ELSEVIER Computational Statistics & Data Analysis 29 (1999) 387–410

COMPUTATIONAL STATISTICS & DATA ANALYSIS

Hyperparameter estimation in forecast models

Hedibert Freitas Lopes [a,b,*], Ajax R. Bello Moreira [c], Alexandra Mello Schmidt [c]

[a] *Instituto de Matemática, Universidade Federal do Rio de Janeiro, Brazil*
[b] *Institute of Statistics and Decision Sciences, Duke University, USA*
[c] *Instituto de Pesquisa Econômica Aplicada, Brazil*

Received 1 June 1998; received in revised form 1 September 1998; accepted 1 October 1998

Abstract

A large number of non-linear time series models can be more easily analyzed using traditional linear methods by considering explicitly the difference between parameters of interest, or just parameters, and hyperparameters. One example is the class of conditionally Gaussian dynamic linear models. Bayesian vector autoregressive models and non-linear transfer function models are also important examples in the literature. Until recently, a two-step procedure was broadly used to estimate such models. In the first step maximum likelihood estimation was used to find the best value of the hyperparameter, which turned to be used in the second step where a conditionally linear model was fitted. The main drawback of such an algorithm is that it does not take into account any kind of uncertainty that might have been brought, and usually was, to the modeling at the first step. In other words and more practically speaking, the variances of the parameters are underestimated. Another problem, more philosophical, is the violation of the *likelihood principle* by using the sample information twice. In this paper we apply sampling importance resampling (SIR) techniques (Rubin, 1988) to obtain a numerical approximation to the full posterior distribution of the hyperparameters. Then, instead of conditioning in a particular value of that distribution we integrate the hyperparameters out in order to obtain the marginal posterior distributions of the parameters. We used SIR to model a set of Brazilian macroeconomic time-series in three different, but important, contexts. We also compare the forecast performance of our approach with traditional ones.

Keywords: Hyperparameter; Posterior distribution; Sampling importance resampling (SIR); Litterman's prior; Dynamic modeling; Bayes factor

* Correspondence address. Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251, USA. Tel.: +1 919 684 8088; e-mail: hedibert@stat.duke.edu.

0167-9473/99/$ – see front matter
PII: S0167-9473(98)00078-4

1. Introduction

We will consider throughout the paper the following class of dynamic models:

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{F}_t'(\boldsymbol{\lambda})\boldsymbol{\theta}_t + \boldsymbol{v}_t, \quad \boldsymbol{v}_t \sim N(0, \boldsymbol{V}_t(\boldsymbol{\lambda})), \\ \boldsymbol{\theta}_t &= \boldsymbol{G}_t(\boldsymbol{\lambda})\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N(0, \boldsymbol{W}_t(\boldsymbol{\lambda})), \\ \boldsymbol{\lambda} &\sim p(\boldsymbol{\lambda}), \end{aligned} \tag{1}$$

where $\boldsymbol{y}$ is a vector of time-series and $\boldsymbol{\theta}$ is the state vector (matrix), or simply *parameters*, and $\boldsymbol{\lambda}$ is the vector (matrix) of *hyperparameters*. Therefore, conditionally on $\boldsymbol{\lambda}$, model (1) is usually represented by the quadruple $\{\boldsymbol{F}, \boldsymbol{G}, \boldsymbol{V}, \boldsymbol{W}\}_t$. Further details concerning dynamic linear models (DLMs) and Bayesian hierarchical models can be found in West and Harrison (1997) and Gamerman and Migon (1998).

The distinction between hyperparameters, $\boldsymbol{\lambda}$, and parameters $\boldsymbol{\theta}$, has been widely made in the theoretical and practical forecasting model literature in order to facilitate the modeling from an analytical and, even more effectively, computational viewpoint. For example, West and Harrison (1997) use hyperparameters to define the evolution variances, $\boldsymbol{W}_t(\boldsymbol{\lambda})$ through discount factors. Harvey (1989) follows a similar approach from a classical perspective. Finally, hyperparameters have also special attention in Bayesian vector autoregressive (BVAR) models through priors for the autoregressive coefficients, $\boldsymbol{W}_0(\boldsymbol{\lambda})$ in model (1) without parameter evolution (Litterman, 1986).

In cases where the posterior distribution of the hyperparameters is unimodal and highly concentrated around its maximum the marginal posterior for the parameter of interest can be well approximated by the conditional posterior given the value of the hyperparameter that maximizes the predictive likelihood (Bernardo and Smith, 1994). In realistic situations, however, such distribution does not have an unique mode, or it is not well concentrated around its maximum value. Therefore, it is a problem-specific issue whether or not and under which circumstances the MLE used in the *empirical Bayes* can replace the estimator obtained from the sampling importance resampling (SIR) techniques.

In the literature many authors, like Carter and Kohn (1994) and Frühwirth-Schnatter (1994) have used Markov chain Monte Carlo (MCMC) methods to obtain the posterior distributions of the parameters and hyperparameters. Nevertheless, the number of unknown quantities can be quite large resulting in various computational problems. The approach we are proposing is divided in three simple and non-iterative steps. In the first step, DLMs are proposed conditionally on a set of hyperparameters values drawn from the prior. In the second step the SIR techniques are used to obtain the posterior distributions of the hyperparameters. Finally, in the third step we integrate out numerically the hyperparameters in order to obtain the posterior distribution of the parameters.

Our approach is more effective in models where the number of hyperparameters is much smaller than the number of parameters. Fortunately, in important econometric problems, as will be seen later on in this paper, the number of hyperparameters is considerably small, ranging from one to five in general. Unconditional models

obtained from SIR techniques are compared to conditional ones obtained by conditioning on maximum likelihood estimator (MLE) or posterior modes.

Our approach has been applied to various trade balance forecast models from the Brazilian economy. The macroeconomics variables used are export, non-oil imports, gross national product (GNP) and real exchange rate taken from the Brazilian economy and available upon request.

We have analyzed a BVAR model using univariate and multivariate estimation. Therefore, the hyperparameters are the coefficients of the *Litterman's prior*. In the multivariate case we have used the fact that a VAR model can be seen as a DLM of common components (DLMCC) as described in Quintana (1985), Barbosa (1989) and West and Harrison (1997).[1] Secondly, a second-order DLM with seasonal components is considered, where the hyperparameters in this case are the discount factors of the trend and the seasonal components. Finally, we have explored a DLM with transfer function. For this model we allow the exchange rate effect on the exports (imports) to occur as a saturation effect. In doing so, the hyperparameter is the quantity measuring the velocity of saturation, which turns out to be the coefficient of a first-order transfer function.

In the first application we find that the effect of Litterman's prior on the forecasts decreases as the sample size increases, as asymptotic theory insures. For this reason it is an empirical matter to verify the robustness of the model to the choice of priors in models which have small samples. In order to make this evaluation a procedure which makes use of the cumulative Bayes factor (CBF) is proposed. It compares the model estimated by Litterman's prior with one which uses a non-informative prior. By using this scheme it was possible to evaluate, through the sample, the Litterman's prior effect on the BVAR models mentioned above. This follows closely what has been suggested by Geweke (1994) to compare the forecast performance of econometric models.

The rest of the paper is developed in the following way. Section 2 describes how sampling importance re-sampling methods can be used to approximate the posterior of the hyperparameter, and ultimately the posterior distribution of the parameters. In Section 3, a simulated exercise is presented to investigate the sensitivity of the posterior distribution to the prior sample size. Section 4 presents the results obtained from the BVAR model, the decomposition model and the transfer function model. Final conclusions are presented in Section 5.

2. Methodology

2.1. SIR techniques in conditionally Gaussian DLM

Recalling, the class of models we considered is represented by Eq. (1), where the first hierarchical level corresponds to a DLM as described in Appendix A. Particularly

[1] These models were estimated as static models that are numerically equal to dynamic models without stochastic components in the transition equation.

in BVAR models, the second hierarchical level corresponds to the Litterman's prior which is defined at the very first period ($t=0$). For other traditional DLM common hyperparameters are discount factors, saturation parameters of the transfer function, among others. Model 1 sheds some light to the fact that the hyperparameters are defined as constants through the sampling period.[2]

According to the SIR techniques (Rubin, 1988; Smith and Gelfand, 1992), samples from the posterior distribution of $\boldsymbol{\lambda}$, $p(\boldsymbol{\lambda} \mid D_T)$, can be obtained by sampling from $p(\boldsymbol{\lambda})$ and then computing for each sampled value $\boldsymbol{\lambda}_i$ its likelihood marginal, $L(\boldsymbol{\lambda}_i; D_T)$, where $D_T = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T, D_0\}$ is the observed data up to time T, and D_0 is the set of prior information. Schematically, the method is as follows:

Step 1: Sample $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_n$ from $p(\boldsymbol{\lambda})$;
Step 2: Sample $\boldsymbol{\lambda}^*$ from the discrete distribution function on $\{\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_n\}$, where

$$P(\boldsymbol{\lambda}^* = \boldsymbol{\lambda}_j) = q_j = \frac{L(\boldsymbol{\lambda}_j; \boldsymbol{y})}{\sum_{k=1}^{n} L(\boldsymbol{\lambda}_k; \boldsymbol{y})}, \quad j = 1, 2, \ldots, n. \tag{2}$$

Step 3: Repeat Eq. (2) M times.

It can be proved that the $\boldsymbol{\lambda}^*$'s form a sample from a distribution that approximates the posterior $p(\boldsymbol{\lambda} \mid D_T)$ as n increases.

The marginal likelihood $L(\boldsymbol{\lambda}; D_T)$ corresponds to the predictive density function of $\boldsymbol{y}$ conditional on $\boldsymbol{\lambda}$, which in the DLM context has the following analytical form (see Appendix A for the details):

$$\begin{aligned} L(\boldsymbol{\lambda}; D_T) &= L(\boldsymbol{\lambda}; \boldsymbol{y}, D_0) = p(\boldsymbol{y} \mid \boldsymbol{\lambda}, D_0) \\ &= \int p(\boldsymbol{y} \mid \boldsymbol{\theta}, D_0)\, p(\boldsymbol{\theta} \mid \boldsymbol{\lambda})\, \mathrm{d}\boldsymbol{\theta} = E_{\theta \mid \lambda}[L(\boldsymbol{\theta}; \boldsymbol{y} \mid D_0)]. \end{aligned}$$

Therefore, the posterior distribution of the hyperparameter is given by the following expression:

$$p(\boldsymbol{\lambda} \mid D_T) = \frac{L(\boldsymbol{\lambda}; D_T)\, p(\boldsymbol{\lambda})}{\int L(\boldsymbol{\lambda}; D_T)\, p(\boldsymbol{\lambda})\, \mathrm{d}\boldsymbol{\lambda}}. \tag{3}$$

If the prior for $\boldsymbol{\lambda}$ is constant then Eq. (5) is just the normalized likelihood. After obtaining the posterior distribution of $\boldsymbol{\lambda}$, it is straightforward to make posterior

[2] A simple example, but quite complex to be analytically considered is as follows:

$$(y_t \mid \theta, y_{t-1}) \sim \mathrm{N}(\theta y_{t-1}, \sigma^2),$$
$$(\theta, \sigma^2) \sim \mathrm{Normal} - IG(0, \sigma^2 \lambda, \alpha, \beta),$$
$$(\lambda) \sim \mathrm{Beta}(\zeta, \xi)$$

for known $\alpha, \mathrm{beta}, \zeta$ and ξ. A similar example can be found in Berger (1985), where the essential part of the posterior distribution for λ can be analytically computed. However, for more general cases it is necessary to make use of numerical methods to obtain summaries of the posterior distribution.

inference concerning any function g, of $\boldsymbol{\theta}$:

$$\begin{aligned} E(g(\boldsymbol{\theta}) \mid D_T) &= \int_{\boldsymbol{\theta}} g(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D_T) \, \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\lambda}} g(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D_T, \boldsymbol{\lambda}) p(\boldsymbol{\lambda} \mid D_T) \, \mathrm{d}\boldsymbol{\lambda} \, \mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\lambda}} \left\{ \int_{\boldsymbol{\theta}} g(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid D_T, \boldsymbol{\lambda}) \, \mathrm{d}\boldsymbol{\theta} \right\} p(\boldsymbol{\lambda} \mid D_T) \, \mathrm{d}\boldsymbol{\lambda} \\ &= \int_{\boldsymbol{\lambda}} E(g(\boldsymbol{\theta}) \mid \boldsymbol{\lambda}, D_T) p(\boldsymbol{\lambda} \mid D_T) \, \mathrm{d}\boldsymbol{\lambda} \end{aligned}$$

which can be approximated by re-sampling $\{\boldsymbol{\lambda}_i\}$ from $p(\boldsymbol{\lambda} \mid D_T)$ to obtain [3]

$$E(g(\boldsymbol{\theta}) \mid D_T) \cong \sum_{i=1}^{n} E(g(\boldsymbol{\theta}) \mid \boldsymbol{\lambda}_i, D_T). \tag{4}$$

For g equal to, for instance, the identity function or a future value, $\boldsymbol{y}_{t+1}$, the expression in Eq. (4) approximates $E(\theta \mid D_t)$ and $E(\boldsymbol{y}_{t+1} \mid D_T)$, respectively. The same idea can be used to make long-run forecasts when the regressors variables are not stochastic.

2.1.1. Adaptation to VAR models

In the VAR models, however, the regressors are stochastic and a recursive procedure is used. [4] For example in a VAR model of order p, VAR(p), the forecast for more than one period ahead can be shown to be (conditional on $\boldsymbol{\lambda}$):

$$\hat{\boldsymbol{y}}_{t+h}(\boldsymbol{\lambda}) = \sum_{k=1}^{p} E(\boldsymbol{\theta}_{t,k} \mid D_T, \boldsymbol{\lambda}) \hat{\boldsymbol{y}}_{t+h-k}(\boldsymbol{\lambda}), \tag{5}$$

where $E(\boldsymbol{\theta}_{t,k} \mid D_T, \boldsymbol{\lambda})$ is a matrix of parameters estimated at period t related to lagged k. The expected value of the forecast unconditional on $\boldsymbol{\lambda}$, could be given by

$$E[\hat{\boldsymbol{y}}_{t+h} \mid D_T] \cong \sum_{i}^{M} \hat{\boldsymbol{y}}_{t+h}(\boldsymbol{\lambda}_i), \quad \boldsymbol{\lambda}_i \sim p(\boldsymbol{\lambda} \mid D_T). \tag{6}$$

In order to obtain the predictive performance of the model we have considered the following conditional forecast:

$$\tilde{\boldsymbol{y}}_{t+1}(\boldsymbol{\lambda}) = E(\boldsymbol{y}_{t+1} \mid D_t, \boldsymbol{\lambda}), \tag{7}$$

$$\tilde{\boldsymbol{y}}_{t+h}(\boldsymbol{\lambda}) = \sum_{k=1}^{p} E(\boldsymbol{\theta}_{t,k} \mid D_t, \boldsymbol{\lambda}) \tilde{\boldsymbol{y}}_{t+h-k}(\boldsymbol{\lambda}) \tag{8}$$

[3] The posterior distribution of the hyperparameters is defined just at the very last period of time, because they were estimated with the use of all data. The parameters of the intermediate periods are defined in the same way $p(\boldsymbol{\theta}_t \mid D_T, \boldsymbol{\lambda})$.

[4] See, for instance, Lopes and Ehlers (1997) for a forecasting algorithm in Bayesian VAR models.

and defined $\tilde{\boldsymbol{e}}_{t+h}(\tilde{\boldsymbol{\lambda}}) = \boldsymbol{y}_{t+h} - \tilde{\boldsymbol{y}}_{t+h}(\boldsymbol{\lambda})$ to be the conditional forecast error. Making use of this errors an approximation to the mean square error (MSE) can be constructed, as in Eq. (9) below, or for any other measure of forecast performance, such as the mean absolute error (MAE) or the Theil-U index. [5]

$$E[\tilde{\boldsymbol{e}}^2_{t+h}] \cong \sum_{i=1}^{M} \tilde{\boldsymbol{e}}^2_{t+h}(\boldsymbol{\lambda}_i), \quad \boldsymbol{\lambda}_i \sim p(\boldsymbol{\lambda} \mid D_T), \tag{9}$$

which is an improper predictive capacity measure because it considers D_t, the relevant set to estimate the forecast in t, and $D_T = \bigcup_t D_t$ to estimate the posterior distribution of the hyperparameters. Despite of this, such statistics were constructed in order to be comparable with those equivalent to the conditional model.

2.2. Relation between SIR and Monte Carlo methods

For each hyperparameter value $\{\boldsymbol{\lambda}_j\}$ sampled from the prior distribution $p(\boldsymbol{\lambda})$ we obtain $L(\boldsymbol{\lambda}_j;)$, and then calculate the weights q_i, from Eq. (2). We use these weights as probabilities of a discrete distribution over $\{\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots\}$ to perform a resampling, which turns out to be an approximate sample from $p(\boldsymbol{\lambda} \mid D_T)$. Therefore, as showed in the previous subsection, the posterior draws of the hyperparameters can be used to forecast, smooth and compute any other statistic related to the marginal posterior distribution of the parameters of interest.

The proportion of the sample from the prior that has significant weight (q_i) (see Eq. (2)) depends on (i) the number of hyperparameters and (ii) the distance between the prior and posterior distributions of the hyperparameters. In many applications the prior is quite far from the posterior distribution, which implies that only a small proportion of the draws from the prior has significant weight. This tends to happen when the number of hyperparameters is too large, in which case the SIR techniques may not be feasible. In our applications, and similar real situations, the number of hyperparameters is relatively small, varying from two to eight. It seems to us that this is one of the main limitations of the SIR techniques.

Such limitation can be seen as an advantage when calculating the posterior of quantities of interest that are computationally demanding. In this case we just have to calculate such quantities for a few proportion of the sample that has significant weight. However, if the quantity of interest needs to be calculated in order to obtain $L(\boldsymbol{\lambda}_j; \boldsymbol{y})$ then, resampling would be unnecessary. In this former situation Monte Carlo with importance sampling (MCIS) could be used with the importance function equal to the prior. More details about MCIS can be found in van Dijk and Kloek (1980, 1985), Efron (1982), and Gamerman (1998).

[5] The Theil-U statistic compares the forecast error $\boldsymbol{e}_t$ with that one obtained from a *naive* model, i.e. $E(\boldsymbol{y}_t \mid D_{t-1}) = \boldsymbol{y}_{t-1}$. The statistic is $TU = [\sum \boldsymbol{e}_t^2 / \sum (\boldsymbol{y}_t - \boldsymbol{y}_{t-1})^2]^{1/2}$. For the marginal model the Theil statistic was calculated but the predictive likelihood was not.

The SIR technique and the MCIS, just mentioned, are comparable asymptotically. In the last section we have seen that Rubin (1988) and Smith and Gelfand (1992) present a result which guarantees that $\sum \boldsymbol{\lambda}_j^*/M$ converges to $E(\boldsymbol{\lambda} \mid D_T)$ as n and M go to infinity. On the other hand, the MCIS method guarantees, for example, that

$$\tilde{\boldsymbol{\lambda}} = \frac{\sum_{i=1}^{n} \boldsymbol{\lambda}_i L(\boldsymbol{\lambda}_i, D_T)}{\sum_{i=1}^{n} L(\boldsymbol{\lambda}_i, D_T)} \xrightarrow{n\to\infty} \frac{\int \boldsymbol{\lambda}\, p(\boldsymbol{y} \mid \boldsymbol{\lambda}, D_0)\, p(\boldsymbol{\lambda})\, \mathrm{d}\boldsymbol{\lambda}}{\int p(\boldsymbol{y} \mid \boldsymbol{\lambda}, D_0)\, p(\boldsymbol{\lambda})\, \mathrm{d}\boldsymbol{\lambda}} = E(\boldsymbol{\lambda} \mid D_T).$$

So, when $n, M \to \infty$, both procedures, SIR and MCIS, produce equivalent results. Nevertheless, in the SIR method the Bayesian structure of estimation is clear, besides one can obtain good approximations even when M is small comparatively to n, for example $M = 0.1n$.

Both methods, SIR and MCIS, with a constant or vague prior, produce a posterior distribution, $p(\boldsymbol{\lambda} \mid D_T)$ with a mode that corresponds to the maximum likelihood estimator ($\hat{\lambda}$). However this estimator is not necessarily normally distributed, so the mode and the mean can be very different. Also the second order moments, which are locally measured in the maximum point by the Hessian, can be different from the global measure $V(\lambda \mid D_T)$.[6] Besides, the reparametrization needed for the optimization algorithm makes difficult the estimation of the variance of the hyperparameter.

For the examples shown here the maximum likelihood estimators were calculated using the minimization algorithm of Davidon–Flecher–Powell (Brian, 1984), which makes use of the gradient of the likelihood function, numerically calculated, to obtain the Hessian in each iteration. This algorithm is a generalization of the Newton–Raphson method, which makes use of conjugated directions in order to make the minimization faster. As some hyperparameters are defined just in an interval of the real line, reparametrization of these hyperparameters were adopted, which allow the use of algorithms which have no restrictions to solve problems for that kind of restriction.[7]

2.3. Bayes factor

An old result in Bayesian inference tells us that the effect of the prior is asymptotically negligible, so Litterman's prior would not be a concern and the BVAR model would simply be reduced to a VAR model, computationally speaking.

Nevertheless, since samples are finite in real applications it would be interesting to find out at which period in time the prior information is not significant anymore. In order to verify this, the cumulative Bayes factor (CBF) will be used as defined in West and Harrison (1997). The CBF allows us to compare two models through their

[6] These two measures are nearly the same when the distribution's concavity is approximately constant.

[7] In the case of Litterman's coefficient, which is a variance component, a reparametrization was made by getting the square of the variable used in the algorithm. This guarantees that the search is made just on the positive quadrant. In the discount factor case a reparametrization was made in order to guarantee that the values would lie in a particular segment of the real line, say from p to one, for $p > 0$.

predictive density functions. The CBF will be used to compare a BVAR model with $\boldsymbol{\lambda}$ fixed at $E(\boldsymbol{\lambda} \mid D_T)$ (model M_0) with an equivalent classical VAR model (model M_1). In this way, the logarithm of the CBF, which compares the two models M_0 and M_1 over k observations, is given by

$$\log(H_t(k)) = \log(p(\boldsymbol{y}_{t-k+1}, \ldots, \boldsymbol{y}_t \mid D_0, M_0)) - \log(p(\boldsymbol{y}_{t-k+1}, \ldots, \boldsymbol{y}_t \mid M_1)).$$

More details about the form of the Bayes factor we have used in our applications can be found in the Appendix B.

3. Simulated example

As explained previously, posterior quantities can be obtained through the SIR method based on a result that guarantees that if the number of samples obtained from the prior distribution, or any other importance function, is large enough, the adopted estimators converge in distribution for the posterior ones. However, the result does not indicate the rate of convergence and there is no way to evaluate the optimal sample size for a general problem. In order to evaluate this number, two experiments were conducted where the posterior distribution were analytically known. In both examples the model is normally distributed and conjugated priors are chosen such that the posterior distributions are well defined. In the first example the mean of the normal is resampled for a given variance, while in the second the variance is resampled for a given mean.

These methods tend to give poor results when the prior distribution and the likelihood function are too far from each other. Naturally, if the functions are too close, the prior's sample size would be less than that one if they do not overlap too much. For this reason, the exercise was repeated for different sample sizes and different "*distance*" grades. The distance grade was measured by the probability of being selected from the prior a value that is less or equal to the true value of the parameter.

3.1. Normal model with known variance

The first exercise considers the normal likelihood function with unknown mean θ and fixed variance, $\sigma^2 = 1$. The prior used for θ was the conjugate distribution which is, in this case, a normal distribution with mean μ and variance τ^2. Then if $\boldsymbol{x} = (x_1, x_2, \ldots, x_T)$ is such that x_i's are independent and identically distributed, with $(x_i \mid \theta) \sim \mathrm{N}(\theta, \sigma^2)$ and $\theta \sim \mathrm{N}(\mu, \tau^2)$. It can be shown that, $\theta \mid \boldsymbol{x} \sim \mathrm{N}(\mu_1, \tau_1^2)$, where $\mu_1 = (T\sigma^2\overline{x} + \tau^{-2}\mu)/(T\sigma^{-2} + \tau^{-2})$ and $\tau_1^{-2} = T\sigma^{-2} + \tau^{-2}$.

The SIR method was used to estimate the posterior distribution using the same sample of 50 normal draws with mean $\theta = 0$ and variance $\sigma^2 = 1$. The distance grade between the two distributions was measured by the prior as $P(|\theta| \leq 3)$. So, the larger is this probability the larger will be the overlapping area between prior distribution and likelihood function.

Table 1
Estimation of μ conditioned on σ^2

μ	$P(\lvert\theta\rvert \leq 3)$	True mean	True variance	m	Estimated mean	Estimated variance	Kolmogorov test (P-value)
0.0	0.9773	−0.0283	0.0196	100	−0.0207	0.0205	0.094(0.196)
				1000	−0.0271	0.0205	0.050(0.062)
1.2	0.7257	−0.0048	0.0196	100	−0.0134	0.0163	0.093(0.196)
				1000	0.0142	0.0175	0.096(0.062)
1.5	0.5000	0.0011	0.0196	100	0.1285	0.0482	0.269(0.196)
				1000	−0.0098	0.0212	0.076(0.062)

The quality of the posterior distribution approximation was evaluated by comparing the true distribution with the empirical one by using Kolmogorov's test.[8]

Table 1 presents the results of this exercise. In each case a sample of size n from the prior distribution, $\theta \sim \mathrm{N}(\mu, \tau^2)$, with $\tau^2 = 1$, was selected. Afterwards, the estimation for the posterior distribution was obtained and the mean, variance and Kolmogorov's test were computed. True means and the 1% critical value of Kolmogorov's test are also presented.[9] Initially we have considered a favorable situation where the prior's mean coincides with the likelihood's mode. In the worst situation we have considered, the probability of drawing a value from the prior that is less than or equal to the truth value is just 27%. For each case three sample sizes from the prior were considered, $m = (100, 500, 1000)$. The results show that the SIR method works better if the prior overlaps at least 70% of the relevant region and if at least $m = 1000$.

3.2. Normal model with unknown variance

The second experiment was done for normal distribution with known mean and unknown variance, σ^2. The distribution of the precision, $\phi = 1/\sigma^2$, is a gamma, which is conjugate for a normal distribution with unknown variance. So, if $\boldsymbol{x} = (x_1, x_2, \ldots, x_T)$ is a random sample from $\mathrm{N}(\theta, \phi^{-1})$ and $\phi \sim \mathrm{G}(1/\sigma_0^2, 1)$, it can be shown that $\phi \mid \boldsymbol{x} \sim \mathrm{G}(T/2 + 1/\sigma_0^2, Ts^2/2)$, where $Ts^2 = \sum_{t=1}^{T}(x_t - \bar{x})^2$.

The SIR method was used with a sample of size $T = 100$ from the standard normal, i.e. $\theta = 0$ and $\phi^{-1} = 1$. In each exercise, a sample of size m was drawn from a gamma distribution with parameters σ_0^2 and 1, respectively. So in this way, the mean, variance and the Kolmogorov's test were calculated. The true mean and variance are shown in the table, so as the critical value for Kolmogorov's test. The results show that the posterior mean and the variance of the ϕ can be obtained even in unfavorable situations. But as the "distance" grade between the distributions increases Kolmogorov's test rejects the hypothesis of equality between the true and the estimated posterior distributions of ϕ. Table 2 summarizes this results. It can be concluded that the moments of the posterior's variances are well calculated for any

[8] The statistical test is $D = \max|F - F^*|$, i.e. the maximum of the difference between the true cumulative distribution and the approximated one.

[9] Algorithm for generating random variables can be found in Ripley (1987).

Table 2
Estimation of σ^2 conditioned on θ

σ_0^{-2}	$P(\|\phi\|<1)$	True mean	True variance	m	Estimated mean	Estimated variance	Kolmogorov test (P-value)
1.2	0.54	1.11	0.0241	100	1.11	0.0259	0.200(0.196)
				1000	1.12	0.0214	0.056(0.062)
2.8	0.10	1.15	0.0248	100	1.15	0.0234	0.119(0.196)
				1000	1.14	0.0254	0.060(0.062)
3.3	0.05	1.16	0.0250	100	1.20	0.0230	0.154(0.196)
				1000	1.16	0.0238	0.045(0.062)

distance grade between prior distribution and likelihood function and for any size of sampling and re-sampling. However, when ϕ is drawn from the true distribution more frequently the Kolmogorov's test rejects the null hypothesis of equality between the distributions.

After these small simulated exercise, which are far from being conclusive we move to the next three sections where we apply our methodology to real economic data.

4. Real data applications

In this section we will apply our approach for three of the most important class of models used to analyze macroeconomic variables. We have used data from the Brazilian economy. In the first one we fit a Bayesian VAR model for a group of brazilian macroeconomic variables. In this case the hyperparameters are the Litterman's coefficients. In our second and third exercises we fit DLM with seasonal components, without and with transfer function.

4.1. Analysis of a BVAR model

In this application we have used a four-dimensional VAR model to analyze the dynamics of Brazilian macroeconomic variables where the time series were (i) exports excluding commodities (EXP), (ii) imports excluding oil (IMP), gross national product (GNP), and exchange rate (EXCH).[10]

Litterman (1986) suggests some practical strategies to elicit the prior distribution for the parameters in a VAR model. In our particular application $\boldsymbol{y}_t = (\mathrm{EXP}, \mathrm{IMP}, \mathrm{GNP}, \mathrm{EXCH})'$ is the vector of macroeconomic time series at time t, such that the VAR(p) is

$$\boldsymbol{y}_t = \boldsymbol{\Phi}\boldsymbol{x}_t + \sum_{l=1}^{p} \boldsymbol{\Gamma}_l \boldsymbol{y}_{t-l} + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathrm{NID}(0, \boldsymbol{\Sigma}), \tag{10}$$

[10] The models were estimated using data from 1975Q1 up to 1995Q4. The variables were log transformed and time periods 1986Q4, 1990Q3 and 1994Q4 were considered expurious (government plans).

where $\boldsymbol{x}_t$ is a vector containing all exogenous variables, such as intercept and seasonal dummies. In general, the prior distribution for $\boldsymbol{\Phi}$ is constant, while for $\boldsymbol{\Gamma}_{l,ij}$, the autoregressive coefficient of order l from the jth variable in the ith equation of the VAR, the prior is normal with zero mean and variance defined by

$$V_{l,ij} = \left(\frac{\lambda_{ij}}{l^d}\right)^2 \frac{\hat{\sigma}_i^{\,2}}{\hat{\sigma}_j^{\,2}}, \tag{11}$$

where the ratio $\hat{\sigma}_i^{\,2}/\hat{\sigma}_j^{\,2}$ unifies the scale among the time series, d is a decay factor (in general, $d=1$), such that lagged terms become less important, a priori. The hyperparameters λ_{ij} are the well-known Litterman's coefficients. Notice that $\lambda_{ij} \to 0$ means that y_j is, a priori, excluded from equation i. On the other hand, $\lambda \to \infty$ can be seen as non-informative prior distribution for all parameters. [11]

The Litterman's coefficients, λ's, can be specified with restriction, for example, imposing the same hyperparameter for the same block in all equations, i.e. $\lambda_{ij} = \lambda_j, \forall i$, or the same hyperparameter for all blocks of each equation, i.e. $\lambda_{ij} = \lambda_i, \forall j$.

In our applications, the hyperparameters' posterior were obtained by using the SIR method. For comparative purposes, we have also estimated them by maximum likelihood. One thousand samples were drawn from the prior distribution of the hyperparameters which was lognormal with $E(\lambda) = 6.14$ and $P(\lambda > 40) = 5\%$. Using the lognormal distribution we restrict the values of λ in the positive real line. [12] The mean and variance of this lognormal distribution were defined considering the fact that the hyperparameter is a Litterman's coefficient and for this reason a value equals to 40 implies a non-informative value, while informative values are all around 6.

We have considered four different scenarios. The first one estimate the four equations as a system with one hyperparameter per block of endogenous variable and respective lagged terms, for all four equations. The second one estimate as a system too, but with the same hyperparameter for the four blocks. The third and fourth cases were estimated considering the hyperparameters specified by equation. The third considers four hyperparameters by equation – one for each block – and the fourth one considers one hyperparameter by equation – one for the four blocks of variables. Therefore 25 distributions are estimated. Forecast performance was evaluated conditionally on: (i) the posterior mean of λ, $E(\lambda \mid D_T)$; (ii) the mode in the discrete distribution case; and (iii) the MLE of λ. For comparison reasons it is also shown an equivalent model without hyperparameters, i.e. a VAR model in the classical form, which is basically a multivariate regression model. [13]

In the case of just one hyperparameter for all the equations and blocks Fig. 1 shows the hyperparameter's posterior distribution. It is clear the gain in information.

[11] Further details about Bayesian estimation of VAR models can be found in Kadiyala and Karlsson (1993, 1997), Koop (1992), Lima et al. (1993) and Lopes (1994).

[12] Other distributions could have been used, such as the truncated normal or even the gamma distribution, since λ can be thought as a variance parameter.

[13] To save space the other results are not presented here.

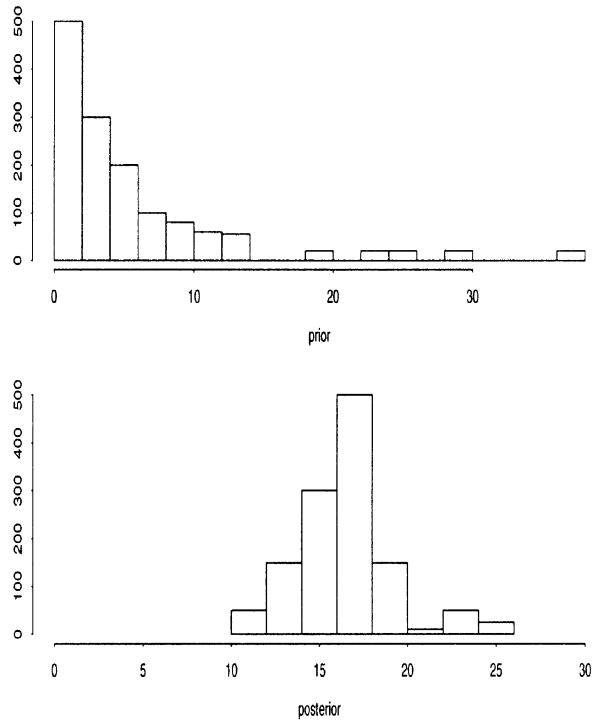

Fig. 1. Histogram of the prior and posterior distributions for Litterman's coefficient.

Table 3
Hyperparameters: Posterior means (with HPD intervals in parenthesis)

	Regressors				
Dependent	Syst	GNP	EXP	IMP	EXCH
GNP	17.7	18.7	5.7	6.8	1.3
	(13.6, 38.9)	(7.3, 49.8)	(0.6, 29.9)	(0.4, 28.6)	(0.1, 5.5)
EXP	9.7	2.0	5.1	3.6	2.7
	(5.9, 20.8)	(0.1, 5.4)	(0.9, 25.9)	(0.2, 10.6)	(0.1, 9.2)
IMP	12.7	8.5	1.1	1.6	2.5
	(7.4, 20.9)	(0.2, 14.5)	(0, 3.5)	(0.1, 6.1)	(0.1, 10.0)
EXCH	15.9	3.0	8.3	2.6	6.3
	(3.3, 38.7)	(0.2, 8.5)	(0.3, 30.7)	(0.1, 14.3)	(0.3, 38.7)
All blocks	16.6	7.2	1.7	0.7	0.8
	(15.7, 27.7)	(7.2, 19.0)	(0.3, 4.3)	(0.1, 2.3)	(0.1, 2.0)

The results for the other cases are summarized in Table 3 which shows the posterior mean and highest posterior density interval (which are described by the percentiles 3 and 97).

In general, it seems that the hyperparameters were estimated with great uncertainty, with maximum density posterior interval hardly wide, and with asymmetric distribution. This asymmetry turns the usual uncertainty estimators invalid. Also the assumption of equality between the hyperparameters estimated with and without

Table 4
Log predictive in the VAR and BVAR under different restrictions (log predictive are comparable through the columns)

Model	Conditional on	System	GNP	EXP	IMP	EXCH
M1	Mean	−2976.8	−2780.9	−2850.3	−2849.3	−2823.1
	Mode	−2977.1	−2780.6	−2848.4	−2846.0	−2820.2
	MLE	−2977.2	−2781.6	−2850.9	−2845.5	−2820.2
M2	Mean	−2976.9	−2782.8	−2850.4	−2847.7	−2821.0
	Mode	−2976.6	−2782.8	−2850.3	−2846.3	−2820.9
	MLE	−2976.9	−2782.5	−2850.4	−2846.3	−2820.9
M0	—	−2991.1	−2795.5	−2859.7	−2861.2	−2835.9

Table 5
Log predictive in the BVAR model conditional on the mean

Model	Estimation	GNP	EXP	IMP	EXCH
M1	Multivariate	−2783.0	−2851.5	−2850.9	−2827.2
	Univariate	−2780.9	−2850.3	−2849.3	−2823.1
M2	Multivariate	−2784.0	−2852.5	−2852.1	−2888.4
	Univariate	−2782.8	−2850.4	−2847.7	−2821.0
M0	—	−2795.5	−2859.7	−2861.2	−2835.9

restrictions is, in most of the cases, rejected. The hyperparameters with respect to a certain block, for example GNP's block, can be compared in the system and in each of the equations through the line. On the other side, the hyperparameters of the different blocks of each one of the equations can be compared through the columns. The following tables make use of the following notation: $(M1)$ to indicate that the equation or the set of equations were estimated with 4 hyperparameters; $(M2)$ to indicate that the equation or the set of equations were estimated with just one hyperparameter and, finally, $(M0)$ to indicate that the model was estimated from a classical point of view.

From Table 4 it can be seen that the logarithm of the predictive likelihoods for the model conditional on different values of the hyperparameters – hyperparameter's posterior mean, mode or the MLE – presents quite similar behavior, on both analysis, the one which considers each equation and the other one which considers the whole system. However, the predictive likelihoods are different when we compare models estimated with restrictions, for example, when $M2$ and $M1$ are compared; the result is in accordance with the comparisons of the HPD intervals for the hyperparameter. This table also shows that the model estimated in the classical form is significantly worse than that one estimated with some prior information.

Table 5 shows that the log predictive is significantly reduced when the introduction of the restriction that the hyperparameter value of a certain block is the same for all the equations. This table shows the logarithm of the predictive likelihood of each equation when they were jointly (multivariate) and separately (univariate) estimated. In the jointly estimation case the restriction of an equal hyperparameter in each block was made. The Theil-U statistic for one-step ahead forecast error is used to

Table 6
Other results from the models estimated by equation and jointly

			Theil-U				$E(y_{t+1} \mid D_t)$			
Estimation	Model	Condit.	GNP	EXP	IMP	EXCH	GNP	EXP0	IMP	EXCH
Univariate	M1	Marginal	0.552	0.837	0.774	1.003	4.946	2.763	3.299	1.521
	M1	Mean	0.541	0.841	0.797	1.055	4.942	2.774	3.297	1.521
	M2	Marginal	0.559	0.845	0.759	1.020	4.941	2.765	3.297	1.527
	M2	Mean	0.556	0.838	0.764	1.017	4.936	2.766	3.297	1.525
Multivariate	M1	Marginal	0.561	0.857	0.813	1.118	4.934	2.801	3.287	1.509
	M1	Mean	0.558	0.857	0.815	1.120	4.932	2.805	3.287	1.508
	M2	Marginal	0.536	0.870	0.839	1.136	4.932	2.811	3.282	1.502
	M2	Mean	0.552	0.870	0.840	1.137	4.931	2.813	3.281	1.502
	M0	—	0.780	1.200	1.156	1.355	4.922	2.836	3.270	1.487

Table 7
Theil-U for the forecast error 3 steps ahead (multivariate case)

Model	Conditional	GNP	EXP	IMP	EXCH
M1	Marginal	0.799	1.075	1.160	1.173
M1	Mean	0.810	1.080	1.170	1.181
M2	Marginal	0.831	1.077	1.188	1.201
M2	Mean	0.833	1.079	1.191	1.204
M0	—	0.936	1.129	1.362	1.302

compare the marginal model (which integrates out all the hyperparameter), with the one conditional on the posterior mean of the hyperparameter obtained from univariate and multivariate estimations.

Here $M0$ is also significantly worst. The results which are shown in Table 6 indicate that restriction effects over the hyperparameters becomes less visible, and also the marginal model is not substantially different from that model conditional on the hyperparameter's mean value. Table 7 shows the point forecasts and the Theil-U value of the forecast errors three steps ahead.

The results presented so far make clear that, when the whole sample is considered, the BVAR models are systematically and significantly better than their classical counterpart. Nevertheless, it is possible, and reasonable, that hyperparameter's effect diminish through time. Fig. 2 shows the CBF with a window of length 12. The plots refer to the comparison between those models conditional on the hyperparameter's mean value and the classical one. Part (a) refers to the system and parts (b)–(d) refer to GNP, non-oil imports and exchange rate, respectively. In all the cases where the CBF is between -2 and 2 the BVAR model is considered analogous to the VAR. The BVAR will be better if CBF is greater than 2 and worst if it is below -2.

At the very beginning of the sample the BVAR is significantly better. However, as times goes by, the BVAR model becomes analogous to the VAR. In this case the CBF changes its slope around the '90, which suggests a structural change which was not considered. In fact, the VAR model is equivalent to a BVAR when prior variance is quite large. In this case, the use of procedures such as monitoring and interventioning should be taken into account.

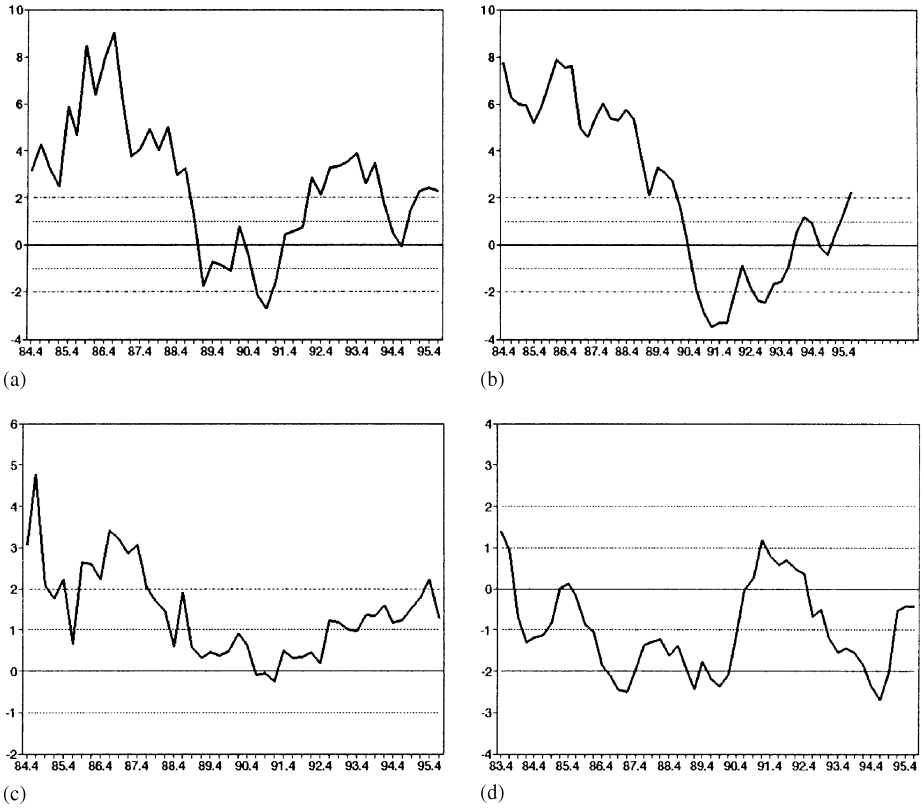

Fig. 2. CBF for: (a) system; (b) GNP; (c) non-oil imports; and (d) exchange rate.

4.2. Trend and seasonal models

In the DLM context the discount factor represents a special kind of hyperparameter and determines the volatility of each state variable. In general there is no unified definition of a discount factor, even though it is thought to be a parameter that measures the amount of information that a single information has through time.[14] Naturally, this factor can be treated as a hyperparameter and estimated by the SIR method or by maximum likelihood. A simple version of the model is given by

$$
\begin{aligned}
&y_t = \mu_t + \phi_{t0} + v_t, \quad v_t \sim \mathrm{N}(0, V_t),\\
&\mu_t = \mu_{t-1} + \beta_{t-1} + \omega_{1t},\\
&\beta_t = \beta_{t-1} + \omega_{2t},\\
&\phi_{tr} = \phi_{t-1,r+1} + \omega_{r+3,t} \quad (r = 0, 1, \ldots, p-2),\\
&\phi_{t,p-1} = \phi_{t-1,0} + \omega_{p+1,t}, \quad \omega \sim \mathrm{N}(0, \boldsymbol{W}_t).
\end{aligned}
$$

where μ_t is the level of the series, β_t is the current rate of change in the level, $\phi = (\phi_0, \ldots, \phi_{p-2})_t$ is a vector of $(p-1)$ seasonal components, V_t is the variance of

[14] Considering the middle life (n) of the information, the discount factor is equal to $(3n-1)/(3n+1)$.

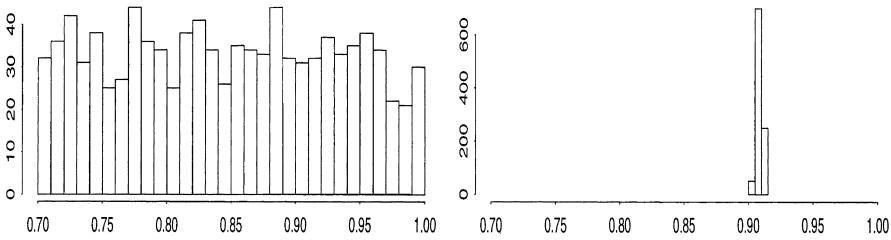

Fig. 3. Priori and posteriori distribution for the trend's discount factor.

Table 8
Posterior hyperparameter analysis

Model	Block	Mean	Mode	MLE	P_3	P_{50}	P_{97}
GNP	Trend	0.903	0.903	0.918	0.891	0.903	0.911
	Seas	0.984	0.998	0.999	0.997	0.998	0.998
IMP	Trend	0.921	0.931	0.914	0.885	0.911	0.935
	Seas	0.997	0.998	1.000	0.994	0.997	0.998
EXP	Trend	0.938	0.930	1.000	0.870	0.930	0.993
	Seas	0.997	0.999	0.999	0.993	0.998	0.999

the model, and $\boldsymbol{W}_t = diag(\boldsymbol{W}_t^{\mu,\beta}, \boldsymbol{W}_t^{\phi})$ is a block diagonal matrix that represents the variance of the states. Using DLM notation, as in the appendix, let $\boldsymbol{C}_t^{\mu,\beta}$ and $\boldsymbol{C}_t^{\phi}$ be, at time t, the posterior covariance matrixes of the trend and seasonal components, respectively. Then $\boldsymbol{W}_t^{\mu,\beta} = \boldsymbol{C}_t^{\mu,\beta}/\lambda_1$ and $\boldsymbol{W}_t^{\phi} = \boldsymbol{C}_t^{\phi}/\lambda_2$. Conditionally on the hyperparameters (λ_1, λ_2), which correspond to the discount factors, the model is a DLM which can be estimated based on the algorithm which is shown in the appendix. In order to estimate the hyperparameter, the methodology presented in Section 2 was adopted. We have used a sample of size 1000 from the prior distribution, which is assumed uniform over the interval $[0.7; 1]$. [15]

The model has been used to forecast the same components of the trade balance forecast model of the former section, excluding exchange rate, which does not present a seasonal pattern. As an illustration, Fig. 3 presents the histogram of the posterior distribution of the discount factor for GNP's trend components.

Table 8 summarizes the information of the hyperparameter's posterior distribution for all models. In all cases the expected value, the median and the mode of the hyperparameter posterior distribution are nearly the same and close to the maximum likelihood estimator. The maximum density posterior interval is approximately equal to the limits of the hyperparameter's set M, where the log predictive is at most 2 units less than the mode. This is an empirical criterion to compare models that work reasonably well. Besides, with approximately 95% of probability the discount factors lie in the interval $[0.87; 1]$.

[15] This prior distribution corresponds to a empirically relevant interval of a discount factor in this case (see West and Harrison (1997) for further details).

Table 9
Predictive analysis

Variable	Model	Theil-U	MAD	LP	Pred(1)	Pred(3)
GNP	Marginal	0.918	0.031	—	4.795	4.923
	Mean	0.917	0.047	−37232.9	4.796	4.923
	Mode	0.917	0.047	−37232.9	4.796	4.923
	MLE	0.871	0.045	37233.2	4.797	4.925
IMP	Marginal	0.931	0.119	—	3.418	3.584
	Mean	0.926	0.119	−39374.4	3.424	3.590
	Mode	0.943	0.120	−39374.6	3.420	3.590
	MLE	0.914	0.117	−39372.4	3.417	3.589
EXP	Marginal	1.102	0.124	—	2.734	2.813
	Mean	1.103	0.124	−38901.5	2.737	2.804
	Mode	1.080	0.121	−38896.8	2.733	2.814
	MLE	1.199	0.174	−38879.5	2.715	2.796

The marginal model, that is the one obtained by integrating out the hyperparameters, is more representative than those conditional on some hyperparameter value because hyperparameter's posterior distribution is considered as a whole in which way model uncertainty is fully and correctly incorporated in the statistical analysis.

From a classical point of view, the model is chosen conditionally on the MLE of the discount factors. So an interesting question that emerges is, do those two approaches generate different results from a forecasting point of view? The table below compares the predictive performance of some of these models, using as a criterion the forecast error one step ahead, the Theil-U and the mean absolute error (MAE), along with the log predictive (LP).

Looking at the predictive performance of the models, it can be seen that both models, the conditional and the marginal ones behave quite similar in this application. These results suggest that the MLE can be used to estimate first-order moments, but what should be stressed is that using SIR the full posterior can be obtained and consequently more information used.

Table 9 indicates that in the import and export cases the MLE presents less significant results compared to those obtained by the mode using the SIR method. A possible reason for this is that the posterior distribution is much more concentrated than the prior, specially in the case of the seasonal's hyperparameter block. In this way the number of draws might have not been sufficient. This characteristic, which is evident in the results, indicates that the SIR procedure should be repeated but with a better definition of the region of the hyperparameter to be sampled in the case of the seasonal block. Some extensions of this method are explored by Schmidt et al. (1998).

4.3. Transfer function models

A general class of non-linear and non-normal Bayesian models is the class of transfer function models. Here the methodology discussed in Section 2 will be used

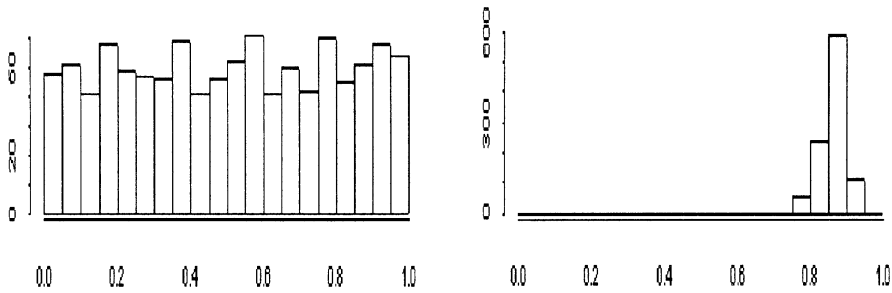

Fig. 4. Saturation's prior and posterior distributions.

to analyze a first-order transfer function model. In other terms the model is,

$$
\begin{aligned}
&y_t = \mu_t + E_t + \phi_{t0} + v_t, \quad v_t \sim \mathrm{N}(0, V_t),\\
&\mu_t = \mu_{t-1} + \beta_{t-1} + \omega_{1t},\\
&\beta_t = \beta_{t-1} + \omega_{2t},\\
&E_t = \lambda E_{t-1} + \gamma_{t-1} x_t + \omega_{3t},\\
&\gamma_t = \gamma_{t-1} + \omega_{4t},\\
&\phi_{tr} = \phi_{t-1,r+1} + \omega^*_{tr} \quad (r = 0, 1, \ldots, p-2),\\
&\phi_{t,p-1} = \phi_{t-1,0} + \omega^*_{t,p-1},\\
&\lambda \sim p(\lambda),
\end{aligned}
$$

where μ_t, β_t and ϕ_t's were defined in the last section, E_t is the transfer function term from x_t to y_t, λ measures the velocity of saturation, or the effect that x_t has in $y_t, y_{t+1}, \ldots$; and γ takes into account the magnitude of such effect. The relationship between y_t and x_t can also be seen by $E_t = \gamma_t(1 + \lambda L + \lambda L^2 + \cdots)x_t$.

Conditional on λ, our previous model becomes a well-known Bayesian DLM, with analytical solution. Therefore, λ can be considered a hyperparameter in the context we have already described. In this application we analyze the effect of the exchange rate (x_t) on the exports (or imports), y_t. We have also estimated λ by maximum likelihood in order to compare with our Bayesian procedure. Our prior for λ is restricted on the interval $[0, 1]$, which discard possible explosive cases. We have used a uniform prior on that interval to express complete lack of knowledge about λ. When export was the dependent variable we have estimated also the model in the non-linear form using the methodology proposed by (Migon and Harrison, 1985). [16]

Fig. 4 shows the posterior distribution for λ when considering exports the dependent variable and using a sample of size 1000 from the prior. We can observe that the posterior distribution is almost concentrated around 0.9, which indicates high persistency effect of exchange rate on exports.

[16] The non-linearity affects just the calculation of the prior distribution. The non-linear terms are approximated by the first term of the Taylor expansion around the posterior mean at the last period.

Table 10
Posterior analysis of the transfer function models

Model	Mean	Mode	MLE	Percentiles			Non-linear
				3%	Median	97%	
Imports	0.894	0.904	0.906	0.841	0.896	0.936	—
Exports	0.861	0.861	0.876	0.875	0.783	0.867	0.716

Table 11
Predictive analysis of the transfer function models

Variable	Predictive	Summaries			
	Model	Theil-U	DAM	LP	Pred(1)
Imports	Marginal	0.909	0.119	—	3.412
	Mean	0.910	0.119	−38409.2	3.412
	Mode	0.908	0.119	−38409.2	3.409
	MLE	0.908	0.119	−38409.2	3.408
Exports	Marginal	0.971	0.116	—	2.564
	Mean	0.972	0.116	−37926.6	2.565
	Mode	0.972	0.116	−37926.6	2.565
	MLE	0.970	0.116	−37926.6	2.565

Table 10 summarizes the posterior analysis for λ, while Table 11 summarizes the predictive analysis for imports and exports. The MLE and the posterior mode are quite similar, and the predictiveness of the models are also similar.

5. Conclusions

The use of prior modeling of hyperparameter to make dynamic modeling more flexible and powerful has been used in three empirical economical situations, and the posterior analysis has been conducted through the use of SIR techniques. To make the work richer we have compared our methodology to the maximum likelihood approach.

In our applications, the two approaches seem to be quite similar, but not in the BVAR case. Similar in the sense that conditioning on the MLE or on the posterior mean of the hyperparameter gives the same general results. However, we cannot generalize our empirical findings to other models, neither to other empirical datasets. Furthermore, more research needs to be done to provide effective answers for many open questions.

Nevertheless, some useful observations could be listed. First of all, the MLE is less expensive, computationally speaking, than the SIR; but on the other hand, its results are *local maximums*, which are hard to estimate since, in general, reparametrization is needed to restrict the range of the hyperparameters. The SIR presents *global results*,

which is an advantage over MLE methods. Of course, our procedure is subject to our challenges of Monte Carlo methods; for instance, we should have problems when the prior density and likelihood function do not overlap or when the posterior is to be concentrated at least around one of the hyperparameters. Therefore, care should be taken before applying such methods to problems with very little prior information. Other MCMC methods have been applied to dynamic models as well, such as the forward-filtering backward-smoothing proposed by Frühwirth-Schnatter (1994) and Carter and Kohn (1994) simultaneously, or the particle filtering algorithm proposed by Pitt and Shephard (1997). However, for the class of models we are considering it seems easier to use a non-iterative method, such as the SIR instead of iterative ones, such as the Gibbs sampler which might involve concerns as convergence to the posterior distribution. Recently, Gamerman and Moreira (1998) have proposed hybrid algorithms that mix the flexibility of SIR methods with powerful MCMC techniques.

To sum up, from a Bayesian viewpoint it does not make any sense to analyze those models using MLE. However, in some practical applications it does not matter which method has been used for computations. The main difference appears when analyzing the results. Even though our empirical results are quite similar to those using maximum likelihood, the Bayesian interpretation is clearer and precise and should be the one considered.

All numerical results presented in this paper were obtained by using an user-friendly software, named PRVWIN, which is now available at http://www.ipea.gov.br or upon request.

Acknowledgements

We are grateful to Stanley Azen (Editor), Peter Müller, Jonathan Stroud and an anonymous referee for comments and suggestions that considerably improved the article. The first author was partially supported by grants from IPEA and CAPES, Brazil.

Appendix A. DLM of common components

A dynamic linear model of common components appears when modeling a vector of time series with respect to the same common components, which might be regressors, indicator variables, seasonal components, and so on and so forth. A vector autoregressive model presents this behavior, i.e. all equations have, as regressors, the same set of lagged variables. Therefore, it is important to understand how the Bayesian approach works on this situation. See West and Harrison (1997), for details and references.

Let us assume q univariate time series Y_{tj} $(j=1,\ldots,q)$, each one following a standard univariate linear dynamic model, $\{\boldsymbol{F}_t, \boldsymbol{G}_t, V_t\sigma_j^2, \boldsymbol{W}_t\sigma_j^2\}$, where $\boldsymbol{F}_t(n\times 1)$ vector of common regressors is given. Then, the observation and system equations are,

respectively,

$$Y_{tj} = \boldsymbol{F}_t'\boldsymbol{\theta}_{tj} + v_{tj}, \quad v_{tj} \sim \mathrm{N}(0, V_t\sigma_j^2),$$

$$\boldsymbol{\theta}_{tj} = \boldsymbol{G}_t\boldsymbol{\theta}_{t-1,j} + \omega_{tj}, \quad \omega_{tj} \sim \mathrm{N}(0, \boldsymbol{W}_t\sigma_j^2).$$

Note that $\boldsymbol{F}_t, \boldsymbol{G}_t, V_t, \boldsymbol{W}_t$ are common to all time series, but each series has its vector of states $\boldsymbol{\theta}_{tj}$. In matrix representation we have,

$$\boldsymbol{Y}_t' = \boldsymbol{F}_t'\boldsymbol{\Theta}_t + \boldsymbol{v}_t', \quad \boldsymbol{v}_t \sim \mathrm{N}(0, V_t\boldsymbol{\Sigma}),$$

$$\boldsymbol{\Theta}_t = \boldsymbol{G}_t\boldsymbol{\Theta}_{t-1} + \boldsymbol{\Omega}_t, \quad \boldsymbol{\Omega}_t \sim \mathrm{N}(0, \boldsymbol{W}_t, \boldsymbol{\Sigma}),$$

where $\boldsymbol{Y}_t' = (Y_{t1}, \ldots, Y_{tq}), \boldsymbol{v}_t = (v_{t1}, \ldots, v_{tq}), \boldsymbol{\Theta}_t = (\boldsymbol{\theta}_{t1}, \ldots, \boldsymbol{\theta}_{tq}), \boldsymbol{\Omega}_t = (\boldsymbol{\omega}_{t1}, \ldots, \boldsymbol{\omega}_{tq})$ and $\boldsymbol{\Sigma} = \{\sigma_{ij}\}$, $\sigma_{jj} = \sigma_j^2$.

Also, let us suppose that the prior information could be written jointly as,

$$(\boldsymbol{\Theta}_0 \mid \boldsymbol{\Sigma}, D_0) \sim \mathrm{N}(\boldsymbol{m}_0, \boldsymbol{C}_0, \boldsymbol{\Sigma}),$$

$$(\boldsymbol{\Sigma} \mid D_0) \sim WI(\boldsymbol{S}_0, n_0)$$

which implies on matrix-variate marginal for $\boldsymbol{\Theta}_0$,

$$(\boldsymbol{\Theta}_0 \mid D_0) \sim T_{n_0}(\boldsymbol{m}_0, \boldsymbol{C}_0, \boldsymbol{S}_0).$$

The following theorem summarizes the sequential aspect of the Bayesian approach for dynamic linear models of common components.

Theorem. *The joint posterior distribution for* $(\boldsymbol{\Theta}, \boldsymbol{\Sigma})$ *for each time* t *can be described as follows*:

(a) *Posterior on* $t-1$

$$(\boldsymbol{\Theta}_{t-1} \mid \boldsymbol{\Sigma}, D_{t-1}) \sim \mathrm{N}(\boldsymbol{m}_{t-1}, \boldsymbol{C}_{t-1}, \boldsymbol{\Sigma}),$$

$$(\boldsymbol{\Sigma} \mid D_{t-1}) \sim WI(\boldsymbol{S}_{t-1}, n_{t-1}),$$

$$(\boldsymbol{\Theta}_{t-1} \mid D_{t-1}) \sim T_{n_{t-1}}(\boldsymbol{m}_{t-1}, \boldsymbol{C}_{t-1}, \boldsymbol{S}_{t-1}).$$

(b) *Prior on* t

$$(\boldsymbol{\Theta}_t \mid \boldsymbol{\Sigma}, D_{t-1}) \sim \mathrm{N}(\boldsymbol{a}_{t-1}, \boldsymbol{R}_{t-1}),$$

$$(\boldsymbol{\Sigma} \mid D_{t-1}) \sim WI(\boldsymbol{S}_{t-1}, n_{t-1}),$$

$$(\boldsymbol{\Theta}_t \mid D_{t-1}) \sim T_{n_t-1}(\boldsymbol{a}_{t-1}, \boldsymbol{R}_{t-1}, \boldsymbol{S}_{t-1}),$$

where $\boldsymbol{a}_t = \boldsymbol{G}_t\boldsymbol{m}_{t-1}, \boldsymbol{R}_t = \boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t' + \boldsymbol{W}_t$.

(c) *Predictive Distribution on* t

$$(\boldsymbol{Y}_t \mid \boldsymbol{\Sigma}, D_{t-1}) \sim \mathrm{N}(\boldsymbol{f}_t, Q_t\boldsymbol{\Sigma}),$$

$$(\boldsymbol{Y}_t \mid D_{t-1}) \sim T_{n_t}(\boldsymbol{f}_t, Q_t\boldsymbol{S}_{t-1}),$$

where $\boldsymbol{f}_t' = \boldsymbol{F}_t'\boldsymbol{a}_t, Q_t = \boldsymbol{F}_t'\boldsymbol{R}_t\boldsymbol{F}_t + V_t$.

(d) *Posterior on t*

$$(\boldsymbol{\Theta}_t \mid \boldsymbol{\Sigma}, D_t) \sim \mathrm{N}(\boldsymbol{m}_t, \boldsymbol{C}_t, \boldsymbol{\Sigma}),$$

$$(\boldsymbol{\Sigma} \mid D_t) \sim WI(\boldsymbol{S}_t, n_t),$$

$$(\boldsymbol{\Theta}_t \mid D_t) \sim T_{n_t}(\boldsymbol{m}_t, \boldsymbol{C}_t, \boldsymbol{S}_t),$$

where $\boldsymbol{m}_t = \boldsymbol{a}_t + \boldsymbol{A}_t \boldsymbol{e}_t', \boldsymbol{C}_t = \boldsymbol{R}_t + \boldsymbol{A}_t \boldsymbol{A}_t' Q_t, n_t = n_{t-1} + 1, \boldsymbol{S}_t = n_t^{-1}(n_{t-1}\boldsymbol{S}_{t-1} + \boldsymbol{e}_t \boldsymbol{e}_t'/Q_t)$, $\boldsymbol{A}_t = \boldsymbol{R}_t \boldsymbol{F}_t / Q_t$ *and* $\boldsymbol{e}_t = \boldsymbol{Y}_t - \boldsymbol{f}_t$.

Therefore, by the theorem, when $\boldsymbol{\Sigma}$ is assumed to be unknown,[17] the joint predictive distribution is

$$p(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T \mid D_0) \propto [n_{t-1} + \boldsymbol{e}_t^T (Q_t \boldsymbol{\Sigma})^{-1} \boldsymbol{e}_t]^{-(n_{t-1}+q)/2} |Q_t \boldsymbol{S}_{t-1}|^{-1/2}.$$

The matrix $\boldsymbol{W}_t$ can be defined as being proportional to $\boldsymbol{C}_t$. Then, the discount factor literature could be applied, i.e. $\boldsymbol{W}_t = \delta \boldsymbol{C}_t \delta'$ is block-diagonal, and the vector δ determines the volatility of the parameters, and can be defined arbitrarily or, in our exercises, estimated as a hyperparameter.

Appendix B. Bayes factor

The two models for $\boldsymbol{Y}$ can be compared in a straightforward way by means of Bayes factor. The Bayes factor from the model M_0 to the model M_1 is defined as the predictive ratio for the two models, i.e.

$$H_t = \frac{p(\boldsymbol{Y}_t \mid D_{t-1}, M_0)}{p(\boldsymbol{Y}_t \mid D_{t-1}, M_1)}.$$

In our analyses we have compared models with the same location, but with greater scale. Roughly speaking, the predictive distribution of the dynamic models we have considered in the paper has the form,

$$M_0 \equiv (\boldsymbol{Y}_t \mid D_{t-1}) \sim T_{n_{t-1}}(\boldsymbol{f}_t, Q_t \boldsymbol{S}_{t-1}),$$

$$M_1 \equiv (\boldsymbol{Y}_t \mid D_{t-1}) \sim T_{n_{t-1}}(\boldsymbol{f}_t, kQ_t \boldsymbol{S}_{t-1}).$$

[17] If $\boldsymbol{\Sigma}$ is known the joint predictive distribuiton is

$$p(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T \mid \boldsymbol{\Sigma}, D_0) \propto \prod_{t-1}^{T} \left\{ |Q_t \boldsymbol{\Sigma}|^{-1/2} \exp\left\{ -\frac{1}{2} \sum_{t-1}^{T} \boldsymbol{e}_t^T (Q_t \boldsymbol{\Sigma})^{-1} \boldsymbol{e}_t \right\} \right\}.$$

Then the Bayes factor we have used has the following form:

$$H_t = \frac{[n_{t-1} + \boldsymbol{e}_t^T(Q_t\boldsymbol{S}_{t-1})^{-1}\boldsymbol{e}_t]^{-(n_{t-1}+q)/2}|Q_t\boldsymbol{S}_{t-1}|^{-1/2}}{[n_{t-1} + \boldsymbol{e}_t^T(kQ_t\boldsymbol{S}_{t-1})^{-1}\boldsymbol{e}_t]^{-(n_{t-1}+q)/2}|kQ_t\boldsymbol{S}_{t-1}|^{-1/2}}$$

$$= k^{q/2}\left[\frac{n_{t-1} + \boldsymbol{e}_t^T(Q_t\boldsymbol{S}_{t-1})^{-1}\boldsymbol{e}_t}{n_{t-1} + \boldsymbol{e}_t^T(kQ_t\boldsymbol{S}_{t-1})^{-1}\boldsymbol{e}_t}\right]^{-(n_{t-1}+q)/2}.$$

Appendix C. Hypotheses testing in DLM of common components

Let us define,

$$\boldsymbol{\Theta} = (\,\boldsymbol{\theta}_1 \quad \ldots \quad \boldsymbol{\theta}_q\,) = \begin{pmatrix} \boldsymbol{\theta}_1^* \\ \vdots \\ \boldsymbol{\theta}_n^* \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Theta}_k \\ \boldsymbol{\Theta}_r \end{pmatrix},$$

where $\boldsymbol{\Theta}_k$ is $(k \times q)$, $\boldsymbol{\Theta}_r$ is $(r \times q)$ and $n = k + r$.

The main interest is to verify whether or not the last r components of $\boldsymbol{F}_t$ are jointly and statistically significant, i.e.

$$\mathrm{H}_0 : \boldsymbol{\Theta}_r^{\mathrm{T}} = 0.$$

We are using the transpose of $\boldsymbol{\Theta}$ to make the development easier. It follows that

$$vec(\boldsymbol{\Theta}_r^{\mathrm{T}}) \sim T_v(vec(\boldsymbol{m}_r^{\mathrm{T}}), \boldsymbol{C}_{22} \otimes \boldsymbol{S})$$

since $vec(\boldsymbol{\Theta}^{\mathrm{T}}) \sim T_v(vec(\boldsymbol{m}^{\mathrm{T}}), \boldsymbol{C} \otimes \boldsymbol{S})$ and

$$vec(\boldsymbol{m}^{\mathrm{T}}) = \begin{pmatrix} vec(\boldsymbol{m}_k^{\mathrm{T}}) \\ vec(\boldsymbol{m}_r^{\mathrm{T}}) \end{pmatrix} \quad \text{and} \quad \boldsymbol{C} = \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{pmatrix}.$$

Finally,

$$P\left(F_{rq,v} \geq \frac{1}{rq}\boldsymbol{m}_r(\boldsymbol{C}_{22}^{-1} \otimes \boldsymbol{S}^{-1})\boldsymbol{m}_r^{\mathrm{T}}\right)$$

can be interpreted as a sort of p-value for the F-test of the hypothesis H_0. As a matter of illustration, we can think of $\boldsymbol{m}_r$ as $(\alpha_1, \alpha_2, \ldots, \alpha_q,\ \beta_1, \beta_2, \ldots, \beta_q)$ where α's and β's are the constant terms and time trend for the first q equations, respectively.

References

Barbosa, E.P., 1989. Dynamic Bayesian models for vector time series analysis and forecasting. Ph.D. Thesis, University of Warwick, UK.

Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis, 2nd ed. Springer, Berlin.

Bernardo, J.M., Smith, A.F.M., 1994. Bayesian Theory. Wiley, New York.

Brian, B., 1984. Basic Optimization Methods. School of Mathematical Science, University of Bradford.

Carter, C.K., Kohn, R., 1994. On Gibbs sampling for state space models. Biometrika 81, 541–553.
van Dijk, H.K., Kloek, T., 1980. Further experience in Bayesian analysis using Monte Carlo integration. J. Econometrics 14, 307–328.
van Dijk, H.K., Kloek, T., 1985. Experiments with some alternatives for simple importance sampling in Monte Carlo integration. In Bayesian Statistics, vol. 2, pp. 511–530.
Efron, B., 1982. The Jackknife, the Bootstrap and Other Resampling Plans. Society for Industrial and Applied Mathematics.
Frühwirth-Schnatter, S., 1994. Data augmentation and dynamic linear models. J. Time Ser. Anal. 15, 183–202.
Gamerman, D., 1998. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Chapman & Hall, London.
Gamerman, D., Migon, H.S., 1998. Statistical Inference: an Integrated Approach. Edward Arnold.
Gamerman, D., Moreira, A.R.B., 1998. Bayesian analysis of econometric time series models using hybrid integration rules. Journal of econometrics (submitted).
Geweke, J., 1994. Bayesian comparison of econometric models. Working Paper 532, Federal Reserve Bank of Minneapolis.
Harvey, A.C., 1989. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, Cambridge.
Kadiyala, K.R., Karlsson, S., 1993. Forecasting with generalized Bayesian vector autoregressions. J. Forecasting 12, 365–378.
Kadiyala, K.R., Karlsson, S., 1997. Numerical methods for estimation and inference in Bayesian VAR-models. J. Appl. Econometrics 12, 99–132.
Koop, G., 1992. Aggregate shocks and macroeconomic fluctuations: a Bayesian approach. J. Appl. Econometrics 7, 395–411.
Lima, E.C.R., Migon, H.S., Lopes, H.F., 1993. Efeitos dinâmicos dos choques de oferta e demanda agregadas sobre o nível de atividade econômica do brasil (in portuguese). Rev. Bras. Economia 47, 177–204.
Litterman, R.B., 1986. Forecasting with Bayesian vector autoregressions: five years of experience. J. Business Econom. Statist. 4, 25–38.
Lopes, H.F., 1994. Aplicações de modelos autoregressivos vetoriais Bayesiano. Master Thesis, Federal University of Rio de Janeiro.
Lopes, H.F., Ehlers, R.S., 1997. Bayesian forecasting (the levels) of vector autoregressive log-transformed time series. Technical Report, Laboratory of Statistics, Federal University of Rio de Janeiro.
Migon, H.S., Harrison, P.J., 1985. An application of non-linear Bayesian forecasting to television advertising. In Bayesian Statistics, vol. 2, pp. 681–696.
Pitt, M.K., Shephard, N., 1997. Filtering via simulation: auxiliary particle filters. Mimeo.
Quintana, J.M., 1985. A dynamic linear matrix-variate regression model. Research Paper, Department of Statistics, University of Warwick, UK.
Rubin, D.B., 1988. Using the Sir algorithm to simulate posterior distributions. In Bayesian Statistics, vol. 3, pp. 395–402.
Schmidt A.M., Gamerman, D., Moreira, A.R.B., 1998. An adaptive resampling scheme for cycle estimation. J. Appl. Statist., submitted.
Smith, A.F.M., Gelfand, A.E., 1992. Bayesian statistics without tears: a sampling-resampling perspective Amer. Statist. 46, 84–88.
West, M., Harrison, J., 1997. Bayesian Forecasting and Dynamic Models, 2nd ed. Springer, Berlin.