Brazilian Journal of Probability and Statistics (2003), 17, pp. 91–105.

Expected posterior priors in factor analysis

Hedibert Freitas Lopes
Universidade Federal do Rio de Janeiro

Abstract: Bayesian inference in factor analytic models has received renewed attention in recent years, partly due to computational advances but also partly to applied focuses generating factor structures as exemplified by recent work in financial time series modeling. The focus of our current work is to investigate the commonly overlooked problem of prior specification and sensitivity in factor models. We accomplish that by implementing Pérez and Berger's (1999). Expected Posterior (EP) prior distributions. As opposed to alternative objective priors, such as Jeffreys' prior and Bernardo's prior, EP prior has several important theoretical and practical properties, with its straightforward computation through MCMC methods and coherence when comparing multiple models perhaps the most important ones.

Key words: Bayes' factors; Bayesian inference; expected posterior prior; latent factor models; model selection criteria; model uncertainty.

1 Introduction

Methodological innovations and real-world applications of factor analysis, and latent structure models more generally, have developed rapidly in recent years, partly due to increased access to appropriate computational tools. In particular, iterative MCMC simulation methods have very naturally opened up access to fully Bayesian treatments of factor analytic models, as developed and applied in, for example, Geweke and Zhou (1996) and Polasek (1997), with a historical view presented in Lopes and West (2004). The growing range of developments and creative applications in increasingly complex models, and with larger data sets in higher dimensions, justifies the view that computational advances have been critically enabling; the near future will very likely see much broader use of factor analysis in routine applied statistical work. Formal inference on the number of factors itself has been relatively ignored in the Bayesian literature, with most of the studies assuming that the number of factors is a modelling choice. Applied work typically studies sensitivity of predictions and variations/ambiguities of interpretations as the number of factors is varied as a control parameter. Lopes and West (2004), for instance, presents situations where overestimating the number of factor causes multimodality in the posterior distributions of most of the the factor model's parameters.

In Lopes and West (2004) we introduced, developed and explored MCMC methods for factor models that treat the number of factors as unknown. Building on prior work on MCMC methods for a given number of factors, we introduced a customised reversible jump Markov chain Monte Carlo (RJMCMC, see Green's (1995) algorithm for moving between models with different numbers of factors. We compared our sampler with alternative stochastic approximation methods. Most of these methods are derived to compute normalizing constants, such as the bridge sampling Meng and Wong (1996); the candidate estimator Chib (1995); the harmonic mean estimator Newton and Raftery (1994); Gelfand and Dey's estimator Gelfand and Dey (1994) and the well-know Laplace-Metropolis estimator Lewis and Raftery (1997), which later on will be of practical computational usage.

In the present work, our main goal is to explore the sensitiviness of factor models to prior hyperparameters. In other words, we want to study how model selection in the factor analysis context is affected by decreasing the amount of prior information. Is is well known that Bayesian model comparison is cumbersome when the prior is relatively uninformative. Many solutions have been proposed in the last two decades to overcome this problem, such as Spiegelhalter and Smith's (1982) imaginary samples, Berger and Pericchi's (1996) intrinsic Bayes factors and O'Hagan's (1995) fractional Bayes factor, among others. We will restrict our attention to Pérez and Berger's (2002) expected posterior prior, which is also an intrinsic prior Berger and Pericchi (2001), mainly due to its foundational and computational aspects, to be discussed shortly.

Section 2 defines the basic factor model framework, notation and structure, and briefly reviews and discusses issues of model specification and Bayesian analysis of the factor model when the number of factors is specified, based on standard Gibbs sampling. Expected posterior priors in the factor model context is introduced in Section 3, while Section 4 presents some comparative studies based on synthetic data and an application to international exchange rate markets. Section 5 concludes the paper with summary comments.

2 Factor model structure and specification

2.1 Basic model form

Data on m related variables are considered to arise through random sampling from a zero-mean multivariate normal distribution denoted by $N(\mathbf{0}, \mathbf{\Omega})$ where $\mathbf{\Omega}$ denotes the $m \times m$ non-singular variance matrix. A random sample of size T is denoted by $\{\boldsymbol{y}_t, t = 1, \ldots, T\}$. For any specified positive integer $k \leq m$, the standard $k-$factor model relates each $\boldsymbol{y}_t$ to an underlying $k-$vector of random variables $\boldsymbol{f}_t$, the common factors, via

$$\boldsymbol{y}_t = \boldsymbol{\beta}\boldsymbol{f}_t + \boldsymbol{\epsilon}_t \tag{2.1}$$

where (i) the factors $\boldsymbol{f}_t$ are independent with $\boldsymbol{f}_t \sim N(\mathbf{0}, \boldsymbol{I}_k)$, (ii) the $\boldsymbol{\epsilon}_t$ are independent normal $m-$vectors with $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \mathbf{\Sigma})$, and $\mathbf{\Sigma} = \text{diag}(\sigma_1^2, \cdots, \sigma_m^2)$,

(iii) $\boldsymbol{\epsilon}_t$ and $\boldsymbol{f}_s$ are independent for all t and s, and (iv) $\boldsymbol{\beta}$ is the $m \times k$ factor loadings matrix.

Under this model, the variance-covariance structure of the data distribution is constrained; we have $\boldsymbol{\Omega} = V(\boldsymbol{y}_t|\boldsymbol{\Omega}) = V(\boldsymbol{y}_t|\boldsymbol{\beta}, \boldsymbol{\Sigma})$ given by $\boldsymbol{\Omega} = \boldsymbol{\beta}\boldsymbol{\beta}' + \boldsymbol{\Sigma}$. The model implies that, conditional on the common factors, the observable variables are uncorrelated: hence the common factors explain all the dependence structure among the m variables. For any elements y_{it} and y_{jt} of $\boldsymbol{y}_t$ and conditionally on $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, we have the characterising moments, (i) $\text{var}(y_{it}|\boldsymbol{f}) = \sigma_i^2$, (ii) $\text{cov}(y_{it}, y_{jt}|\boldsymbol{f}) = 0$, (iii) $\text{var}(y_{it}) = \sum_{l=1}^{k} \beta_{il}^2 + \sigma_i^2$, and (iv) $\text{cov}(y_{it}, y_{jt}) = \sum_{l=1}^{k} \beta_{il}\beta_{jl}$.

In practical problems, especially with larger values of m, the number of factors k will often be small relative to m, so that much of the variance-covariance structure is explained by the common factors. The *uniquenesses,* or *idiosyncratic variances,* σ_i^2 measure the residual variability in each of the data variables once that contributed by the factors is accounted for. The model (2.1) can be written as

$$\boldsymbol{y} = \boldsymbol{F}\boldsymbol{\beta}' + \boldsymbol{\epsilon} \tag{2.2}$$

where $\boldsymbol{y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_T)'$, $\boldsymbol{F} = (\boldsymbol{f}_1, \cdots, \boldsymbol{f}_T)'$ and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_T)'$ are matrices of dimension $(T \times m), (T \times k)$ and $(T \times m)$, respectively. The elements $\boldsymbol{\epsilon}$ and $\boldsymbol{F}$ are mutually independent matrix normal random variables, as in Dawid (1981), Press (1982) and West and Harrison (1997), chapter 16. The notation, as in Dawid (1981), is simply $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{I}_T, \boldsymbol{\Sigma})$. We then have densities

$$p(\boldsymbol{y}|\boldsymbol{F}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-T/2}\text{etr}(-0.5\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon}\boldsymbol{\epsilon}') \tag{2.3}$$

and, marginalising over $\boldsymbol{F}$,

$$p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Omega}|^{-T/2}\text{etr}(-0.5\boldsymbol{\Omega}^{-1}\boldsymbol{y}'\boldsymbol{y}) \tag{2.4}$$

where $\text{etr}(\mathbf{A}) = \exp(\text{trace}(\mathbf{A}))$ for any matrix $\mathbf{A}$. The likelihood function (2.3) will be subsequently used in Gibbs sampling for the parameters of a factor model with k fixed, whereas the likelihood form (2.4) is extensively used in the RJMCMC algorithm proposed in Lopes and West (2003).

2.2 Model structure and identification issues

As is well-known, the $k-$factor model must be further constrained to define a unique model free from identification problems. First we address the standard issue that the model is invariant under transformations of the form $\boldsymbol{\beta}^* = \boldsymbol{\beta}\boldsymbol{P}'$ and $\boldsymbol{f}_t^* = \boldsymbol{P}\boldsymbol{f}_t$, where $\boldsymbol{P}$ is any orthogonal $k \times k$ matrix. There are many ways of identifying the model by imposing constraints on $\boldsymbol{\beta}$, including constraints to orthogonal $\boldsymbol{\beta}$ matrices, and constraints such that $\boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}$ is diagonal (see Seber, 1984, for example). The alternative preferred here is to constrain so that $\boldsymbol{\beta}$ is a block lower triangular matrix, assumed to be of full rank, with diagonal elements strictly positive. This form is used, for example, in Geweke and Zhou (1996) and Aguilar and West (2000), and provides both identification and, often, useful

interpretation of the factor model. In this form, the loadings matrix has $r = mk - k(k-1)/2$ free parameters. With m non-zero σ_i parameters, the resulting factor form of $\mathbf{\Omega}$ has $m(k+1) - k(k-1)/2$ parameters, compared with the total $m(m+1)/2$ in an unconstrained (or $k = m$) model; leading to the constraint that $m(m+1)/2 - m(k+1) + k(k-1)/2 \geq 0$, which provides an upper bound on k. For example, $m = 6$ implies $k \leq 3$, $m = 12$ implies $k \leq 7$, $m = 20$ implies $k \leq 14$, $m = 50$ implies $k \leq 40$, and so on. Even for small m, the bound will often not matter as relevant k values will not be so large. In realistic problems, with m in double digits or more, the resulting bound will rarely matter. Finally, note that the number of factors can be increased beyond such bounds by setting one or more of the residual variances σ_i to zero.

A question arises about the full-rank assumption for $\boldsymbol{\beta}$. This was addressed in Geweke and Singleton (1980) who shown that, if $\boldsymbol{\beta}$ is rank deficient, then the model is unidentified. Specifically, if $\boldsymbol{\beta}$ has rank $r < k$ there exists a matrix $\boldsymbol{Q}$ such that $\boldsymbol{\beta Q} = \mathbf{0}$, $\boldsymbol{Q'Q} = \boldsymbol{I}$ and, for any orthogonal matrix $\boldsymbol{M}$, such that $\mathbf{\Omega} = \boldsymbol{\beta\beta'} + \mathbf{\Sigma} = (\boldsymbol{\beta} + \boldsymbol{MQ'})'(\boldsymbol{\beta} + \boldsymbol{MQ'}) + (\mathbf{\Sigma} - \boldsymbol{MM'})$. This translation invariance of $\mathbf{\Omega}$ under the factor model implies lack of identification and, in application, induces symmetries and potential multimodalities in resulting likelihood functions. This issue relates intimately to the question of uncertainty of the number of factors, discussed further below.

A final question concerns the ordering of the y_{it} variables and the connection between a chosen ordering and the specific form of the factor loading matrix above. The order of variables is a modelling decision that has no effect on the resulting theoretical model nor on predictive inferences under the model. Given the k-factor model (2.1) specified and appropriate for the $\boldsymbol{y}$ with variables in a specific order, alternative orderings are trivially produced via $\boldsymbol{Ay}_t$ for some rotation matrix $\boldsymbol{A}$. Model (2.1) then transforms to a similar factor model for the reordered data $\boldsymbol{Ay}_t$ with the same latent factors but transformed loadings matrix $\boldsymbol{A\beta}$. This new loadings matrix does not have the lower triangular structure. However, we can always find an orthonormal matrix $\boldsymbol{P}$ such that $\boldsymbol{A\beta P'}$ is lower triangular, and so simply recover the factor model in precisely the form (2.1) with the same probability structure for the underlying latent factors $\boldsymbol{Pf}_t$. This result confirms that the order of the variables in $\boldsymbol{y}_t$ is theoretically irrelevant assuming that k is properly chosen. However, when it comes to model estimation, the order of variables has a determining effect on the choice of k, and the interaction between variable order and model fitting can be quite subtle, as we illustrate in examples below.

2.3 Elements of prior specification

To complete the model specification we require classes of priors for the model parameters $\boldsymbol{\beta}$ and $\mathbf{\Sigma}$. Our reported analyses are based on very diffuse but proper priors with the following ingredients. For the factor loadings, we take independent priors such that $\beta_{ij} \sim N(0, C_0)$ when $i \neq j$, and $\beta_{ii} \sim N(0, C_0)\mathbf{1}(\beta_{ii} > 0)$ for the upper-diagonal elements of positive loadings $i = 1, \cdots, k$. The latter simply truncates the basic normal prior to restrict the diagonal elements to positive values.

Analysis now requires only that we specify the variance parameter C_0, which we take to be rather large in the studies below. See Table 1.

For each of the idiosyncratic variances σ_i^2 we assume a common inverse gamma prior, and take the variances to be independent. Specifically, the σ_i^2 are independently modelled as $\sigma_i^2 \sim IG(\nu/2, \nu s^2/2)$ with specified hyperparameters ν and s^2. Here s^2 is the prior mode of each σ_i^2 and ν the prior degrees of freedom hyperparameter. Our examples below assume low values of ν to produce diffuse though proper priors. Note that we eschew the use of standard improper reference priors $p(\sigma_i^2) \propto 1/\sigma_i^2$. Such priors lead to the Bayesian analogue of the so-called *Heywood problem* Martin and McDonald (1981), Ihara and Kano (1995). In terms of these variance parameters, likelihood functions in factor models are bounded below away from zero as σ_i^2 tends to zero, so inducing singularities in the posterior at zero. Proper priors that decay to zero at the origin obviate this problem and induce proper posteriors. See Table 1.

2.4 MCMC methods in a $k-$factor model

With a specified $k-$factor model, Bayesian analyses using MCMC methods are straightforward. We simply summarise the main ingredients here, referring to Lopes and West (2004) for further details. MCMC analysis involves iteratively simulating from sets of conditional posterior distributions which, in this model, are standard forms. A basic method simulates from the conditional posteriors for each of $\boldsymbol{F}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ in turn, utilising the following sets of full conditional posteriors arising from our model as specified. These are as follows.

First, the factor model in (2.2) can be seen as a standard multivariate regression model with "parameters" $\boldsymbol{F}$ when $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ and k are fixed (e.g., Press, 1982, Box and Tiao, 1973, Broemeling, 1985 and Zellner, 1971). It easily follows that the full conditional posterior for $\boldsymbol{F}$ factors into independent normal distributions for the $\boldsymbol{f}_t$, whose mean vector and covariance matrix are $(\boldsymbol{I}_k + \boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t$ and $(\boldsymbol{I}_k + \boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta})^{-1})$, respectively. Second, the full conditional posterior for $\boldsymbol{\beta}$ also factors into independent margins for the non-zero elements of the rows of $\boldsymbol{\beta}$, as follows. For rows $i = 1, \ldots, k$, write $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ii})'$ for just these non-zero elements. For the remaining rows $i = k+1, \ldots, m$, write $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ik})$. Similarly, for $i = 1, \ldots, k$ denote by $\boldsymbol{F}_i$ the $T \times i$ matrix containing the first i columns of $\boldsymbol{F}$, and for all i let $\boldsymbol{y}_i$ be the column i of $\boldsymbol{y}$. Finally, it is trivially deduced that full conditional posterior for the elements of $\boldsymbol{\Sigma}$ reduces to a set of m independent inverse gammas, with $\sigma_i^2 \sim IG((\nu+T)/2, (\nu s^2 + d_i)/2)$ where $d_i = (\boldsymbol{y}_i - \boldsymbol{F}\boldsymbol{\beta}_i')'(\boldsymbol{y}_i - \boldsymbol{F}\boldsymbol{\beta}_i')$. Then we have full conditionals as follows: for $i = 1, \ldots, k$, $\boldsymbol{\beta}_i \sim N(\boldsymbol{m}_i, \boldsymbol{C}_i)\mathbf{1}(\beta_{ii} > 0)$ where $\boldsymbol{m}_i = \boldsymbol{C}_i\left(C_0^{-1}\mu_0\mathbf{1}_i + \sigma_i^{-2}\boldsymbol{F}_i'\boldsymbol{y}_i\right)$ and $\boldsymbol{C}_i^{-1} = C_0^{-1}\boldsymbol{I}_i + \sigma_i^{-2}\boldsymbol{F}_i'\boldsymbol{F}_i$. For $i = k+1, \ldots, m$, $\boldsymbol{\beta}_i \sim N(\boldsymbol{m}_i, \boldsymbol{C}_i)$ where $\boldsymbol{m}_i = \boldsymbol{C}_i\left(C_0^{-1}\mu_0\mathbf{1}_k + \sigma_i^{-2}\boldsymbol{F}'\boldsymbol{y}_i\right)$ and $\boldsymbol{C}_i^{-1} = C_0^{-1}\boldsymbol{I}_k + \sigma_i^{-2}\boldsymbol{F}'\boldsymbol{F}$. These distributions are easy to sample from.

3 Expected posterior priors

In this section we explore default analysis of factor models using Expected Posterior priors distributions Pérez and Berger (2002). Our main motivation is the difficulty in assessing prior information in factor models, both because it involves multivariate prior assessment and because of the highly nonlinear structure that drives the parameters in a factor model. More specifically, we want to study how sensitive the predictive distribution is to the values of the hyperparameters C_0, ν and νs^2 from section 2.3, or more generally, to the prior distribution chosen for the loadings, $\boldsymbol{\beta}$, and for the idiosyncratic variances, $\boldsymbol{\Sigma}$.

It is broadly known that, as far as estimation and posterior analysis is concerned, the effect of the prior decreases as the sample size increases. Also, even in the case where improper priors are used, it is most likely that the posteriors will be proper. However, when it comes to model comparison, the prior's weight does not cancel out so easily and in the worst scenario, when priors are improper, Bayes factors are as large or as small as the decision makers might please. On the other hand, there are many situations where the investigator wants to use as little prior information as possible, either because he does not have it or because he wants to avoid further scrutiny from competing investigators that might suggest that his work is subjective.

In factor analysis there might be situations where a psychologist wants, in an exploratory factor analysis, to identify possible traits that his patients share. In order to do so, one might want to consider factor models with different number of factors and then pick that one with the highest posterior probability and then draw his conclusions (eventually after trying a couple of rotations). If the researcher has no or quite scarce prior information he will find it hard to assess prior distributions and might be willing to use default methods, especially when the procedure is used routinely.

Many attempts have being made in order to overcome such fallacy. Among others, Spiegelhalter and Smith's (1982) imaginary samples, Berger and Pericchi's (1996) intrinsic Bayes factors and O'Hagan's (1995) fractional Bayes factor, are just some examples of how important (partial) solutions for this problem might be. In a recent article, Pérez and Berger (2002) propose what they call *Expected Posterior Priors*, which are obtained as

$$\pi_k^*(\boldsymbol{\theta}_k) = \int \pi_k^N(\boldsymbol{\theta}_k|\boldsymbol{y}^*)m^*(\boldsymbol{y}^*)d\boldsymbol{y}^* \tag{3.1}$$

where y^* is a training sample and m^* is a measure on the training sample space. Here $\pi_k^N(\boldsymbol{\theta}_k|\boldsymbol{y}^*)$ is the posterior density for $\boldsymbol{\theta}_k$ based on the "vague" or noninformative (usually improper) prior and under model $\mathcal{M}_k$. In our notation, $\boldsymbol{\theta}_k$ corresponds to the loading matrix and the uniquenesses variances in a i-factor model.

The key feature of this method comes from the fact that the Bayes factor of M_k to $M_{k'}$ is given by

$$B_{kk'}^*(\boldsymbol{y}) = \frac{m_{\pi_k^*}(\boldsymbol{y})}{m_{\pi_{k'}^*}(\boldsymbol{y})}$$

$$
\begin{aligned}
&= \frac{\int p_k(\boldsymbol{y}|\boldsymbol{\theta}_k)\pi_k^*(\boldsymbol{\theta}_k)d\boldsymbol{\theta}_k}{\int p_{k'}(\boldsymbol{y}|\boldsymbol{\theta}_{k'})\pi_{k'}^*(\boldsymbol{\theta}_{k'})d\boldsymbol{\theta}_{k'}} \\
&= \frac{\int m_k^N(\boldsymbol{y}|\boldsymbol{y}^*)m^*(\boldsymbol{y}^*)d\boldsymbol{y}^*}{\int m_{k'}^N(\boldsymbol{y}|\boldsymbol{y}^*)m^*(\boldsymbol{y}^*)d\boldsymbol{y}^*}
\end{aligned} \tag{3.2}
$$

where $m_k^N(\boldsymbol{y}|\boldsymbol{y}^*) = m_k^N(\boldsymbol{y},\boldsymbol{y}^*)/m_k^N(\boldsymbol{y}^*)$. Notice that all multiplicative constant factors cancel out. Additionally, it is coherent using this Bayes factor, $B_{kk'}^* = 1/B_{k'k}^*$, and $B_{kk'}^* = B_{kl}^* B_{lk'}^*$. Besides coherence, other attractive properties of EP priors are: (i) only one measure m^* needs to be specified, (ii) MCMC output can be promptly used to its computation (see equation (3.9) below), and (iii) improper priors can be used since scaling constant cancel out (see equation (3.2) above). Further results can be found in Pérez and Berger (2002) or Pérez (1998), for a more detailed approach.

3.1 Computational issues

In order to compute the Bayes factor in (3.2) for the factor model at hand, we need first to obtain an approximation for $m_{\pi_k^*}$. But, firstly, let us rewrite $m_{\pi_k^*}$ as a function of m_k^N and m^*. Notice that

$$
\begin{aligned}
m_{\pi_k^*}(\boldsymbol{y}) &= \int p_k(\boldsymbol{y}|\boldsymbol{\theta}_k)\pi_k^*(\boldsymbol{\theta}_k)d\boldsymbol{\theta}_k \\
&= \int p_k(\boldsymbol{y}|\boldsymbol{\theta}_k)\left[\int \pi_k^N(\boldsymbol{\theta}_k|\boldsymbol{y}^*)m^*(\boldsymbol{y}^*)d\boldsymbol{y}^*\right]d\boldsymbol{\theta}_k \\
&= \int \underbrace{\left[\int \frac{p_k(\boldsymbol{y},\boldsymbol{y}^*|\boldsymbol{\theta}_k)\pi_k^N(\boldsymbol{\theta}_k)}{m_k^N(\boldsymbol{y}^*)}d\boldsymbol{\theta}_k\right]}_{m_k^N(\boldsymbol{y},\boldsymbol{y}^*)/m_k^N(\boldsymbol{y}^*)} m^*(\boldsymbol{y}^*)d\boldsymbol{y}^* \\
&= m_k^N(\boldsymbol{y})\int \frac{m_k^N(\boldsymbol{y}^*|\boldsymbol{y})}{m_k^N(\boldsymbol{y}^*)}m^*(\boldsymbol{y}^*)d\boldsymbol{y}^*
\end{aligned} \tag{3.3}
$$

and, accordingly, the Bayes factor $B_{kk'}^*(\boldsymbol{y})$ can be written as

$$
B_{kk'}^*(\boldsymbol{y}) = \frac{m_k^N(\boldsymbol{y})}{m_{k'}^N(\boldsymbol{y})}\frac{\int \frac{m_k^N(\boldsymbol{y}^*|\boldsymbol{y})}{m_k^N(\boldsymbol{y}^*)}m^*(\boldsymbol{y}^*)d\boldsymbol{y}^*}{\int \frac{m_{k'}^N(\boldsymbol{y}^*|\boldsymbol{y})}{m_{k'}^N(\boldsymbol{y}^*)}m^*(\boldsymbol{y}^*)d\boldsymbol{y}^*} \tag{3.4}
$$

Expression (3.3) gives us a hint of how to obtain an MCMC approximation for $m_{\pi_k^*}(\boldsymbol{y})$. Usually, $\boldsymbol{y}^*$ will be much smaller than $\boldsymbol{y}$, because it will be based on a minimal training sample Pérez and Berger (2002). Initially, let $\boldsymbol{y}_1^*,\ldots,\boldsymbol{y}_R^*$ be random draws taken from $m^*(\boldsymbol{y}^*)$. An approximation for the integral in (3.3) is

$$
\hat{I} = \frac{1}{R}\sum_{r=1}^{R}\frac{m_k^N(\boldsymbol{y}_r^*|\boldsymbol{y})}{m_k^N(\boldsymbol{y}_r^*)} \tag{3.5}
$$

and according to well established MC results, Geweke (1989), it can be shown that

$$\hat{I} \rightarrow \int \frac{m_k^N(\boldsymbol{y}^*|\boldsymbol{y})}{m_k^N(\boldsymbol{y}^*)} m^*(\boldsymbol{y}^*) d\boldsymbol{y}^*$$

as R goes to infinity. Therefore,

$$\hat{m}_{\pi_k^*}(\boldsymbol{y}) = \frac{m_k^N(\boldsymbol{y})}{R} \sum_{r=1}^{R} \frac{m_k^N(\boldsymbol{y}_r^*|\boldsymbol{y})}{m_k^N(\boldsymbol{y}_r^*)}$$

approximates $m_{\pi_k^*}(\boldsymbol{y})$ in the same sense. The question that remains now is how such predictive measures can be computed. Rephrasing it, the question is how normalizing constants can be computed. In order to compute $\hat{m}_{\pi_k^*}(\boldsymbol{y})$, we need first to compute $m_k^N(\boldsymbol{y})$, $m_k^N(\boldsymbol{y}_r^*|\boldsymbol{y})$ and $m_k^N(\boldsymbol{y}_r^*)$ for $r = 1, \ldots, R$. Firstly, it is easy to see that

$$m_k^N(\boldsymbol{y}^*|\boldsymbol{y}) = \int p(\boldsymbol{y}^*|\boldsymbol{\theta}_k) \pi_k^N(\boldsymbol{\theta}_k|\boldsymbol{y}) d\boldsymbol{\theta}_k \tag{3.6}$$

where, having at hand draws from the posterior distribution under model i, say $\boldsymbol{\theta}_k^{(1)}, \ldots, \boldsymbol{\theta}_k^{(M)}$, this integral can be approximated by

$$\hat{m}_k^N(\boldsymbol{y}^*|\boldsymbol{y}) = \frac{1}{M} \sum_{m=1}^{M} p(\boldsymbol{y}^*|\boldsymbol{\theta}_k^{(m)}) \tag{3.7}$$

which has the same asymptotic properties as $\hat{I}$. On the other, to compute $m_k^N(\boldsymbol{y})$ (or $m_k^N(\boldsymbol{y}^*)$) we need to rely on other kinds of approximations. However, now we need an estimator that (i) is minimally reliable, (ii) is based on a sample from the posterior distribution and (iii) is fast enough to be used many times. In Lopes and West's (2004) simulations and applications, the Laplace-Metropolis estimator was shown to be considerably reliable, besides being straightforwardly implementable.

3.1.1 Laplace-Metropolis estimator

The Laplace-Metropolis estimator combines analytic posterior approximations with MCMC output to modify traditional Laplace asymptotics, Tierney and Kadane (1986). Discussed in Lewis and Raftery (1997), the resulting estimator has the form

$$\hat{p}_{LM} = \hat{m}_k^N(\boldsymbol{y}) = (2\pi)^{d/2} |\boldsymbol{\Psi}_k|^{1/2} p_k(\boldsymbol{y}|k, \tilde{\boldsymbol{\theta}}_k) p_k^N(\tilde{\boldsymbol{\theta}}_k) \tag{3.8}$$

where $\tilde{\boldsymbol{\theta}}_k$ maximizes $p(\boldsymbol{y}|k, \boldsymbol{\theta}_k)\pi_k^N(\boldsymbol{\theta}_k)$ among the M posterior draws, $\boldsymbol{\Psi}_k$ is the MCMC approximation to the posterior variance of $\boldsymbol{\theta}_k$, and d is the dimension of $\boldsymbol{\theta}_k$. Variations on this method replace $\tilde{\boldsymbol{\theta}}_k$ by the MCMC approximation to the posterior mean. Hence, our final approximation for $m_{\pi_k^*}(\boldsymbol{y})$ is

$$\hat{\hat{m}}_{\pi_k^*}(\boldsymbol{y}) = \frac{|\boldsymbol{\Psi}|^{1/2} p_k(\boldsymbol{y}|\tilde{\boldsymbol{\theta}}_k) \pi_k^N(\tilde{\boldsymbol{\theta}}_k)}{RM} \sum_{r=1}^{R} \frac{\sum_{m=1}^{M} p(\boldsymbol{y}_r^*|\boldsymbol{\theta}_{k,r}^{(m)})}{|\boldsymbol{\Psi}_{k,r}|^{1/2} p_k(\boldsymbol{y}_r^*|\tilde{\boldsymbol{\theta}}_{k,r}) \pi_k^N(\tilde{\boldsymbol{\theta}}_{k,r})} \tag{3.9}$$

where $\tilde{\boldsymbol{\theta}}_{k,r}$ and $\boldsymbol{\Psi}_{k,r}$ are analogous to $\tilde{\boldsymbol{\theta}}_k$ and $\boldsymbol{\Psi}_k$, respectively, but based on the sample $\boldsymbol{\theta}_{k,r}^{(1)}, \ldots, \boldsymbol{\theta}_{k,r}^{(M)}$ taken from $\pi_k^N(\boldsymbol{\theta}|\boldsymbol{y}_r^*)$, for $r = 1, \ldots, R$. Notice that $\pi_k^N(.)$ appears in both the numerator and denominator of (3.9), so cancelling out any multiplicative constant factor.

4 Applications

We investigated two distinct simulated situations, a single-factor and a three-factor models, with the main goal of studying model selection robustness to prior hyperparameter choice when performing the estimation with and without considering expected posterior prior distributions. As for a real data analysis, we apply EP priors to model the exchange rates for six countries Lopes and West (2004).

4.1 A one-factor model

We simulated $T = 200$ observations from a $k = 1$-factor model for a vector $\boldsymbol{y}$ of $m = 7$ variables. The factor loading matrix and the uniquenesses variances are fixed at $\boldsymbol{\beta}_1 = (0.995, 0.975, 0.949, 0.922, 0.894, 0.866, 0.837)$ and $diag(\boldsymbol{\Sigma}_1) = (0.01, 0.05, 0.10, 0.15, 0.20, 0.25,$
$0.30)$, respectively. We have run the Gibbs sampler for $M_0 = 5000$ iterations, after which every tenth draw has been stored in a sample of size 1000 for further computation. In this model, i.e. when $k = K_{max} = 3$, the maximum number of parameters is 25. Therefore, we have chosen to make $\boldsymbol{y}^*$ represent a sample with 26 observations taken from a multivariate normal distribution with zero mean. The covariance matrix used was the sample covariance matrix based on $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{200}$. Notice that this would correspond to a factor model with no factor model at all, which in a sense represents the basic model.

As prior hyperparameters we have tried three different sets summarized in Table 1. With these choices we believe we can express decreasing amounts of information present in the prior distributions.

Table 1 Prior hyperparameters. $i = 1, \ldots, 7$ and $k = 1, \ldots, 3$

	μ_{ki}	C_{k0}	ν_{ki}	$\nu_{ki}s_{ki}^2$
I	0	1	2.2	0.1
II	0	10	2.2	0.01
III	0	100	2.2	0.0001

Finally, we have chosen $R = 1000$ in order to compute the predictive distributions. Table 2 summarizes our findings. It can be seen that the predictive density varies considerably when computed with different priors, while it is more stable or "robust" when based on expected posterior prior distributions. If the prior model probabilities are uniform, then posterior model probabilities, $\hat{P}_k^N(k)$ and $\hat{\hat{P}}_{\pi_k^*}(k)$

can also be computed. From Table 2 it seems that in the context of the simulated example either prior set up leads to the same answer. This observed robustness to the prior reflects the propriety of the intrinsic prior generated by the expected posterior prior setup. Additionally, it is well know that predictive densities are computationally unstable even when the prior is proper but relatively vague.

Table 2 Comparative results (one-factor model). $\hat{m}_k^N(\boldsymbol{y})$ and $\hat{\hat{m}}_{\pi_k^*}(\boldsymbol{y})$ are given by equations (3.8) and (3.9), respectively. Also, $\hat{P}_k^N(k)$ and $\hat{\hat{P}}_{\pi_k^*}(k)$ are the posterior model probabilities, which are computed from Bayes factors, B_ij, by $P(k) = (1 + \sum_{j=1}^{3} B_{jk})^{-1}$ for $k = 1, 2, 3$.

Prior	$\hat{m}_k^N(\boldsymbol{y})$			$\hat{\hat{m}}_{\pi_k^*}(\boldsymbol{y})$		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
(I)	-933.7	-941.1	-948.6	-906.9	-908.4	-911.0
(II)	-950.2	-962.4	-970.9	-907.1	-911.1	-914.3
(III)	-989.4	-1009.3	-1030.4	-910.8	-916.3	-928.1
	$\hat{P}_k^N(k)$			$\hat{\hat{P}}_{\pi_k^*}(k)$		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
(I)	0.999	0.001	0.000	0.804	0.182	0.014
(II)	1.000	0.000	0.000	0.980	0.019	0.001
(III)	1.000	0.000	0.000	0.996	0.004	0.000

4.2 A three-factor model

We introduce a second but relatively unimportant factor in order to see whether or not the EP prior captures such characteristic. We simulated $T = 200$ observations from a $k = 3$-factor model for a vector $\boldsymbol{y}$ of $m = 7$ variables. The factor loading matrix and the uniquenesses variances are fixed at $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$ and $\boldsymbol{\beta}_1' = (0.795, 0.0, 0.0, 0.722, 0.0, 0.0, 0.722)$, $\boldsymbol{\beta}_2' = (0.0, 0.775, 0.0, 0.0, 0.794, 0.0, 0.0)$ and $\boldsymbol{\beta}_3' = (0.0, 0.0, 0.749, 0.0, 0.0, 0.766, 0.0)$, and $diag(\boldsymbol{\Sigma}) = (0.37, 0.40, 0.44, 0.48, 0.37, 0.41, 0.48)$, respectively. We ran the Gibbs sampler for $M_0 = 1000$ iterations, after which $M = 1000$ draws were stored for further computation. In this model the maximum number of parameters is still 25 and, as before, $\boldsymbol{y}^*$ represents a sample of size 26 taken from a multivariate normal distribution with zero mean. The covariance matrix was computed accordingly. We used the same three different prior sets and $R = 1000$. Our findings are summarized in Table 3. The posterior probabilities were all one hundred percent in favor of the $k = 3$-factor model. Notice that, even though all methods pointed out towards the correct model, more robust estimates were found for the various hyperparameter sets by using the expected posterior priors regardless of the choice of the probability measure, m_k^*.

Table 3 Comparative results (three-factor model). $\hat{m}_k^N(\boldsymbol{y})$ and $\hat{\hat{m}}_{\pi_k^*}(\boldsymbol{y})$ are given by equations (3.8) and (3.9), respectively. Also, $\hat{P}_k^N(k)$ and $\hat{\hat{P}}_{\pi_k^*}(k)$ are the posterior model probabilities. $m^* \sim N(\mathbf{0}, \boldsymbol{S})$, where S is the sample covariance matrix. $m_k^* \sim N(\mathbf{0}, \widehat{\boldsymbol{\Omega}}_k)$, where $\widehat{\boldsymbol{\Omega}}_k$ is the maximum likelihood estimator of $\boldsymbol{\beta}_k\boldsymbol{\beta}_k' + \boldsymbol{\Sigma}_k$ based on a k-factor model.

	$m^* \sim N(\mathbf{0}, \boldsymbol{S})$					
Prior	$\hat{m}_k^N(\boldsymbol{y})$			$\hat{\hat{m}}_{\pi_k^*}(\boldsymbol{y})$		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
(I)	-1954.4	-1915.9	-1884.5	-1923.5	-1887.2	-1853.5
(II)	-1978.8	-1944.3	-1918.6	-1928.6	-1886.7	-1856.2
(III)	-2022.1	-1988.2	-1970.5	-1926.3	-1874.4	-1840.5
	$m^* = m_1^*$					
Prior	$\hat{m}_k^N(\boldsymbol{y})$			$\hat{\hat{m}}_{\pi_k^*}(\boldsymbol{y})$		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
(I)	-1954.4	-1915.9	-1884.5	-1929.0	-1893.0	-1862.9
(II)	-1978.8	-1944.3	-1918.6	-1921.6	-1896.2	-1868.5
(III)	-2022.1	-1988.2	-1970.5	-1929.6	-1880.2	-1855.7
	$m^* = m_2^*$					
Prior	$\hat{m}_k^N(\boldsymbol{y})$			$\hat{\hat{m}}_{\pi_k^*}(\boldsymbol{y})$		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
(I)	-1954.4	-1915.9	-1884.5	-1922.1	-1881.5	-1858.0
(II)	-1978.8	-1944.3	-1918.6	-1925.7	-1880.7	-1861.4
(III)	-2022.1	-1988.2	-1970.5	-1924.9	-1874.2	-1850.5
	$m^* = m_3^*$					
Prior	$\hat{m}_k^N(\boldsymbol{y})$			$\hat{\hat{m}}_{\pi_k^*}(\boldsymbol{y})$		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
(I)	-1954.4	-1915.9	-1884.5	-1924.0	-1884.0	-1850.0
(II)	-1978.8	-1944.3	-1918.6	-1922.4	-1887.2	-1858.4
(III)	-2022.1	-1988.2	-1970.5	-1922.6	-1874.2	-1844.8

4.3 Exchange rates

As a real situation application, in this section we explore the nuances of applying the expected posterior prior distribution when modeling changes in monthly international exchange rate returns for six countries in British pounds. This dataset was previously analysed in West and Harrison (1997, pp. 610–618). More recently, Lopes and West (2004) revisited this dataset when comparing, empirically, the sampling performances of various alternative ways of computing normalizing constants.

The currencies are the US dollar (US), Japanese yen (JAP), Canadian dollar (CAN), French franc (FRA), Italian lira (ITA), and the (West) German (Deutsch) mark (GER). We the application that follows, we have used the variables in this order. The data span the period from 1/1975 to 12/1986 inclusive. Each series has been standardised with respect to its sample mean and standard deviation over the period for comparability. It is well known in the factor analyis literature that such linear transformations does not affect the modeling process and factor structure analysis. According to Lopes and West's (2004) analysis, a two factor model is preferred, i.e. it had higher posterior probability, when the prior information was fairly informative. There we set $\mu_{k0} = 0, C_{k0} = 1$ and $\nu_{ki} = 2.2, \nu_{ki} s_{ki}^2 = 0.1$ for $i = 1, \ldots, 6$ and $k = 1, \ldots, 3$. Even though we have considered a variety of hyperparameters in that work and found out, at least empirically, that the effect of such hyperparameters was vacuous, we are now able to compare the previous analysis with one where little or none prior information is available and predictive distributions are computed according to Pérez and Berger's (2002) expected posterior prior methodology. The prior distributions (I-III) are the same used in the simulated cases. The Gibbs sampler and the computation of normalizing constants were done in the same way, with $M = 1000$ and $R = 1000$. Table 4 summarizes our findings.

As can be seen, there is not much difference when we look at the posterior probability for each model (here, the number of factors). However, it is clear how unstable the predictive is when based on noninformative priors indicating, as one would have expected, that the predictive is arbitrary in this case. On the other hand, by using expected posterior prior distributions the predictives are fairly robust. We have also tried $m^* \sim N(\mathbf{0}, \widehat{\mathbf{\Omega}})$, where $\widehat{\mathbf{\Omega}}$ is the maximum likelihood estimator of $\boldsymbol{\beta}\boldsymbol{\beta}' + \mathbf{\Sigma}$ based on a one factor model. Despite the fact that intuition would suggest that such choice would rather favor a one-factor model, we have found, at least in this simulation, that this is not the case. As a matter of fact, the results are virtually the same as the one in Table 4.

Table 4 Comparative results for the exchange rate data

Prior	$\hat{m}_k^N(\boldsymbol{y})$			$\hat{\hat{m}}_{\pi_k^*}(\boldsymbol{y})$		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
(I)	-1015.5	-904.4	-910.3	-993.5	-878.4	-884.2
(II)	-1039.1	-936.5	-947.9	-992.7	-880.4	-886.1
(III)	-1073.2	-973.3	-989.5	-997.7	-882.2	-888.4
	$\hat{P}_k^N(k)$			$\hat{\hat{P}}_{\pi_k^*}(k)$		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
(I)	0.0	0.997	0.003	0.0	0.997	0.003
(II)	0.0	1.000	0.000	0.0	0.997	0.003
(III)	0.0	0.994	0.006	0.0	0.998	0.002

5 Concluding comments

The main objective of this work was to investigate the extent to which the prior information about the loading matrix parameters and the idiosyncrasies affect the choice factor models. We accomplish that by implementing Pérez and Berger's (2002). Expected Posterior (EP) prior distributions. More specifically, we wanted to study the extent to which the (lack of) prior information affects model selection in the factor analysis context.

From the simulated examples presented in Section (3), it can be seen that Pérez and Berger's expected posterior prior performs considerably well in factor analysis when the prior information is absent. In both simulation studies model selection was robust to the various prior specifications and, perhaps more importantly, this robustness remained regardless of the choice of the probability measure, m_k^*.

Our findings are by no means based on exhaustive scenarios. Nonetheless, we believe that those simulations are reasonably realistic and in accordance with previous studies in factor analysis (see, for instance, Lopes and West (2004), and Aguilar and West (2000). Moreover, by using the Laplace-Metropolis estimator, most of the calculations are easily performed in any modest machine and took no more than 10 minutes overall.

Looking ahead, we note that related developments in dynamic factor modelling Aguilar and West (2000), and more recently in Lopes and Migon (2002), in financial time series and portfolio studies are focussed almost exclusively on short-term forecasting and the potential improvements available in forecasting moderate to high-dimensional time series using factor structures. Here the assessment of the number of factors and the sensitiveness to the choice of prior hyperparameters are also live issues that are complicated by the time-varying nature of such models that leads to the notion of time-variation in the number of (practically relevant) factors.

Acknowledgements

This research was partially supported by CAPES and was developed during the author's PhD studies at Duke University.

(Received January, 2003. Accepted June, 2003.)

References

Aguilar, O. and West, M. (2000). Bayesian dynamic factor models and variance matrix discounting for portfolio allocation. *Journal of Business and Economic Statistics*, **18**, 338–357.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*,

91, 109–122.

Berger, J. O. and Pericchi, L. (2001). Objective Bayesian methods for model selection: Introduction and comparison. *IMS Lecture Notes-Monograph Series*, **38**, 135–207.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Massachusetts: Addison-Wesley.

Broemeling, L. D. (1985). *Bayesian Analysis of Linear Models*. 472p. New York: Marcel Dekker.

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

Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, **68**, 265–274.

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

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

Geweke, J. F. and Singleton, K. J. (1980). Interpreting the likelihood ratio statistic in factor models when sample size is small. *Journal of the American Statistical Association*, **75**, 133–137.

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

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Ihara, M. and Kano, Y. (1995). Identifiability of full, marginal, and conditional factor analysis model. *Statistics and Probability Letters*, **23**, 343–350.

Lewis, S. M. and Raftery, A. E. (1997). Estimating Bayes' factors via posterior simulation with the Laplace-Metropolis estimator. *Journal of the American Statistical Association*, **92**, 648–655.

Lopes, H. F. and Migon, H. S. (2002). Comovements and contagion in emergent markets: Stock indexes volatilities. In *Case Studies in Bayesian Statistics*, **6**, Gatsonis, C. and Kass, R.E. and Carriquiry, A.L. and Gelman, A. and Verdinelli and Pauler, D and Higdon, D., Eds. Springer-Verlag, 287–302.

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

Martin, J. L. and McDonald, R. P. (1981). Bayesian estimation in unrestricted factor analysis: A treatment for Heywood cases. *Psychometrika*, **40**, 505–517.

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

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

O'Hagan, A. (1995). Fractional Bayes factor for model comparison. *Journal of the Royal Statistical Society, Ser. B*, **57**, 99–138.

Pérez, J. M. (1998). *Development of expected posterior prior distributions for model comparison*. Ph. D. Thesis, Purdue University.

Pérez, J. M. and Berger, J. M. (2002). Expected posterior prior distributions for model selection. *Biometrika*, **89**, 491–512.

Polasek, W. (1997). *Factor analysis and outliers: A Bayesian approach*. Discussion Paper, University of Basel.

Press, S. J. (1982). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*. 2.ed. New York: Krieger.

Seber, G. A. F. (1984). *Multivariate Observations*. New York: Wiley.

Spiegelhalter, D. J. and Smith, A. F. M. (1982). Bayes factor for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society B*, **44**, 377–387.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Society*, **81**, 82–86.

West, M. and Harrison, P. J. (1997). *Bayesian Forecasting and Dynamic Model*. New York: Springer-Verlag.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley.

Hedibert Freitas Lopes
Universidade Federal do Rio de Janeiro
Rio de Janeiro, RJ, Brazil
E-mail: hedibert@im.ufrj.br