theta}}\right)\mathbf{K}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}-\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{d}_{\boldsymbol{\theta}}\right)} \mathrm{d}\boldsymbol{\theta} \\
&= c_2 (2\pi)^{\frac{n}{2}} |\mathbf{K}_{\boldsymbol{\theta}}|^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}\left(\frac{1}{\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \mathbf{d}_{\boldsymbol{\theta}}'\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{d}_{\boldsymbol{\theta}}\right)} \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} |\boldsymbol{\Omega}_\tau|^{-\frac{1}{2}} |\mathbf{G}||\mathbf{K}_{\boldsymbol{\theta}}|^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}\left(\frac{1}{\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \mathbf{d}_{\boldsymbol{\theta}}'\mathbf{K}_{\boldsymbol{\theta}}^{-1}\mathbf{d}_{\boldsymbol{\theta}}\right)}.
\end{aligned}
$$

Appendix B: Gibbs Samplers

This appendix provides the estimation details of the models used in the empirical applications. Gibbs samplers are used to simulate from the posterior distributions of all models in the empirical applications. All the full conditional distributions of each Gibbs sampler are provided below.

The first is the Gibbs sampler for the factor models:

1. Draw from $(\mathbf{f}\,|\,\mathbf{y},\boldsymbol{\beta},\mathbf{A},\boldsymbol{\Sigma},\boldsymbol{\Omega}) \sim \mathcal{N}(\widehat{\mathbf{f}},\mathbf{D}_{\mathbf{f}})$ using the precision sampler in Chan and Jeliazkov (2009), where

$$\mathbf{D}_{\mathbf{f}}^{-1} = \mathbf{I}_T \otimes \boldsymbol{\Omega}^{-1} + \mathbf{I}_T \otimes (\mathbf{A}'\boldsymbol{\Sigma}^{-1}\mathbf{A}), \quad \widehat{\mathbf{f}} = \mathbf{D}_{\mathbf{f}}\left((\mathbf{I}_T \otimes (\mathbf{A}'\boldsymbol{\Sigma}^{-1}))(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right).$$

2. Draw from $(\mathbf{a}_i,\boldsymbol{\beta}_i\,|\,\mathbf{y},\mathbf{f},\boldsymbol{\Sigma},\boldsymbol{\Omega}) \sim \mathcal{N}(\widehat{\boldsymbol{\alpha}}_i,\mathbf{D}_{\boldsymbol{\alpha}_i})$, where $\widehat{\boldsymbol{\alpha}}_i$ and $\mathbf{D}_{\boldsymbol{\alpha}_i}$ depend on whether $i \leqslant k$ or $i > k$. For $i \leqslant k$,

$$\mathbf{D}_{\boldsymbol{\alpha}_i}^{-1} = \mathbf{V}_{\boldsymbol{\alpha}_i}^{-1} + \frac{1}{\sigma_i^2}\sum_{t=1}^{T}\mathbf{z}_{it}'\mathbf{z}_{it}, \quad \widehat{\boldsymbol{\alpha}}_i = \mathbf{D}_{\boldsymbol{\alpha}_i}\left(\mathbf{V}_{\boldsymbol{\alpha}_i}^{-1}\boldsymbol{\alpha}_{0i} + \frac{1}{\sigma_i^2}\sum_{t=1}^{T}\mathbf{z}_{it}'(y_{it}-f_{it})\right),$$

where $\mathbf{z}_{it} = (\mathbf{X}_{it}', f_{1t},\ldots,f_{i-1,t})'$, $\boldsymbol{\alpha}_{0i} = (\boldsymbol{\beta}_{0i}',\mathbf{a}_{0i}')'$ and $\mathbf{V}_{\boldsymbol{\alpha}_i} = \mathrm{diag}(\mathbf{V}_{\boldsymbol{\beta}_i},\mathbf{V}_{\mathbf{a}_i})$. For $i > k$,

$$\mathbf{D}_{\boldsymbol{\alpha}_i}^{-1} = \mathbf{V}_{\boldsymbol{\alpha}_i}^{-1} + \frac{1}{\sigma_i^2}\sum_{t=1}^{T}\mathbf{z}_{it}'\mathbf{z}_{it}, \quad \widehat{\boldsymbol{\alpha}}_i = \mathbf{D}_{\boldsymbol{\alpha}_i}\left(\mathbf{V}_{\boldsymbol{\alpha}_i}^{-1}\boldsymbol{\alpha}_{0i} + \frac{1}{\sigma_i^2}\sum_{t=1}^{T}\mathbf{z}_{it}'y_{it}\right),$$

where $\mathbf{z}_{it} = (\mathbf{X}_{it}', f_{1t},\ldots,f_{kt})'$.

3. Draw from $(\sigma_i^2\,|\,\mathbf{y},\mathbf{f},\mathbf{A},\boldsymbol{\beta},\boldsymbol{\Omega}) \sim \mathcal{IG}\left(\nu_{\sigma_i^2} + \frac{T}{2}, S_{\sigma_i^2} + \frac{1}{2}\sum_{t=1}^{T}s_{it}^2\right)$, where s_{it} is the i-th element of $\mathbf{s}_t = \mathbf{y}_t - \mathbf{X}_t\boldsymbol{\beta} - \mathbf{A}\mathbf{f}_t$.

4. Draw from $(\omega_j^2 \,|\, \mathbf{y}, \mathbf{f}, \mathbf{A}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \sim \mathcal{IG}\left(\nu_{\omega_j^2} + \frac{T}{2}, S_{\omega_j^2} + \frac{1}{2}\sum_{t=1}^{T} f_{jt}^2\right)$, where f_{jt} is the j-th element of $\mathbf{f}_t$.

5. Repeat Steps (1)-(4) N times.

Next, we present the Gibbs sampler for the state space model in (7)–(8). Both the constant coefficients VAR and TVP-VAR are special cases of this model.

1. Draw from $(\boldsymbol{\gamma} \,|\, \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\Omega}) \sim \mathcal{N}(\widehat{\boldsymbol{\gamma}}, \mathbf{D}_{\boldsymbol{\gamma}})$, where
$$\mathbf{D}_{\boldsymbol{\gamma}}^{-1} = \mathbf{W}'(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{W} +' \mathbf{V}_{\boldsymbol{\gamma}}^{-1}, \quad \widehat{\boldsymbol{\gamma}} = \mathbf{D}_{\boldsymbol{\gamma}}\left(\mathbf{W}'(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \mathbf{V}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{\gamma}_0\right).$$

2. Draw from $(\boldsymbol{\beta} \,|\, \mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}, \boldsymbol{\Omega}) \sim \mathcal{N}(\widehat{\boldsymbol{\beta}}, \mathbf{D}_{\boldsymbol{\beta}})$, using the precision sampler in Chan and Jeliazkov (2009), where
$$\mathbf{D}_{\boldsymbol{\beta}}^{-1} = \mathbf{X}'(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})\mathbf{X} + \mathbf{H}'\mathbf{S}^{-1}\mathbf{H}, \quad \widehat{\boldsymbol{\beta}} = \mathbf{D}_{\boldsymbol{\beta}}\left(\mathbf{X}'(\mathbf{I}_T \otimes \boldsymbol{\Sigma}^{-1})(\mathbf{y} - \mathbf{W}\boldsymbol{\gamma}) + \mathbf{H}'\mathbf{S}^{-1}\mathbf{H}\boldsymbol{\alpha}\right).$$

3. Draw from
$$(\boldsymbol{\Sigma} \,|\, \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Omega}) \sim \mathcal{IW}\left(T + \nu_{\boldsymbol{\Sigma}}, \mathbf{S}_{\boldsymbol{\Sigma}} + \sum_{t=1}^{T}(\mathbf{y}_t - \mathbf{W}_t\boldsymbol{\gamma} - \mathbf{X}_t\boldsymbol{\beta}_t)(\mathbf{y}_t - \mathbf{W}_t\boldsymbol{\gamma} - \mathbf{X}_t\boldsymbol{\beta}_t)'\right).$$

4. Draw from $(\omega_i^2 \,|\, \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}) \sim \mathcal{IG}\left(\nu_i + \frac{T-1}{2}, S_i + \frac{1}{2}\sum_{t=2}^{T}(\beta_{it} - \beta_{i,t-1})^2\right)$.

5. Repeat Steps (1)-(4) N times.

Finally, the following is the Gibbs sampler for the semiparametric regression:

1. Draw from $(\boldsymbol{\theta} \,|\, \mathbf{y}, \boldsymbol{\beta}, \sigma^2, \tau) \sim \mathcal{N}(\widehat{\boldsymbol{\theta}}, \mathbf{D}_{\boldsymbol{\theta}})$, using the precision sampler in Chan and Jeliazkov (2009), where
$$\mathbf{D}_{\boldsymbol{\theta}}^{-1} = \frac{1}{\sigma^2}\mathbf{D}'\mathbf{D} + \mathbf{G}'\boldsymbol{\Omega}_{\tau}^{-1}\mathbf{G}, \quad \widehat{\boldsymbol{\theta}} = \mathbf{D}_{\boldsymbol{\theta}}\left(\frac{1}{\sigma^2}\mathbf{D}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right).$$

2. Draw from $(\boldsymbol{\beta} \,|\, \mathbf{y}, \boldsymbol{\theta}, \sigma^2, \tau) \sim \mathcal{N}(\widehat{\boldsymbol{\beta}}, \mathbf{D}_{\boldsymbol{\beta}})$, where
$$\mathbf{D}_{\boldsymbol{\beta}}^{-1} = \frac{1}{\sigma^2}\mathbf{X}'\mathbf{X} + \mathbf{V}_{\boldsymbol{\beta}}^{-1}, \quad \widehat{\boldsymbol{\beta}} = \mathbf{D}_{\boldsymbol{\beta}}\left(\frac{1}{\sigma^2}\mathbf{X}'(\mathbf{y} - \mathbf{D}\boldsymbol{\theta}) + \mathbf{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}_0\right).$$

3. Draw from
$$(\sigma^2 \,|\, \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\theta}, \tau) \sim \mathcal{IG}\left(\nu_{\sigma^2} + \frac{n}{2}, S_{\sigma^2} + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{D}\boldsymbol{\theta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{D}\boldsymbol{\theta})\right).$$

4. Draw from $(\tau \,|\, \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) \sim \mathcal{IG}\left(\nu_\tau + \frac{m-2}{2}, S_\tau + \frac{1}{2}\sum_{i=3}^{m}\eta_i^2\right)$, where η_i is the i-th element of $\boldsymbol{\eta} = \mathbf{G}\boldsymbol{\theta}$.

5. Repeat Steps (1)-(4) N times.

References

C. A. Abanto-Valle, D. Bandyopadhyay, V. H. Lachos, and I. Enriquez. Robust Bayesian analysis of heavy-tailed stochastic volatility models using scale mixtures of normal distributions. *Computational Statistics and Data Analysis*, 54(12):2883–2898, 2010.

L. Bauwens and J. V. K Rombouts. On marginal likelihood computation in change-point models. *Computational Statistics and Data Analysis*, 56(11):3415–3429, 2012.

M. Belmonte, G. Koop, and D. Korobilis. Hierarchical shrinkage in time-varying coefficients models. *Journal of Forecasting*, 2014. Forthcoming.

A. Berg, R. Meyer, and J. Yu. Deviance information criterion for comparing stochastic volatility models. *Journal of Business and Economic Statistics*, 22(1):107–120, 2004.

F. Canova. Modelling and forecasting exchange rates with a Bayesian time-varying coefficient model. *Journal of Economic Dynamics and Control*, 17:233–261, 1993.

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

G. Celeux, F. Forbes, C. P. Robert, and D. M. Titterington. Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4):651–674, 2006.

J. C. C. Chan and E. Eisenstat. Marginal likelihood estimation with the Cross-Entropy method. *Econometric Reviews*, 2014. Forthcoming.

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

J. C. C. Chan, G. Koop, and S. M. Potter. A new model of trend inflation. *Journal of Business and Economic Statistics*, 31(1):94–106, 2013.

X. Chen, A. Kontonikas, and A. Montagnoli. Asset prices, credit and the business cycle. *Economics Letters*, 117:857–861, 2012.

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

S. Chib, E. Greenberg, and I. Jeliazkov. Estimation of semiparametric models in the presence of endogeneity and sample selection. *Journal of Computational and Graphical Statistics*, 18:321–348, 2009.

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

E. Eisenstat, J. C. C. Chan, and R. W. Strachan. Stochastic model specification search for time-varying parameter VARs. *Econometric Reviews*, 2014. Forthcoming.

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

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

J. Geweke and G. Zhou. Measuring the pricing error of the arbitrage pricing theory. *The Review of Financial Studies*, 9:557–587, 1996.

B. Kline and J. L. Tobias. The wages of BMI: Bayesian analysis of a skewed treatment-response model with nonparametric endogeneity. *Journal of Applied Econometrics*, 23 (6):767–793, 2008.

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

G. Koop and D. Korobilis. Forecasting inflation using dynamic model averaging. *International Economic Review*, 53(3):867–886, 2012.

G. Koop and D. J. Poirier. Bayesian variants of some classical semiparametric regression techniques. *Journal of Econometrics*, 123(2):259–282, 2004.

G. Koop, D. J. Poirier, and J. Tobias. Semiparametric Bayesian inference in multiple equation models. *Journal of Applied Econometrics*, 20(6):723–747, 2005.

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

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

Y. Li, T. Zeng, and J. Yu. Robust deviance information criterion for latent variable models. *SMU Economics and Statistics Working Paper Series*, 2012.

H. F. Lopes and E. Salazar. Bayesian model uncertainty in smooth transition autoregressions. *Journal of Time Series Analysis*, 27(1):99–117, 2006.

W. J. McCausland. The HESSIAN method: Highly efficient simulation smoothing, in a nutshell. *Journal of Econometrics*, 168(2):189–206, 2012.

W. J. McCausland, S. Miller, and D. Pelletier. Simulation smoothing for state-space models: A computational efficiency analysis. *Computational Statistics and Data Analysis*, 55:199–212, 2011.

R. B. Millar. Comparison of hierarchical Bayesian models for overdispersed count data using DIC and Bayes factors. *Biometrics*, 65(3):962–969, 2009.

H. Mumtaz and P. Surico. Evolving international inflation dynamics: World and country-specific factors. *Journal of the European Economic Association*, 10(4):716–734, 2012.

F. Nardari and J. T. Scruggs. Bayesian analysis of linear factor models with latent factors, multivariate stochastic volatility, and APT pricing restrictions. *Journal of Financial and Quantitative Analysis*, 42(4):857–892, 2007.

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

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B*, 64(4): 583–639, 2002.

J. J. J. Wang, S. T. B. Choy, and J. S. K. Chan. Modelling stochastic volatility using generalized t distribution. *Journal of Statistical Computation and Simulation*, 83(2): 340–354, 2013.

E. Ward. A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. *Ecological Modelling*, 211(1–2):1–10, 2008.

N. Xiao, J. Zarnikau, and P. Damien. Testing functional forms in energy modelling: An application of the Bayesian approach to US electricity demand. *Energy Economics*, 29 (2):158–166, 2007.