Invariant Inference and Efficient Computation in the Static Factor Model

Joshua Chan, Roberto Leon-Gonzalez & Rodney W. Strachan

Invariant Inference and Efficient Computation in the Static Factor Model

Joshua Chan

Centre for Applied Macroeconomic Analysis,

University of Technology Sydney

Roberto Leon-Gonzalez

National Graduate Institute for Policy Studies and

Rimini Center for Economic Analysis

Rodney W. Strachan[1]

Centre for Applied Macroeconomic Analysis,

University of Queensland and

Rimini Center for Economic Analysis[2]

January 2017

[1]Corresponding author: Rodney Strachan, School of Economics, The University of Queensland, St Lucia 4072 QLD, Australia, email: r.strachan@uq.edu.au

ABSTRACT

Factor models are used in a wide range of areas. Two issues with Bayesian versions of these models are a lack of invariance to ordering of and scaling of the variables and computational inefficiency. This paper develops invariant and efficient Bayesian methods for estimating static factor models. This approach leads to inference that does not depend upon the ordering or scaling of the variables, and we provide arguments to explain this invariance. Beginning from identified parameters which are subject to orthogonality restrictions, we use parameter expansions to obtain a specification with computationally convenient conditional posteriors. We show significant gains in computational efficiency. Identifying restrictions that are commonly employed result in interpretable factors or loadings and, using our approach, these can be imposed ex-post. This allows us to investigate several alternative identifying (non-invariant) schemes without the need to respecify and resample the model. We illustrate the methods with two macroeconomic datasets.

Keywords: Bayesian, Markov Chain Monte Carlo, Reduced rank regression.

1 Introduction

Factor models have proven useful in many areas including psychology, genomics, epidemiology, economics and finance and significant advances in computation using Bayesian approaches (for example, Geweke and Zhou (1996), Aguilar and West (2000) and Chib, Nardari, and Shephard (2006)) have made Bayesian analysis of such models feasible for a range of applications. Two problems that have hampered Bayesian inference in factor models are, first, the models are not invariant to different ordering of the variables (see, for example, Lopes and West (2004)) and, second, poor efficiency of computation algorithms (e.g., Chib *et al*. (2006)).

This paper makes a number of contributions. i) This paper presents an invariant specification. That is, the specification will result in inference that does not depend upon the ordering of the variables. ii) We use parameter expansions to develop an algorithm that is both easy to implement and computationally efficient. The resulting posteriors have relatively simple normal forms. Further, as with the extant non-invariant specifications, our specification is overparameterised. However, we follow the rules of Liu and Wu (1999) to ensure efficiency gains. It is not clear that extant specifications do follow these rules and this may explain to some degree the poor sampling. iii) Finally, we provide a formal explanation for why extant specifications are not invariant, that is, why the evidence in the model can change when the order of the variables changes. In doing so, we demonstrate that there is not an identification problem so much as a specification problem in these models.

Reordering of variables involves groups of transformations of the parameters in the model. We therefore use group theory to show why existing specifications are not invariant to reordering of the variables. Work to date considering invariance has taken one of two approaches. The first approach attempts to resolve the issue by averaging over orderings (see for example Geweke (1996) and Frühwirth-Schnatter and Lopes (2010)). To estimate k we would need to estimate all orderings for all values of k. Averaging over orderings shows promise in small dimensional settings, but as

applications often have many variables (sometimes hundreds), the number of potential orderings to average over increases into the trillions making an averaging approach computationally infeasible or at best challenging. For example, Forni, Giannone, Lippi and Reichlin (2009) investigate some 89 series and find there are between $k = 12$ and $k = 18$ factors. In this case, if we were to use a non-invariant Bayesian approach and average over all orderings we would need to average over more than 237 trillion for $k = 12$ and 3 million trillion models for $k = 18$. It would seem more practical and feasible to only have one invariant model to consider for each k.

Another approach, therefore, is to develop a single model that does not depend upon the orderings. Examples of work taking this approach, besides our paper, are Bhattacharya and Dunson (2011), Aßmann, Boysen-Hogrefe, and Pape (2012) and Kaufmann and Schumacher (2012). Our approach differs from these in that we explicitly take the perspective of the factor model as a reduced rank regression model, such as in Bai and Ng (2002), and use previous work that utilizes the geometry of that model to develop an invariant model specification and inferential framework. Taking this perspective leads us to the view that, contrary to general belief, there is not an identification problem in the factor model but rather there is a problem with the specification used.

The invariant specification we propose uses a singular value decomposition as in Hoff (2007). This approach is related to the principal components specification commonly used in the frequentist literature (see, for example, Connor and Korajczyk (1986) and Bai and Ng (2002)) and efficient computation is achieved by combining this with and extension of the parameter expansion approach of Ghosh and Dunson (2009) in the static factor model and Koop, Léon- González and Strachan (2010 & 2012) in the vector error correction model and instrumental variables model. The parameter expansions are chosen to obtain a specification that is simple to implement, in fact simpler than standard extant specifications. A further benefit of this expanded specification is that the resulting sampler is efficient. This parameter expansion may be viewed as a generalization of the Ghosh and Dunson (2009) approach to computing factor models but with the added benefit of invariance.

In Section 2 we present the identified parameters in the invariant Bayesian specification of the static factor model and the priors for this model from a singular value decomposition (SVD) (contribution i). We then introduce the full parameter expansion using invariant transformations to obtain the prior for the 'expanded' model (contribution ii). We present the posterior, sampling algorithm and the posterior probability estimation for this 'expanded' model. In Section 3 we briefly outline the features of the static factor model and discuss relationships, in particular mappings, among existing identification schemes, two of which are popular non-invariant specifications used in the Bayesian literature and one invariant specification used in the frequentist literature. This discussion allows us to present the source of invariance (contributions iii, iv and v). Section 4 presents several applications including one small application to six exchange rates to demonstrate the effect of reordering and efficiency of the proposed sampling algorithm in this paper. Section 5 provides some concluding comments and potential extensions.

2   The invariant static factor model

In this section we present the invariant specification of the static factor model. In the model stacked over time, we will show that the product of the matrix of all factors and the loading matrix forms a reduced rank matrix. The row and column space of this matrix are identified (as are various norms). There is a smooth relationship between these spaces and appropriate orthonormal frames which we take advantage of to achieve an invariant specification. The parameters identified under the likelihood are the elements of the singular value decomposition of a reduced rank matrix, and therefore some of the identified parameters are orthonormal k-frames that belong to the Stiefel manifold. As discussed in Section 3, most of the literature have attempted to estimate parameters which are discontinuous transformations of the identified parameters, and it is this discontinuity that causes lack of invariance. In contrast, we build on the theory of invariant measures and distributions on special manifolds (e.g. Amari (1985), Chikuse (2003)) and first specify priors directly on the identified

parameters. In order to facilitate computations we then introduce non-identified parameters that allow us to define diffeomorphic transformations from the identified and non-identified parameters (the parameter expansions) to parameters with computationaly convenient supports (Real space) and distributions (normal). Because the transformations are diffeomorphic the resulting approach preserves the invariant (order independent) inference, while allowing for a much more efficient algorithm for computations. The prior that we propose is defined in such a way that conditional posteriors belong to standard families and allow for simple and efficient computations. Although previous work by Hoff (2007) proposes an invariant approach that works directly with the identified parameters, we show in the supplemental material B that thanks to the parameter expansions the efficiency gains of our approach can be very large.

We can write the factor model as a reduced rank regression model for a $1 \times n$ vector $y_t \in \mathbb{R}^n$ with $k < n$ factors as

$$y_t = f_t \Lambda + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \Sigma), \qquad \text{for } t = 1, ..., T \tag{1}$$

where f_t is a $1 \times k$ vector, Λ is an $k \times n$ matrix, and ε_t is a $1 \times n$ vector with a diagonal covariance matrix denoted by Σ. By stacking observations equation (**??**) can be equivalently written as $y = F\Lambda + \varepsilon = \Pi + \varepsilon$, where y and ε are $T \times n$ matrices with $E(vec(\varepsilon) vec(\varepsilon)') = \Sigma \otimes I_T$, F is a $T \times k$ matrix such that the matrix $\Pi = F\Lambda$ has rank k. In this section we develop priors for Λ and F by beginning with the parameters that are identified under the likelihood and, via a series of parameter expansions, we obtain the prior and posterior for the expanded model.

The reduced rank model in (**??**) has the same structure as a one-mode analysis used in psychometrics (see, for example, Magnus and Neudecker (1988)) for which frequentist approaches to estimation are proposed. Bayesian inference in other reduced rank models, such as the cointegrating vector error correction model and the overidentified simultaneous equations model, has been extensively explored and this literature is informative on how to approach the analysis of this model. Bayesian approaches most relevant to this paper are Strachan and Inder (2004), Koop,

Léon-González and Strachan (2010 & 2012).

Taking a singular value decomposition (SVD) of the reduced rank matrix $F\Lambda$, we have

$$F\Lambda = U_1 S_1 V_1' \tag{2}$$

$$U_1 \in V_{k,T} \quad V_1 \in V_{k,n}$$

$$S_1 = diag(s_1, s_2, \ldots, s_k)$$

where $s_i > s_{i+1} > 0$ for all i and $V_{m,n}$ denotes the Stiefel manifold such that $V_{m,n} = \{H(n \times m); H'H = I_m\}$ (for discussion, see Muirhead (1982)). All of the parameters U_1, S_1 and V_1 are identified up to sign and have, respectively, $Tk - \frac{k(k+1)}{2}$, k and $nk - \frac{k(k+1)}{2}$ free elements.

In what follows, we define the trace of a square matrix A as $tr(A)$. We specify priors for U_1, S_1 and V_1 with the form

$$\frac{f(S_1, V_1)(dS_1)\left(U_1'dU_1\right)\left(V_1'dV_1\right)c_O}{c_N c_U}$$

$$f(S_1, V_1) \propto \exp\left\{-\frac{c_\lambda}{2}tr\left(V_1'\underline{M}^{-1}V_1 S_1^2\right)\right\} 2^{-k} |S_1|^{n-k} \prod_{i<j}^{k}\left(s_i^2 - s_j^2\right),$$

$$\frac{c_N}{c_O} = \int f(S_1, V_1)(dS_1)(V_1'dV_1), \quad c_U = \int (U_1'dU_1) = \frac{2^k \pi^{\frac{Tk}{2}}}{\Gamma_k\left(\frac{T}{2}\right)}$$

$$c_O = \int (C'dC) = \frac{2^k \pi^{\frac{k^2}{2}}}{\Gamma_k\left(\frac{k}{2}\right)}, \quad \Gamma_k\left(\frac{m}{2}\right) = \pi^{k(k-1)/2} \prod_{i=1}^{k}\Gamma\left[\frac{m-i+1}{2}\right],$$

where $\underline{M}$ is a matrix that can be fixed equal to the identity matrix I_n for a prior that is invariant to ordering only, or equal to Σ for a prior that is invariant to both ordering and scale transformation. The prior for U_1 is uniform on the Stiefel manifold (for further discussion see James (1954)) and the prior for V_1 is uniform also when $\underline{M} = I_n$ but not when $\underline{M} = \Sigma$. We give an explicit expression for c_N below. The diagonal elements of the matrix S_1 have a 'standard' prior which we will show implies that, marginally on F, Λ follows a Matrix variate t-distribution (Gupta and Nagar, 2000, p. 134) with degrees of freedom equal to $(T+1-k-n)$, zero mean and $var(vec(\Lambda)) = \underline{M} \otimes I_k \frac{1}{c_\lambda} \frac{1}{(T-n-k-1)}$. The priors for U_1, S_1 and V_1 are all proper. The term c_λ is included to permit shrinkage of Λ towards

zero or a more diffuse prior if desired.

2.1 Parameter expansions

We introduce the nonidentified parameters by two parameter expansions. From the first expansion, we obtain a normal form for the loading matrix and the second expansion results in a normal prior for the factors. These expansions do not affect the proper, independent priors for U, S and V, uniform priors for U and V, and the standard prior for S.

Map from the SVD parameters $(U_1, S_1. V_1)$ to the expanded parameters (U_1, Λ^*) by introducing the orthogonal matrix $C \in O(k)$ via the transformation

$$U_1 S_1 V_1' = U_1 C' C S_1 V_1' = U \Lambda^*$$

$$U_1 C' = U, \qquad C S_1 V_1' = \Lambda^*. \tag{3}$$

The expanding parameter C is given a uniform distribution on $O(k) : (C'dC)$. The transformation $\Lambda^* = C S_1 V_1'$ is a singular value decomposition of Λ^* such that Jacobian of this transformation can be calculated using results in, for example, James (1954, p.71). The transformation of measures for (??) is $\left(U_1' dU_1\right) = (U' dU)$, such that the prior now becomes

$$\frac{f(S_1)(dS_1)\left(U_1' dU_1\right)\left(V_1' dV_1\right)(C' dC)}{c_N c_U} = \frac{p^*(\Lambda^*)(d\Lambda^*)(U' dU)}{c_N c_U}$$

$$p^*(\Lambda^*) = \exp\left\{-\frac{c_\lambda}{2} tr\left(\underline{M}^{-1} \Lambda^{*'} \Lambda^*\right)\right\},$$

$$c_N = \int p^*(\Lambda^*)(d\Lambda^*) = \int f(S_1)(dS_1)(V_1' dV_1)(C' dC) = \left(\frac{2\pi}{c_\lambda}\right)^{\frac{nk}{2}} \left|\underline{M}\right|^{\frac{k}{2}},$$

and so $\Lambda^* = C S_1 V_1'$ has a normal prior distribution such that $p^*(\Lambda^*)$ has a form proportional to the density of a zero mean normal distribution with covariance matrix $\frac{1}{c_\lambda}(\underline{M} \otimes I_k)$. The semi-orthogonal matrix U has a uniform distribution over $V_{k,T}$. In the new parameterization, $F\Lambda = U\Lambda^*$, the matrix Λ^* is has a 'nice' form and prior but U is restricted to be semi-orthogonal, and so it would be

difficult to obtain draws from the posterior.

To give the parameters a more computationally convenient form and prior distributions, we transform from (U, Λ^*) to (F, Λ) via the second parameter expansion. Introduce the $k \times k$ rank k matrix κ with $\frac{k(k+1)}{2}$ free parameters. κ may be, for example, lower triangular or symmetric. This matrix is used to obtain the following transformations:

$$U\Lambda^* = U\kappa\kappa^{-1}\Lambda^* = F\Lambda$$

where $F = U\kappa$ and $\Lambda = \kappa^{-1}\Lambda^*$. It is easier to work with the transformation $A = \kappa'\kappa = F'F$ and write the Jacobian of the bijective transformation from (A, U, Λ^*) to (A, F, Λ) (e.g. Muirhead (1982), p. 58, 66) as

$$\frac{p^* \left(\Lambda^*\right) (dA) \left(U'dU\right) (d\Lambda^*)}{c_N c_U} = \frac{p \left(\Lambda, F\right) J \left(F\right) (d\Lambda) (dF)}{c_N c_U}$$

$$J \left(F\right) = 2^k \left|F'F\right|^{-(T-n-k-1)/2}, \qquad p \left(\Lambda, F\right) = \exp\left\{-\frac{c_\lambda}{2}tr\left(\underline{M}^{-1}\Lambda'F'F\Lambda\right)\right\}$$

Clearly the presence of the determinant $\left|F'F\right|$ in the above Jacobian would complicate computation, particularly as we prefer to have a more convenient form such as a normal distribution for F. Fortunately, we are free to choose the distribution of A and so we let this matrix have a Wishart Distribution with degrees of freedom such that the prior for A is proportional to

$$\exp\left\{-\frac{1}{2}tr\left(A\right)\right\} \left|A\right|^{(T-n-k-1)/2} = \exp\left\{-\frac{1}{2}tr\left(F'F\right)\right\} \left|F'F\right|^{(T-n-k-1)/2}.$$

When we introduce this into the full prior we obtain the following expression of the measure

$$\frac{\left|A\right|^{(T-n-k-1)/2} \exp\left\{-\frac{1}{2}tr\left(A\right)\right\} p^* \left(\Lambda^*\right) (dA) \left(U'dU\right) (d\Lambda^*)}{c_N c_U c_A} \tag{4}$$

$$= \exp\left\{-\frac{1}{2}tr\left(F'F\right)\right\} \left|F'F\right|^{(T-n-k-1)/2} p \left(\Lambda, F\right) \left|F'F\right|^{-(T-n-k-1)/2} (d\Lambda) (dF) c$$

$$= \exp\left\{-\frac{1}{2}tr\left(F'F\right)\right\} p \left(\Lambda, F\right) (d\Lambda) (dF) c$$

where $c = \frac{2^k}{c_N c_U c_A}$ and define c_A as

$$c_A = \int |A|^{(T-n-k-1)/2} \exp\left\{-\frac{1}{2}tr\,(A)\right\}(dA) = 2^{(T-n)k/2}\Gamma_k\left(\frac{T-n}{2}\right).$$

We can summarize the transformations used to this point as

$$\Pi = U_1 S_1 V_1' = U_1 C' C S_1 V_1' = U\Lambda^* = U\kappa\kappa^{-1}\Lambda^* = F\Lambda \tag{5}$$

The resulting joint prior distribution for F and Λ is given by

$$p(\Lambda, F)(d\Lambda)(dF) = \exp\left\{-\frac{1}{2}tr\,(F'F)\right\}\exp\left\{-\frac{c_\lambda}{2}tr\left(\underline{M}^{-1}\Lambda'F'F\Lambda\right)\right\}(d\Lambda)(dF)\,c. \tag{6}$$

Recall that are (Λ, F) are unrestricted matrices and note also that $p(\Lambda, QF) = p(\Lambda, F)$ for any orthogonal matrix Q, which confirms that the prior for the space of U_1 is uniform. Integrating (**??**) with respect to F we get that the marginal prior for Λ is a matrix variate t-distribution (e.g. Gupta and Nagar, 2000, p. 134) with zero mean and $var(vec(\Lambda)) = \underline{M} \otimes I_k \frac{1}{c_\lambda} \frac{1}{(T-n-k-1)}$. The resulting conditional priors have convenient normal forms such that they will be conjugate with the usual specification for the model for y. That is, the conditional prior for $\lambda = vec\,(\Lambda)|F$ is normal with zero mean and covariance matrix $\underline{V}_\lambda = \underline{M}\frac{1}{c_\lambda} \otimes (F'F)^{-1}$. The conditional prior for $f = vec\,(F)|\Lambda$ is normal with zero mean and covariance matrix $\underline{V}_F = \left[I_k + c_\lambda\Lambda\underline{M}^{-1}\Lambda'\right]^{-1} \otimes I_T$. For the chosen transformations and distributions for the unidentified parameters, C and κ, we have results from Liu and Wu (1999) ensuring the sampler will converge. Specifically, the transformations we use form locally compact groups and the priors for the expanding parameters correspond to Haar measures. Further, the expanding parameters are independent of the identified parameters.

Parameter expansions have been used in earlier work in factor models to produce more efficient and simple sampling schemes (see, for a recent example, Ghosh and Dunson, 2009) and to accelerate the EM algorithm in factor models (Liu, Rubin and Wu, 1998, Ročková and George, 2015). The approach in this paper is an application of that developed in Liu (1994) and Liu and Wu (1999) and shares some of the features of Ghosh and Dunson (2009). However, a contribution of this paper that distinguishes it from this earlier work is to use parameter expansion to also achieve invariant inference. This builds upon earlier work on estimation of reduced rank models

(Koop, Léon-González and Strachan (2010) & (2012)) which is natural as the factor model can be represented as a particular type of reduced rank regression model.

2.2 Posterior Computations

In this section we extend the model to allow exogenous variables, provide priors for the other parameters in the model and discuss approaches to computing k. The static factor model is often specified with m exogenous variables collected into the $(T \times m)$ matrix X. After the parameter expansions in the previous section we obtained the matrix of factors F and the loading matrix Λ. The model can now be written as

$$Y = X\beta + F\Lambda + \varepsilon, \quad \varepsilon \sim N(0, \Sigma \otimes I_T) \tag{7}$$

$$f = vec(F)|\lambda \sim N\left(0, \underline{V}_F\right), \quad \lambda = vec(\Lambda)|f \sim N\left(0, \underline{V}_\lambda\right), \tag{8}$$

$$b = vec(\beta) \sim N\left(0, \Sigma \otimes (X'X)^{-1} \frac{1}{c_\beta}\right) \tag{9}$$

where $\underline{V}_F = \left[I_k + c_\lambda \Lambda \underline{M}^{-1} \Lambda'\right]^{-1} \otimes I_T$ and $\underline{V}_\lambda = \underline{M}\frac{1}{c_\lambda} \otimes (F'F)^{-1}$. We assume each diagonal element of $\Sigma = diag\left\{\sigma_i^2\right\}$ has an inverse gamma prior

$$p\left(\sigma_i^2\right) \propto \left(\sigma_i^2\right)^{-\frac{\nu+2}{2}} \exp\left\{-\frac{\nu}{2\sigma_i^2 \underline{\mu}_i}\right\}\left(d\sigma_i^2\right)$$

such that

$$p(\Sigma) \propto |\Sigma|^{-\frac{\nu+2}{2}} \exp\left\{-\frac{\nu}{2}tr\left(\Sigma^{-1}\underline{\Omega}\right)\right\}(d\Sigma) \tag{10}$$

where $\underline{\Omega} = diag\left\{\frac{1}{\underline{\mu}_1}, \frac{1}{\underline{\mu}_2}, \ldots, \frac{1}{\underline{\mu}_n}\right\}$. To preserve scale invariance requires that the prior for Σ also be scale invariant. One option is to use the Jeffrey's prior and set $\underline{\nu} = 0$ in (??). Another approach is to set each $\underline{\mu}_i$ to be some function of the scale of the data. For example, we could set $\underline{\mu}_i = \frac{1}{\widehat{\sigma}_i^2}$ where $\widehat{\sigma}_i^2$ is the sample variance of y_i. By setting $\underline{\Omega} = diag\left\{\widehat{\sigma}_1^2, \widehat{\sigma}_2^2, \ldots, \widehat{\sigma}_n^2\right\}$, the prior in (??) permits both of these options. The conditional posteriors $f|\lambda, b, \Sigma$, $\lambda|f, b, \Sigma$ and $b|f, \lambda, \Sigma$ are normal while the

elements on the diagonal of $\Sigma | f, \lambda, b$ are inverse gamma distributions. The precise form of these conditional posteriors are given in the Appendix.

One important question in factor models is the number of factors, k. The model with no factors occurs at the point $\Lambda = 0$ and at this point the factors are excluded from the likelihood. Therefore we are able to use the Savage-Dickey density ratio (SDDR) to compute the Bayes factors for $k = 0$ to $k = k^*$, B_{0,k^*}, as (Verdinelli and Wasserman (1995)):

$$B_{0,k^*} = \frac{m_0}{m_{k^*}} = \frac{p(\Lambda = 0 | y)}{p(\Lambda = 0)} c_{VW}$$

where m_{k^*} is the marginal likelihood for the model with k^* factors, $p(\Lambda = 0 | y)$ is the marginal posterior of Λ at the point $\Lambda = 0$, $p(\Lambda = 0)$ is the marginal prior for Λ evaluated at the same point, and c_{VW} is the correction factor proposed by Verdinelli and Wasserman (1995). When $\underline{M} = I_n$ the correction factor c_{VW} is equal to one, and the marginal prior of Λ is a matrix variate t-distribution so that the value of the prior ordinate can be calculated as $p(\Lambda = 0) = c(2\pi)^{Tk/2}$, where c was defined next to expression (**??**). When $\underline{M} = \Sigma$, the conditional prior of Σ given $\Lambda = 0$ depends on k, and therefore the correction factor c_{VW} becomes different from one. As shown in the Supplemental material C, when $\underline{M} = \Sigma$ the ratio $c_{VW} / p(\Lambda = 0)$ is given by:

$$\frac{c_{VW}}{p(\Lambda = 0)} = \left(\frac{\pi}{c_\lambda}\right)^{nk/2} \frac{\Gamma_k\left(\frac{T-n}{2}\right)}{\Gamma_k\left(\frac{T}{2}\right)} \left(\frac{\Gamma\left(\frac{\underline{\nu}+T}{2}\right)}{\Gamma\left(\frac{\underline{\nu}+k+T}{2}\right)}\right)^n \prod_{i=1}^n \left(\frac{\widetilde{h}_{ii}}{2}\right)^{\frac{k}{2}} \tag{11}$$

where $\widetilde{h}_{ii}$ is the ith diagonal element of $\widetilde{H} = \underline{\nu}(\underline{\Omega}) + (Y - X\widehat{\beta})'(Y - X\widehat{\beta})$, and $\widehat{\beta} = (1 - \sqrt{c_\beta / (1 + c_\beta)})(X'X)^{-1}X'Y$. With a sequence of G draws from the posterior, we can compute the conditional posterior $p(\Lambda | \beta, F, \Sigma, y)$ at $\Lambda = 0$ to estimate the required ratio as:

$$\widehat{B}_{0,k^*} = \frac{\frac{1}{G}\Sigma_{i=1}^G p\left(\Lambda = 0 | \beta^{(i)}, F^{(i)}, \Sigma^{(i)}, y\right)}{p(\Lambda = 0)} c_{VW},$$

where $i = 1, .., G$ indicates the draws from the posterior.

Note that we are able to use the SDDR because the point $\Lambda = 0$ belongs to the parameter space. The SDDR however cannot be used in the context of most previous literature because the

identifying restrictions often imposed on Λ imply that $\Lambda = 0$ is no longer a point in the parameter space. Conversely, although the method of Chib (1995) can be used to calculate marginal likelihoods in the non-invariant specifications, it is less suited to our specification because the accuracy of the method relies on being able to estimate the posterior density accurately at a point of high posterior density. This task is slightly more difficult in our context because we have introduced non-identified parameters, which makes the augmented posterior density more disperse around the mode. However, the SDDR method has the advantage that it does not require further calculations beyond the basic estimation algorithm. The accuracy of both methods to calculate marginal likelihoods decreases with the dimension of y and the number of factors. In such situations one can use alternative methods such as that of Chan and Eisenstat (2015) or calculate predictive likelihoods (Geweke and Amisano, 2010), as we illustrate in the empirical applications of Section 4.

In the following section we provide the technical details for the several contributions of this paper. The reader who prefers not to read the technical details in Section 3 and interested only in applying the approach may prefer to skip to Section 4.

3   An explanation for non-invariance with discussion

In this section we provide an explanation for non-invariance (contribution iii from the introduction). We provide an informal explanation for the invariance followed by a theorem, the proof for which we leave to the paper's Supplemental material D. To support this discussion, we describe existing invariant specifications closely related to the one we propose. We then outline some standard non-invariant identification schemes that have been used in Bayesian analysis of factor models to indicate the source of the non-invariance.

The cause of the lack of invariance can be explained informally as follows. In the support for the loading matrix, Λ, are points where subsets of columns form singular matrices. The collections of these points form manifolds in the support. Changing the order of the variables in y_t induces a

transformation, g, of the loading matrix. These points of singularity pose no problems for the transformation for the invariant specification as g is smooth, continuous and homoemorphic in this case. From group theory, we know these conditions ensure invariance. For the non-invariant specifications, the transformation is discontinuous and so non-homeomorphic at these points of singularity, thus losing invariance. We demonstrate the importance of this discontinuity for inference on the number of factors with a simulation in the Supplemental material A.

The frequentist approach using principle components (as in Bai and Ng (2002)) is an invariant specification very close to ours and that of Hoff (2007). In the frequentist approach, we take a singular value decomposition of the $T \times n$ data matrix y and the factors may be associated with the k eigenvectors associated with the k largest eigenvalues of yy'. Bayesian computation of this model is not straightforward, but achieved in this paper and Hoff (2007) in very different approaches. The existence of an identified and invariant specificatoin leads to the conclusion that non-invariance is not a problem of identification. Rather, it is a specification problem.

In this section we also show how to map from the principal components specification to the other model specifications used in the literature thereby demonstrating how our specification can permit any of the others.

Noninvariant specifications: In an identification scheme used, for example, in Geweke and Zhou (1996), $y_t = f_t^+ \Lambda^+ + \varepsilon_t$, where $f_t^{+\prime} \sim N(0, I)$, $\Lambda^+ = \begin{bmatrix} \Lambda_1^+ & \Lambda_2 \end{bmatrix}$, the $k \times k$ matrix Λ_1^+ is restricted to be upper triangular with positive elements on the diagonal and Λ_2 is an $k \times (n - k)$ unrestricted matrix. The first factor is identified as it is the only factor entering the first variable. The next factor is identified as, besides the already identified first factor, it is the only factor entering the second variable, and so on. This structure requires the researcher to assume some knowledge about the order of the variables with respect to the factors which imply the first k rows of the loading matrix are linearly independent.

We can readily transform to other non-invariant specifications (and back). For example, collect the diagonal elements of Λ_1^+ into the diagonal of an $k \times k$ matrix Λ_d^+. Next, define $f_t^1 = f_t^+ \Lambda_d^+$ and

$\Lambda^1 = \left(\Lambda_d^+\right)^{-1}\Lambda^+$ and $\Lambda_1^1 = \left(\Lambda_d^+\right)^{-1}\Lambda_1^+$ such that $f_t^+\Lambda^+ = f_t^+\Lambda_d\Lambda_d^{-1}\Lambda^+ = f^1\Lambda^1$ and let $\Omega^1 = \Lambda_d^{+\prime}\Lambda_d^+$ be a diagonal covariance matrix for the errors in the factor process. In this identification scheme, Λ^1 is upper triangular with ones on the diagonal. This specification has been used in, for example, Chib, Nardari and Shephard (2006) and Aguilar and West (2000) and requires similar assumptions about the orders of the variables as the previous specification.

Denote by i a particular ordering of the variables. We will use the notation $\mathfrak{T}^{+,i}$ and $\mathfrak{T}^{1,i}$ for the class of non-invariant identification schemes resulting from the restrictions to Λ^+ and Λ^1 respectively for order i.

The identification issues with this model are well understood. Because Σ is diagonal, the unrestricted Λ, Ω and Σ contain a total of $nk + n + k(k+1)/2$ parameters which can exceed the $n(n+1)/2$ parameters in $E(y_t'y_t)$. Even if we assume $\Omega = I_k$, we must restrict $k \le (n-1)/2$ as otherwise there would not be enough information to estimate all of the parameters (Geweke and Zhou (1996)). There remains a second identification problem as we can rotate f_t and obtain an observationally equivalent vector f_t^*. That is, if $U \in O(k) \equiv \{U : U(k \times k), U'U = I_k\}$ the orthogonal group, then rotate the parameters by $f_t^* = f_t U$ and $\Lambda^* = U'\Lambda$ and we can see that the rotated parameters (f_t^*, Λ^*) and (f_t, Λ) are not distinguishable since they both enter the likelihood as products: $f_t^*\Lambda^* = f_t\Lambda$. For this reason restrictions are imposed upon Λ and Ω to permit estimation. Note that this argument uses a specification of f_t and Λ that is overidentified and this permits the observationally equivalent rotations. For the following discussion it is useful to denote the first k columns of Λ by Λ_1. Generally we will take the matrix Λ_1 to represent the columns that are restricted to permit identification of the factors and these may not always be the first k columns.

Notice that these specifications have the same dimension of parameter space as Σ is the same, f_t is always $1 \times k$, and the free elements of $[\Lambda^+, \Omega^+]$ and $\left[\Lambda^1, \Omega^1\right]$ all have dimension $(n-k)k + \frac{k(k+1)}{2} = nk - \frac{k(k-1)}{2}$. With all f_t, the total number of free parameters is then $(T+n)k - \frac{k(k-1)}{2}$. These non-invariant identification schemes (and priors) used in Bayesian approaches result in specifications that have more parameters than can be identified from the likelihood. Thus they are still not

identified under the likelihood, but are identified under the posterior due to the informative priors on the overidentified parameters.

The source of the lack of invariance: A drawback of these specifications is that they assume Λ_1 is nonsingular which has implications for evidence on alternative values for *k*. Reordering the variables involves groups of transformations with the discontinuity mentioned above occuring at the point $|\Lambda_1| = 0$. The simulation study in the Supplemental material A illustrates the issue and demonstrates that this effect is neither ubiquitous nor a measure zero event. That is, the effect does not occur at all points of the parameter space but nor is it only relevant at the point $|\Lambda_1| = 0$. In particular, when $|\Lambda_1|$ is relatively close to 0, the ordering assumption can have a large influence, up to a tenfold impact in our simulations, on the marginal likelihood. Further, the numerical standard error of the estimated marginal likelihood substantially increases, up to 40 times in our simulations, when the ordering implicitly assumes that $|\Lambda_1| \neq 0$ but in the data generating process $|\Lambda_1|$ is close to 0. Choosing correctly the number of factors is also important for estimating the variance decomposition accurately, as we illustrate in the Supplemental material E. Given the above evidence, it would be sensible to choose the ordering in which the posterior mass for $|\Lambda_1|$ is far from zero. Unfortunately, this point cannot be known a priori which brings us back to the solutions of either averaging over all orderings or devising an invariant specification.

The non-invariant identification scheme $\mathfrak{T}$ impose an order on the variables.

Definition 1 *Denote by $\mathfrak{T}^i$ ordering i of the variables and denote a different ordering j by $\mathfrak{T}^j$.*

Definition 2 *In the ordering $\mathfrak{T}^i$ the identification of the factors is achieved by placing restrictions on the submatrix Λ_1^i.*

Theorem 3 (Discontinuity) *The transformation from $\mathfrak{T}^i$ to $\mathfrak{T}^j$ has a discontinuity at $\left|\Lambda_1^j\right| = 0$.*

Proof. See Supplemental material D. ∎

To demonstrate this discontinuity with a simple example, consider a (2×1) vector $\Lambda = (\cos(\theta), \sin(\theta)) = (\lambda_1, \lambda_2)$ where $\theta \in [0, \pi)$. The non-invariant specification discussed above implies $\widehat{\Lambda} = \left(1, \widehat{\lambda_2}\right)$ where $\widehat{\lambda_2} = \tan(\theta)$ under one ordering, $\mathfrak{T}^1$, with support $(-\infty, \infty)$ which clearly includes $\widehat{\lambda_2} = 0$. Changing the order of the variables we obtain $\widetilde{\Lambda} = \left(\widetilde{\lambda_1}, 1\right)$ where $\widetilde{\lambda_1} = \frac{1}{\tan(\theta)} = \frac{1}{\lambda_2}$ is ordering $\mathfrak{T}^2$ which now excludes the point $\widehat{\lambda_2} = 0$. In this case there is a discontinuity from $\mathfrak{T}^1$ to $\mathfrak{T}^2$ at $\widehat{\lambda_2} = 0$.

The discontinuity induced by the non-invariance also manifests in transforming from θ to $\widehat{\lambda_2}$. This transformation maps from the support $[0, \pi)$ to $(-\infty, \infty)$. Transforming back from $\widehat{\lambda_2}$ to θ, however, transforms from the support $(-\infty, \infty)$ to $[0, \frac{\pi}{2}) \cap (\frac{\pi}{2}, \pi]$. The point $\theta = \frac{\pi}{2}$ is excluded. In other words, there is a discontinuity from $\mathfrak{T}$ to $G_{1,1}$. If, however, we used the specification $\Lambda = (\cos(\theta), \sin(\theta)) = (\lambda_1, \lambda_2)$ where $\theta \in [0, \pi)$, then in this simple case the support of the angle θ maps homeomorphically to the one dimensional Grassmann manifold, $G_{1,1}$, and changing the order of the variables maps from $G_{1,1}$ to $G_{1,1}$ without discontinuities.

More generally, transforming from $sp(\Lambda')$ to Λ^+ (or to Λ^1) does not preserve the topology as the deformation 'punches holes' in the form at points where $|\Lambda_1| = 0$. The transformation from the space to Λ^+ is discontinuous at this point. Similarly, transforming between different orderings, changing between Λ_1 and $\widetilde{\Lambda}_1$, involves discontinuities at $|\Lambda_1| = 0$. Without preservation of the topology, the metric and therefore measure are not preserved hence the evidence on k under alternative normalizations need not be the same. This effect is naturally most pronounced when the true parameter values are near the point of discontinuity. As our small simulation exercise in Supplementary material A showed, if we move far enough away from this point the effect is mitigated.

The invariant specification used in this paper is closely related to that used in the method of principal components (pc) to estimate the factors as in, for example, Bai and Ng (2002) (hereafter BN). This specification is also invariant to ordering. Denote this model by

$$y_t = f_t^{pc} \Lambda^{pc} + \varepsilon_t.$$

Stacking the observations over T we obtain

$$y = F^{pc}\Lambda^{pc} + \varepsilon$$

where $y = \left(y_1', y_2', \ldots, y_T'\right)'$, $\varepsilon = \left(\varepsilon_1', \varepsilon_2', \ldots, \varepsilon_T'\right)'$ and $F^{pc} = \left(f_1^{pc\prime}, f_2^{pc\prime}, \ldots, f_T^{pc\prime}\right)'$.

In BN, the matrix F^{pc} is estimated as proportional to the matrix of eigenvectors[3] associated with the k largest eigenvalues of the matrix yy'. In other words, they take a SVD of y as

$$y = USV' = U_1 S_1 V_1' + U_2 S_2 V_2'$$

where $U = [U_1 \quad U_2] \in O(T)$ are the eigenvectors of $yy' = US^2U'$ and $V = [V_1 \quad V_2] \in O(n)$ are the eigenvectors of $y'y = VS^2V'$ and $U'U = V'V = I_n$. Setting the factors and loading matrices to $F^{pc} = U_1$ and $\Lambda^{pc} = S_1 V_1'$, then Λ^{pc} has the unusual property of being orthogonal (but not orthonormal) as $\Lambda^{pc}\Lambda^{pc\prime} = S_1^2$. The orthonormal structure of F^{pc} and orthogonal structure of Λ^{pc} imply that the parameters are identified up to sign. That is, $F^{pc}\Lambda^{pc} = F^{pc}\kappa\kappa\Lambda^{pc} = F^{pc,*}\Lambda^{pc,*}$ where F^{pc} and $F^{pc,*}$, and Λ^{pc} and $\Lambda^{pc,*}$ will have the same structure only for a diagonal matrix κ with the ith diagonal element equal to +1 or -1.[4] This lack of identification is resolved by fixing the sign of, say, the first row of U to be non-negative (but not simply positive). Such a restriction gives a particular orientation of the vectors in U in their space, but in no way restricts the space they span and, as the transformation involving κ is homeomorphic, we can regard U, S and V as identified for any practical purposes and, in particular, our purpose. As there exists a specification that is both identified and invariant to ordering, this raises the question as to whether there is an identification problem. Rather, it appears there is a specification problem that induces a lack of invariance.

In a working paper version of this paper, we show how one can move from one non-invariant specification to another via simple transformations. The existence of this mapping is important

---

[3]In fact, Bai and Ng use $\sqrt{T}$ times the eigenvalue of yy'. This proportional term is not important for the discussion here so we ignore it.

[4]If κ were not of this structure then the restriction $\Lambda^{pc}\Lambda^{pc\prime} = S_1^2$ would be destroyed by the transformation. Thus this restriction implies identification against such a transformation.

as we will show how we can map from our invariant specification to one of those in this section, which implies we can then map to both specifications from ours. The transformation from one non-invariant specification to another, with a given ordering of the variables, is possible because all of the specifications impose the same restriction on the parameter space. That is, they all assume that a particular set of k columns are linearly independent. This assumption does not hold at all points in the parameters space when changing the order of the variables so it is not possible to map homeomorphically between specifications with different variable orderings.

4 Empirical applications

In this section we present two empirical examples to demonstrate the effect of lack of invariance of other approaches and that our approach achieves invariance. We also present evidence on the efficiency of the sampling algorithm.

4.1 Currency Exchange Rates

We use data on international currency exchange rates relative to U.S. dollar over a time period of 1045 business days beginning in January 2007 and ending in December 2010. The returns are computed as $y_{it} = 100(p_{it}/p_{i,t-1} - 1)$, where p_{it} denotes the daily closing spot rate. We first use six exchange rates and at the end of this example we expand the dataset to 20 exchange rates. The six series we analyze first are the Australian Dollar (AUD), Euro (EURO), South Korean Won (KRW), Japanese Yen (JPY), Canadian Dollar (CAD), and British Pound (GBP). These represent some of the most heavily traded currencies over the period. The returns are plotted in Figure **??**.

We fit the data using the static factor model in (**??**) where y_t is the vector of six observed currency returns, $f_t \sim iidN(0, I_k)$ is a $1 \times k$ vector of unobserved factors, Λ is an $k \times n$ matrix of factor loadings. We first impose the identification assumption that Λ is upper triangular with positive diagonal elements. We then compute the marginal likelihoods for four models. Using the

ordering (AUD, EUR, KRW, JPY, CAD, GBP), we compute a single-factor model and then a two-factor model. Next, with the ordering (AUD, KRW, EUR, JPY, CAD, GBP), we again compute a single-factor model and then a two-factor model.

The log marginal likelihoods for the non-invariant specifications are computed via the method of Chib (1995) using 100 parallel chains each of length 50000. The results are reported in Table 1. As the two marginal likelihoods for the models with one factor are almost the same (i.e., ordering made no difference) we only report one of these. The computed marginal likelihoods for the two-factors provide striking evidence of the effect of reordering. The log marginal likelihoods differ by about 142.

Under the ordering (AUD, KRW, EUR, JPY, CAD, GBP) there is a very strong preference for the two factor model over the one factor model with a log Bayes factor of -63.6. However, under the ordering (AUD, EUR, KRW, JPY, CAD, GBP), there is a very strong preference for the one factor model with the log Bayes factor of 78.3. The reordering of the variables has shifted the evidence on the number of factors in the opposite direction. The invariant specification selects a model with two factors over the one factor model with a log Bayes factor of 645 (for $\underline{M} = \Sigma$). The evidence for two factors is therefore overwhelming.

Figure **??** reports the inefficiency factors for blocks of parameters from the model estimated with the Geweke and Zhou specification (*GZ*) and the parameter expanded invariant model (*PX*). The inefficiency measures give an estimate of the number of draws needed to have as much information about the posterior as we would obtain from one independent draw. The smaller the inefficiency factor the better is the sampler. It is clear that the parameters are generally much more efficiently estimated using the expanded parameter specification. These results are consistent with those found in Ghosh and Dunson (2009). The distributions of both the loading matrix, Λ, and the factors, F, are less disperse and lower for the parameter expanded model, but this is also true for the idiosyncratic variances, Σ, and the exogenous variables coefficients, β.

To illustrate how the method performs in higher dimensions, we add to y 14 additional currencies, so that $n = 20$. In order to control for possible serial correlation, we use the model with exogenous variables in (**??**) with the intercept term and one lag of y_t in X_t, so that $m = 21$ and $T = 1043$. Because the SDDR has a high variance in high dimensions, we calculate the marginal likelihood using the adaptive importance sampling method proposed in Chan and Eisenstat (2015). Specifically, an importance sampling density for (β, Λ, Σ) is first obtained by approximating the joint posterior density. Then we compute the importance sampling estimate using the integrated likelihood, that is, the conditional density of the data marginal of F, which is available analytically. For each factor model, 10000 posterior draws are used to compute the importance sampling density. Then, 100000 importance sampling draws are obtained to calculate the marginal likelihood. Table 2 reports log marginal likelihoods and numerical standard errors, showing that the marginal likelihood improves up to the model with 5 factors and then decreases, indicating that the model with 5 factors is the most adequate.

4.2 The Number of Factors Driving US Macroeconomic Indicators

The dataset is obtained from Stock and Watson (2009), which consists of 190 quarterly observations from 1959Q3 to 2006Q4 on $n = 109$ macroeconomic variables. Stock and Watson (2009) provide a detailed list of the data and its transformation in terms of logs and differencing. The dataset includes variables on GDP, industrial production, capacity utilization, purchasing manager's indices, labor force statistics, housing starts, consumer prices, commodity prices, average hourly earnings, productivity, interest rates, yield spreads, exchange rates, stock prices, money bases, business loans, consumer credit and consumer expectations. The data are first standardized.

In Table 3 we report the sums of log predictive likelihoods for the last 10 years using a k-factor model, with $k = 1, \dots, 7$. Alhough Stock and Watson (2009) used a dynamic factor model, here for simplicity we use the static factor model of Section 2. Predictive likelihoods are the one-step ahead

predictive density evaluated at the realized outcome (see Geweke and Amisano 2010 or Geweke, 1996). Due to the large dimension of y_t, predictive likelihoods can be calculated more accurately than the SDDR, especially when parallel computing is available. The best performing model is the one with $k = 6$, which gains as much as 25 points in the log scale with respect to the second best model ($k = 7$). This result is within the range suggested by Stock and Watson (2009), who used the method of Bai and Ng (2002) and found the rank to be between 2 and 10 depending on the criterion used.

5   Concluding remarks

In this paper, we propose a specification for the static factor model that requires no ordering restrictions and so the choice of number of factors cannot depend upon the chosen ordering. By augmenting the posterior with a number of unidentified parameters with appropriate priors, the model can be computed using standard distributions and the draws are relatively efficient.

The specification we propose nests many of the existing and popular specifications used in factor analysis. Thus each of these specifications are attainable directly from the output from estimating our specification.

Although for convenience we have only considered the static factor model, this approach is readily extended to allow dynamics in the state equation. Such an extension would involve using an informative prior on the space $sp(F)$ such as the orthogonal projective Gaussian distribution as used in Koop, Léon- González and Strachan (2011). This would involve transforming from F to $F_c = RF$ where the matrix R captures the dynamics in F_c. For example, if the rows of F follow a

random walk as $f_{c,t} = f_{c,t-1} + f_t = \Sigma_{i=1}^{t} f_i$ then we would define R as

$$R = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}.$$

Alternatively $R = R(\rho)$ may be a function of parameters to allow richer dynamics such as in autoregressive processes. For example, an $AR(1)$ state equation $f_{c,t} = \rho f_{c,t} + f_t = \Sigma_{i=1}^{t} \rho^{t-i} f_i$ and so in which case we define R as

$$R(\rho) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \rho & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ \rho^{t-1} & \rho^{t-2} & \cdots & 1 \end{bmatrix}.$$

The full matrix of factors then becomes $F_c = R(\rho) F$ which implies a prior for a process with a zero mean process with covariance matrix $(I_k + c_\lambda \Lambda \Lambda')^{-1} \otimes (R(\rho)' R(\rho))^{-1}$.

Another implication of the invariant specification, which we have exploited in this paper, is that we are able to compute the Bayes factors for the number of factors using the Savage-Dickey density ratio. This approach requires only the conditional posterior and the conditional prior for Λ. This greatly simplifies the computation of the posterior probabilities. This approach cannot be used in the non-invariant specifications as they exclude the point $\Lambda = 0$ from the support of the loading matrix parameter.

In computing the models in this paper, it became evident that there is a relationship between the computational efficiency and accuracy of marginal likelihood estimates, and the proximity of the posterior to the point of discontinuity. Models that are specified such that the posterior is invariant to reordering tend to have lower numerical standard errors. The accuracy of estimation of the marginal likelihood plays an important role in the confidence we have in the conclusions we make.

This relationship is a topic of current research.

6 References

Aßmann, Boysen-Hogrefe, and Pape (2012) "The Directional Identification Problem in Bayesian Factor Analysis: An Ex-Post Approach" Working paper 2012-11, Department of Economics, Christian-Albrechts-Universität Kiel, Germany

Aguilar, O. and M. West (2000) "Bayesian Dynamic Factor Models and Portfolio Allocation," *Journal of Business & Economic Statistics*, 18, 338-357.

Amari, S. (1985). *Differential-Geometrical Methods in Statistics*. Lecture Notes in Statistics, 28, Springer, New York.

Bai, J. and Ng, S. (2002) "Determining the number of factors in approximate factor models," *Econometrica,* Vol. 70, No. 1, 191-221.

Bhattacharya A. and D. B Dunson (2011) "Sparse Bayesian infinite factor models," *Biometrika*, 98, 291–306.

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

Chib. S. (2001), "Markov Chain Monte Carlo Methods: Computation and Inference," in *Handbook of Econometrics*, J. J. Heckman and E. E. Learner (Eds.), Vol. 5, 3570-3649, Amsterdam, North-Holland.

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

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

Chikuse, Y. (2003). *Statistics on Special Manifolds*. Springer, New York.

Connor, G. and R. Korajczyk (1986) "Performance measurement with the arbitrage pricing theory: A new framework for analysis," *Journal of Financial Economics* 15, 373-394.

Forni M., Giannone D., Lippi M. and L. Reichlin (2009) "Opening the Black Box: Structural Factor Models with large cross-sections", *Econometric Theory* 25, 1319-1347.

Foerster, A., P.G. Sarte and M.W. Watson (2011) "Sectoral versus Aggregate Shocks: A Structural Factor Analysis of Industrial Production," *Journal of Political Economy*, 119, 1-38.

Frühwirth-Schnatter, S. and H. F. Lopes (2010) "Parsimonious Bayesian Factor Analysis when the Number of Factors is Unknown" *Technical Report, Chicago Booth*
    http://faculty.chicagobooth.edu/hedibert.lopes/research/reports.html

Geweke, J. (1996) "Bayesian reduced rank regression in econometrics," *Journal of Econometrics* 75, 121-146.

Geweke, J. and G. Amisano (2010), "Hierarchical Markov normal mixture models with applications to financial asset returns", *Journal of Applied Econometrics,* 26, 1-29.

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

Ghosh, J., and Dunson, D.B. (2009) "Default priors and efficient posterior computation in Bayesian factor analysis," *Journal of Computational and Graphical Statistics* 18, 306-320.

Gupta, A.K. and D.K. Nagar (2000) *Matrix Variate Distributions*. Boca Raton: Chapman & Hall/CRC.

Hoff, P.D.(2007) "Model averaging and dimension selection for the singular value decomposition," *Journal of the American Statistical Association*, 102, 674-685.

James, A. T. (1954) "Normal multivariate analysis and the orthogonal group," *Annals of Mathematical Statistics*, 25, 40–75.

Kaufmann and Schumacher (2012) "Bayesian estimation of sparse dynamic factor models with order-independent identification" Working paper 13.04, Study Center Gerzensee, Foundation of the Swiss National Bank.

Koop G., R. Léon- González and R. W. Strachan (2010) "Efficient posterior simulation for cointegrated models with priors on the cointegration space" *Econometric Reviews* Volume 29, Issue 2, 224-242.

Koop G., R. Léon- González and R. W. Strachan (2011) "Bayesian Inference in the Time Varying Cointegration Model," *The Journal of Econometrics* 165, 210-220.

Koop G., R. Léon-Gonzalez and R. W. Strachan (2012) "Bayesian Model Averaging in the Instrumental Variable Regression Model," *The Journal of Econometrics,* 171, 237-250.

Larsson, R. and M. Villani (2001) "A distance measure between cointegration spaces," *Economics Letters*, 70, 21–27.

Lee, P.M. (2012) *Bayesian Statistics: An Introduction*. Wiley, West Sussex.

Liu, J.S. (1994), "The collapsed Gibbs sampler with applications to a gene regulation problem," *Journal of the American Statistical Association* 89, 958-966.

Liu C., D. B. Rubin, Y. N. Wu (1998) "Parameter Expansion to Accelerate EM: The PX-EM Algorithm," *Biometrika*, Vol. 85, No. 4, pp. 755-770

Liu, J. and Wu, Y. (1999) "Parameter expansion for data augmentation," *Journal of the American Statistical Association*, 94, 1264-1274.

Lopes H. F. and M. West (2004) "Bayesian Model Assessment in Factor Analysis," *Statistica Sinica* 14, 41-67.

Magnus, J.R., Neudecker, H. (1988) *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, New York.

Muirhead, R.J. (1982) *Aspects of Multivariate Statistical Theory*. New York: Wiley.

Ročkováa, V. and E.I. George (2015) "Fast Bayesian Factor Analysis via Automatic Rotations to Sparsity," *Journal of the American Statistical Association*, DOI:10.1080/01621459.2015.1100620

Stock J., and M.W. Watson (2009) "Forecasting in Dynamic Factor Models Subject to Structural Instability". In: N. Shephard and J. Castle (eds), *The Methodology and Practice of Econometrics: Festschrift in Honor of D.F. Hendry*, Oxford University Press, 1-57.

Strachan, R. and Inder, B. (2004) "Bayesian analysis of the error correction model," *Journal of Econometrics* 123, 307-325.

Verdinelli, I. and L. Wasserman (1995) "Computing Bayes Factors using a generalization of the Savage-Dickey Density Ratio," *Journal of the American Statistical Association* 90, 614-618.

Villani, M. (2005) "Bayesian reference analysis of cointegration," *Econometric Theory* 21, 326-357.

Appendix: Conditional Posterior Densities

The likelihood function can be written as

$$L \propto |\Sigma|^{-T/2} \exp\left[-\frac{1}{2}tr\left(\Sigma^{-1}(Y-X\beta-F\Lambda)'(Y-X\beta-F\Lambda)\right)\right]$$

so that the conditional posteriors can be readily derived. First define

$$a = \begin{bmatrix} vec(\beta) \\ vec(\Lambda) \end{bmatrix} = \begin{bmatrix} b \\ \lambda \end{bmatrix}$$

$$W = [I_n \otimes X \quad I_n \otimes F] = (I_n \otimes \omega),$$

$$\omega = [X \quad F],$$

and $f = vec(F)$. Vectorizing the $T \times n$ error matrix ε gives the useful linear forms for f and a. Let $y = vec(Y)$, then let $x = (I_n \otimes X)$ and $l = (\Lambda' \otimes I_T)$, and define $\widetilde{y} = y - xb$ such that we can write

$$\begin{aligned} vec(Y - X\beta - F\Lambda) &= y - (I_n \otimes X)b - (\Lambda' \otimes I_T)f \\ &= \widetilde{y} - lf \\ &= y - (I_n \otimes X)b - (I_k \otimes F)\lambda \\ &= y - Wa \end{aligned}$$

As the vectors f and a have normal priors and enter the likelihood linearly (conditional on the other parameters) the conditional posteriors result from standard computations. Specifically, the conditional posteriors have the following forms:

$$f|\beta,\Lambda,\Sigma,Y \sim N\left(\overline{f},\overline{V}_F\right), \quad a|F,\Sigma,Y \sim N\left(\overline{a},\overline{V}_a\right), \quad \sigma_i^2|a,F \sim \overline{\mu}_i\chi_v^{-2}$$

where

$$\overline{f} = \overline{V}_F \left(\Lambda \Sigma^{-1} \otimes I_T \right)(y - xb) = vec\left((Y - X\beta) \Sigma^{-1} \Lambda' \left[\Lambda \left(\Sigma^{-1} + c_\lambda \underline{M}^{-1} \right) \Lambda' + I_k \right]^{-1} \right),$$

$$\overline{V}_F = \left[\Lambda \left(\Sigma^{-1} + c_\lambda \underline{M}^{-1} \right) \Lambda' + I_k \right]^{-1} \otimes I_T,$$

$$\overline{a} = \overline{V}_a W' \left(\Sigma^{-1} \otimes I_T \right) y = \overline{V}_a \begin{pmatrix} vec(X'Y\Sigma^{-1}) \\ vec(F'Y\Sigma^{-1}) \end{pmatrix},$$

$$\overline{V}_a = \left[W' \left(\Sigma^{-1} \otimes I_T \right) W + \underline{V}_a^{-1} \right]^{-1} = \begin{bmatrix} \Sigma^{-1} \otimes X'X & \Sigma^{-1} \otimes X'F \\ \Sigma^{-1} \otimes F'X & \Sigma^{-1} \otimes F'F \end{bmatrix} + \underline{V}_a^{-1} \Bigg]^{-1}$$

$$\underline{V}_a = \begin{bmatrix} \Sigma \otimes \left(X'Xc_\beta \right)^{-1} & 0 \\ 0 & \underline{M} \otimes (c_\lambda F'F)^{-1} \end{bmatrix}$$

$$y = vec(Y)$$

$$\overline{\mu}_i = h_{ii}$$

where $\overline{\nu} = T + m + \underline{\nu}$, when $\underline{M} = I_n$ and $\overline{\nu} = T + k + m + \underline{\nu}$, when $\underline{M} = \Sigma$. In the above expression, h_{ii} is the ith diagonal element of H, and $H = \varepsilon'\varepsilon + \underline{\nu}\underline{\Omega} + c_\beta\beta'X'X\beta$ when $\underline{M} = I_n$ and $H = \varepsilon'\varepsilon + \underline{\nu}\underline{\Omega} + c_\beta\beta'X'X\beta + c_\lambda\Lambda'F'F\Lambda$ when $\underline{M} = \Sigma$. The notation $\overline{\mu}_i\chi_{\overline{\nu}}^{-2}$ refers to $\overline{\mu}_i$ times an inverse chi-squared with $\overline{\nu}$ degrees of freedom (e.g. Lee (2012, p.377)).

Table 1: Log marginal likelihoods and the corresponding numerical standard errors for the competing models.

Model	log marginal likelihood	numerical standard error
1-factor (AUD, KRW, EUR, JPY, CAD, GBP)	-7572.9	3.40
2-factor (AUD, EUR, KRW, JPY, CAD, GBP)	-7636.5	5.64
2-factor (AUD, KRW, EUR, JPY, CAD, GBP)	-7494.6	1.16

Table 2: Logs of Marginal Likelihood with 20 exchange rates and Numerical Standard Errors.

k	1	2	3	4	5	6	7
Log ML	-19092	-18516	-18397	-18335	-18324	-18367	-18412
NSE	0.08	0.15	0.16	1.24	3.02	4.97	3.86

Table 3: Sums of predictive likelihoods in the k-factor model.

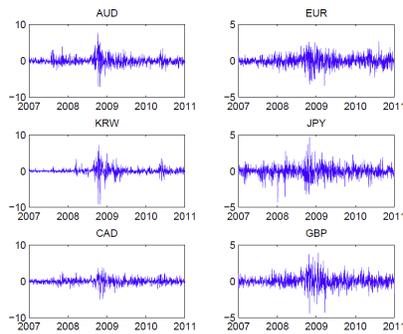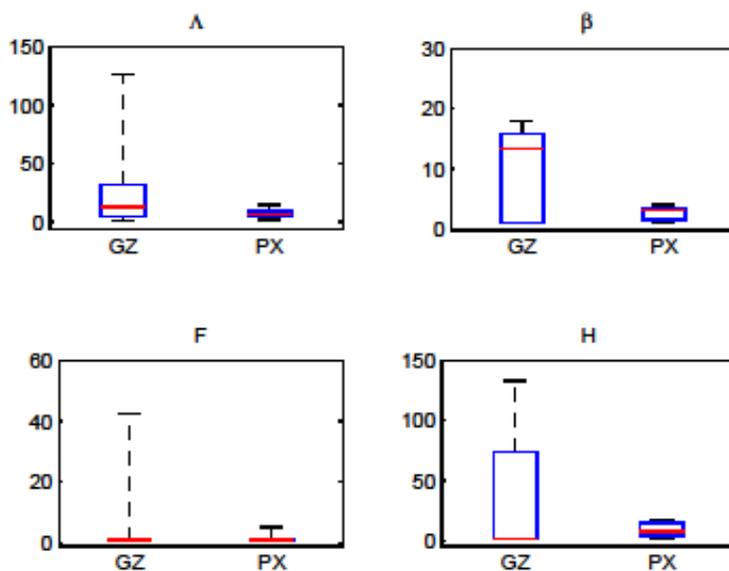| k	1	2	3	4	5	6	7
	-5545.0	-5493.0	-5425.2	-5400.1	-5388.0	-5335.3	-5360.9

Figure 1: Daily returns of the six currencies.

Figure 2: Box and wisker plots of inefficiency factors for blocks of parameters: $(\Lambda, \beta, F, \Sigma)$. The models were estimated using Gibbs sampling. *GZ* refers to the model with the upper triangular Λ_1 with positive elements on the main diagonal, and PX refers to the parameter expansion of the invariant specification.