Bayesian Infinite Factor Models with Non-Gaussian Factors

Margarita Grushanina[*] Sylvia Frühwirth-Schnatter[†]

Abstract

Bayesian factor models represent a very popular tool in the analysis of high-dimensional datasets. The cumbersome task of determining the number of factors has in recent years been addressed in literature by employing nonparametric models for the automatic inference on the number of factors. However, factors are usually assumed to be normally distributed. In reality, this assumption may prove to be too restrictive. Here, the factor model with automatic inference on the number of factors is extended to the non-Gaussian case. We relax the assumption of normality by employing a Laplace prior on factors. Two types of shrinkage priors are considered: the multiplicative gamma process prior and the cumulative shrinkage process, based on a sequence of spike-and-slab-distributions. An estimator of the covariance matrix, used to bound the prior on the idiosyncratic variances away from zero, is adapted to the non-Gaussian case. The models are tested both on simulated data sets as well as on a Eurozone countries inflation rates data set.

Key Words: Factor analysis, multiplicative gamma process, Laplace prior, non-Gaussian factors, adaptive Gibbs sampling, shrinkage, spike-and-slab prior

1. Introduction

In the recent two decades there has been considerable research done in the area of Bayesian factor analysis. However, inference on the true number of factors has for a long time remained a challenging task. The most common approach in the literature is to use various criteria to define the number of factors before running the MCMC algorithm. For example, Bai and Ng (2002) use information criteria to compare models with different factors' cardinalities, Kapetanios (2010) does model comparison using test statistics while Polasek (1997) and Lopes and West (2004) rely on marginal likelihood estimation to determine the true number of factors in the model. As a different approach, Frühwirth-Schnatter and Lopes (2018) suggest a one-sweep algorithm to estimate the true number of factors from overfitting factor models.

Another strand of literature covers models, which do not perform any preliminary inference on the number of factors but instead allow the number of factors to be potentially infinite. The dimension reduction is then achieved by assuming a nonparametric prior on factor loadings which penalises the increase in number of factors. Bhattacharya and Dunson (2011) introduced the multiplicative gamma process (MGP) prior for the increasing penalisation of the loading matrix columns, which has been widely applied, see, e.g., Murphy et al. (2020), among many others. Knowles and Ghahramani (2011) and Rockova and George (2016) use the Indian Buffet Process (IBP) to enforce sparsity on factor loadings. Recently, Legramanti et al. (2020) suggest to employ a sequence of spike-and-slab priors that introduces cumulative shrinkage on the growing number of loading matrix columns.

A common assumption in factor analysis is that the factors are normally distributed. In reality, this assumption does not always hold. There is a growing econometric literature

[*] Institute for Quantitative Economics, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria

[†] Institute for Statistics and Mathematics, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria

on models dealing with non-Gaussian variables (see, for example, Chiu et al. (2016) and Piatek and Papaspiliopoulos (2018)). However, to our knowledge, infinite factor models have so far been mostly applied to models with normal factors. In this paper, we relax the assumption of normality by assigning a Laplace prior to common factors in an infinite factor model.

For the inference on factors' cardinality, we consider two approaches. First, we follow Bhattacharya and Dunson (2011) and assume an MGP prior for our factor loading matrix. Second, we employ the cumulative shrinkage prior (CUSP), introduced in Legramanti et al. (2020) as a more efficient alternative to the MGP prior. We perform comparative analysis of the two approaches both on a simulated dataset and on Eurozone inflation data.

2. Bayesian factor model

2.1 Basic factor model with latent factors

A basic factor model with latent factors is usually written in the form

$$\boldsymbol{y}_t = \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{\epsilon}_t,$$

where each of the p variables $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{pt})'$ in a random sample $\boldsymbol{y} = (\boldsymbol{y}_t, t = 1, \ldots, T)$ of T observations are related to an k-dimensional vector of latent random variables (common factors) $\boldsymbol{f}_t = (f_{1t}, \ldots, f_{kt})'$. $\boldsymbol{\Lambda}$ is the unknown $p \times k$ factor loading matrix with factor loadings Λ_{ij}, k represents the number of factors and it is assumed that $k \ll p$.

Usually the factors are assumed to be orthogonal and normally distributed:

$$\boldsymbol{f}_t \sim N_k(0, \boldsymbol{I}_k). \tag{1}$$

Furthermore, $\boldsymbol{f}_t$ and $\boldsymbol{f}_s$ assumed pairwise independent for $t \neq s$. The idiosyncratic errors are assumed normal and also pairwise independent:

$$\epsilon_t \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}), \qquad \boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2). \tag{2}$$

From (2) it is evident that the p elements of $\boldsymbol{y}_t$ are independent conditional on $\boldsymbol{f}_t$, so all dependence in the model is generated through common factors. It follows from (1) and (2) that the observations $\boldsymbol{y}_t$ also arise from a multivariate normal distribution,

$$\boldsymbol{y}_t \sim N_p(\mathbf{0}, \boldsymbol{\Omega}), \qquad \boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Sigma}.$$

2.2 Relaxing the assumption of Gaussian factors

While the assumption that common factors follow a Gaussian distribution is valid for many applications, there is a number of cases where it can be questioned. In the recent literature on factor models, Piatek and Papaspiliopoulos (2018), for example, use a data set from psychology with a clear evidence of non-Gaussian factors. In economics, time paths of interest rates, stock market prices and returns as well as inflation rates often exhibit non-Gaussian features. Factor models have already proved useful in explaining dependencies and movements of inflation rates in various countries. Neely and Rapach (2011), among others, studied 64 countries' inflation rates in a Bayesian dynamic latent factor model and found that world and regional factors explain a large proportion of price change in individual countries.

In this paper, we analyze a data set on inflation rates in 19 Eurozone countries. The dataset contains monthly inflation rates from February 1997 to October 2019, in total 273 observations. Following standard practice, the time series for each country were de-meaned

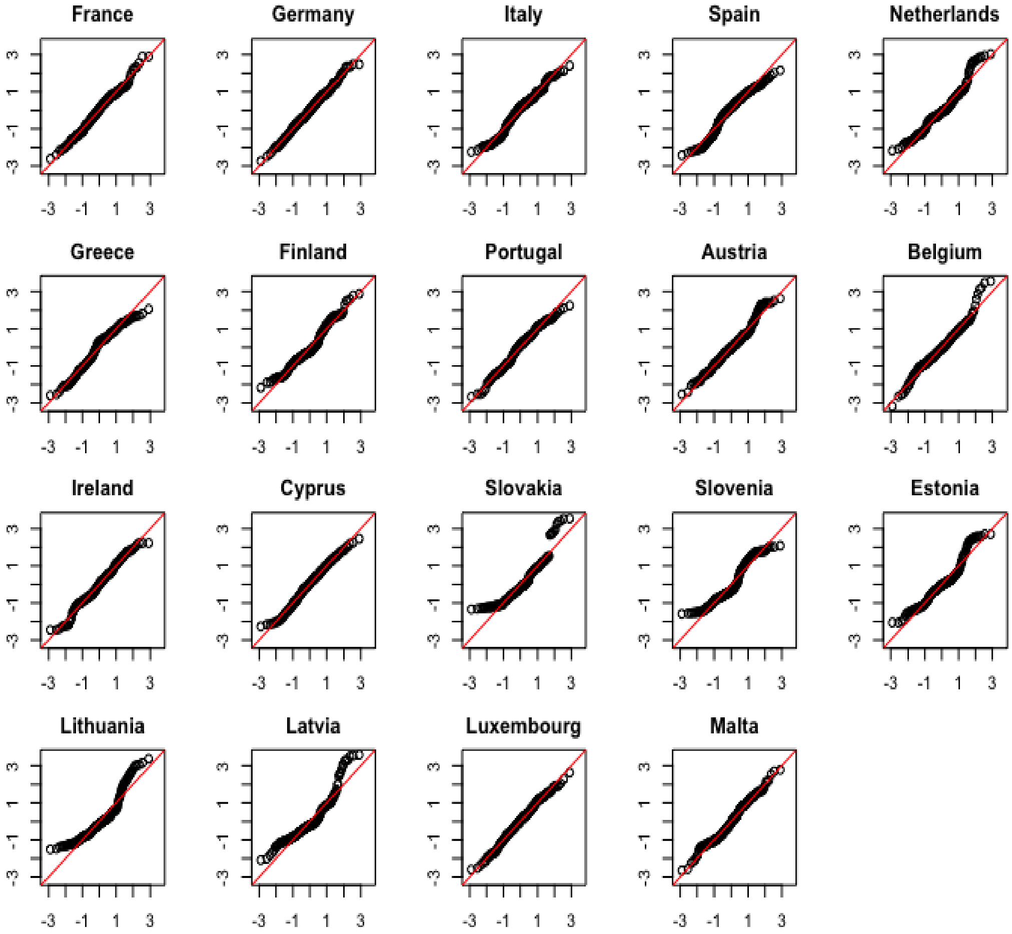

Figure 1: *Eurozone inflation data set. QQ plot comparing the observed inflation rates to a normal distribution.*

and standardised. The QQ plot in Figure 1 shows that for a majority of countries the inflation rates do not look like they are normally distributed. In addition, we performed the Shapiro-Wilk test of normality to test the validity of the assumption that the dataset is normally distributed. Table 1 shows that for at least 14 countries the p-value of the Shapiro-Wilk test is below any reasonable threshold, which allows us to reject the null hypothesis of normality.

2.3 Sparse factor models with Laplace prior

A natural alternative to a Gaussian prior on factors, which also accounts for some level of sparsity, is a Laplace distribution $\mathcal{L}$ with a zero mean and a scale hyperparameter c_j:

$$f_{jt} \sim \mathcal{L}(0, c_j), \qquad j = 1, \dots, k.$$

With the assumption that the factors are independent, the joint density of the factors factorises into individual factor densities:

$$p(\boldsymbol{f}_t) = \prod_{j=1}^{k} \frac{1}{2\,c_j} \exp -\frac{|f_{jt}|}{c_j}.$$

The variance of a Laplace distributed random variable is $2c_j^2$, so to obtain factors with unit variance we need to choose $c_j = 1/\sqrt{2}$ for all factors.

Table 1: *Eurozone inflation data set. Shapiro-Wilk test of normality.*

Country name	Shapiro-Wilk p-value	Country name	Shapiro-Wilk p-value
France	0.05450	Ireland	0.02031
Germany	0.47867	Cyprus	0.07933
Italy	0.00023	Slovakia	0.00000
Spain	0.00000	Slovenia	0.00000
Netherlands	0.00000	Estonia	0.00000
Greece	0.00001	Lithuania	0.00000
Finland	0.00000	Latvia	0.00000
Portugal	0.00515	Luxembourg	0.30377
Austria	0.00087	Malta	0.00692
Belgium	0.00024		

A Laplace distribution can be presented as an infinite mixture of normal distributions, given as a marginal distribution of the bivariate random variable (f_{jt}, w_{jt}), where

$$f_{jt}|w_{jt} \sim N(0, w_{jt}), \qquad w_{jt} \sim \mathrm{Exp}(1/2c_j^2). \tag{3}$$

This representation will be useful in Section 3.3 and 4.2, were we develop adaptive Gibbs samplers for factor models with Laplace priors. The prior variance w_{jt} is latent and can be recovered from Bayes theorem given f_{jt}:

$$p(w_{jt}|f_{jt}) \propto p(f_{jt}|w_{jt})\, p(w_{jt}) \propto w_{jt}^{-1/2} \exp\left(-\frac{f_{jt}^2}{2\, w_{jt}}\right) \exp\left(-\frac{w_{jt}}{2\, c_j^2}\right)$$

Replacing w_{jt} with $\tilde{w}_{jt} = 1/w_{jt}$, we obtain:

$$p(\tilde{w}_{jt}|f_{jt}) \propto \tilde{w}_{jt}^{-3/2} \exp\left(-\frac{f_{jt}^2 \tilde{w}_{jt}}{2}\right) \exp\left(-\frac{1}{2\, c_j^2\, \tilde{w}_{jt}}\right) \tag{4}$$

As a result, $\tilde{w}_{jt}$ given f_{jt} follows an inverse Gaussian distribution, $\tilde{w}_{jt}|f_{jt} \sim \mathrm{InvGau}(\frac{1}{|f_{jt}|\, c_j}, \frac{1}{c_j^2})$.

2.4 Identification issues in factor models

In general, the decomposition $\boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Sigma}$ is not unique. Firstly, this decomposition will also hold for any semi-orthogonal matrix $\boldsymbol{P}$ ($\boldsymbol{P}\boldsymbol{P}' = \boldsymbol{I}$) and $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}\boldsymbol{P}$, $\boldsymbol{g}_t = \boldsymbol{P}'\boldsymbol{f}_t$, in which case the two models

$$\boldsymbol{y}_t = \boldsymbol{\Lambda}\boldsymbol{f}_t + \boldsymbol{\epsilon}_t \quad \text{and} \quad \boldsymbol{y}_t = \boldsymbol{\Lambda}_1\boldsymbol{g}_t + \boldsymbol{\epsilon}_t$$

are observationally equivalent. This problem is known in factor analysis as rotational invariance and additional constraints are required to uniquely identify the model parameters. For Gaussian factors this problem is usually solved by assuming the loading matrix be positive lower triangular, see e.g. Lopes and West (2004) and Frühwirth-Schnatter and Lopes (2018), with the choice of factors' order being an important modeling decision.

One of the advantages of models with non-Gaussian factors is that the factor loading matrix is uniquely identified, usually up to signs and column permutations. In particular for non-Gaussian factors with distributions that are symmetric around zero, such trivial rotations $\boldsymbol{P}$ are the only ones, where $\boldsymbol{f}_t$ and $\boldsymbol{g}_t = \boldsymbol{P}'\boldsymbol{f}_t$ can have the same distribution. This means that rotational invariance is not an issue for non-Gaussian factor models.

Another issue is how to ensure that in the following representation

$$\boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Sigma}, \qquad \boldsymbol{\Omega} = \boldsymbol{\Theta}\boldsymbol{\Theta}' + \boldsymbol{\Sigma_0},$$

$\boldsymbol{\Sigma} = \boldsymbol{\Sigma_0}$ and, hence, the cross-covariance matrix $\boldsymbol{\Lambda}\boldsymbol{\Lambda}' = \boldsymbol{\Theta}\boldsymbol{\Theta}'$ is uniquely identified. This is the problem of variance identification and it can easily fail if the number of factors is too high. The row deletion property of Anderson and Rubin (1956) states that whenever an arbitrary row is deleted from $\boldsymbol{\Lambda}$, two disjoint submatrices of rank k remain. This imposes an upper bound on the number of factors, namely $k \leq \frac{p-1}{2}$, and raises the interesting question, whether variance identification can actually hold in an infinite factor model. The effective number of factors in such models is not, in fact, infinity, but instead either a conservatively chosen upper bound, H, as in Legramanti et al. (2020), or the number of active factors k^* estimated during MCMC sampling as in Bhattacharya and Dunson (2011). Nevertheless, there is no formal means to guarantee that the condition of Anderson and Rubin is going to hold. The value of k^* often varies from iteration to iteration and it theoretically can remain for some time at a rather high level even after the burn-in period, thus influencing the posterior computations of parameters. In sparse factor models, which is the original application of the MGP shrinkage prior, the identification problem becomes even more complicated as additional restrictions on the non-zero elements of the loading matrix are needed to ensure variance identification (see Frühwirth-Schnatter and Lopes (2018)).

However, in many real life applications the true number of factors satisfies $k \ll p$ and in many cases the condition $k \leq \frac{p-1}{2}$ will not be violated after some sufficient burn-in period, thus allowing variance identification.

In addition, identification of factor loadings is not necessary for some applications, such as forecasting, variable selection and estimation of the marginal covariance matrix. For these applications infinite factor models can be a useful tool which results in improved posterior estimations. Another possibility is to use such models for a preliminary inference on the number of factors before running a different model for estimating factor loadings.

3. Infinite Factor Models with Multiplicative Gamma Process Prior

3.1 Prior assumptions

3.1.1 MGP shrinkage prior on factor loadings

The MGP prior on the factor loadings introduced by Bhattacharya and Dunson (2011) has the following form:

$$\lambda_{ih}|\phi_{ih}, \tau_h \sim N(0, \phi_{ih}^{-1}\tau_h^{-1}), \qquad \phi_{ih} \sim \mathcal{G}(\nu_1/2, \nu_2/2), \qquad \tau_h = \prod_{l=1}^{h} \delta_l,$$

$$\delta_1 \sim \mathcal{G}(a_1, b_1), \qquad \delta_l \sim \mathcal{G}(a_2, b_2), \quad l \geq 2,$$

where δ_l $(l = 1, \dots, \infty)$ are independent, τ_h is a global shrinkage parameter for the h-th column, ϕ_{ih} are local shrinkage parameters for the elements of the h-th column.

Bhattacharya and Dunson (2011) state that if $a_2 > 1$ then τ_hs are stochastically increasing with increasing h. Durante (2017) argues that this is not sufficient to guarantee the increasing shrinkage property in a general case. Instead, $a_2 > b_2 + 1$ and $a_2 > a_1$ are necessary and sufficient conditions for the increasing penalization of a high number of factors as long as $a_1 > 0$ and $a_2 > 0$ and the values of a_1 are not excessively high. We follow Bhattacharya and Dunson (2011) and set b_1 and b_2 at 1 and let the data define the values of a_1 and a_2 in a Metropolis-within-Gibbs step imposing a hyperprior of $\mathcal{G}(2,1)$ on both shape parameters.

3.1.2 Prior on idiosyncratic variances

Bhattacharya and Dunson (2011) assume an inverse Gamma prior $\sigma_i^2 \sim \mathcal{G}^{-1}(c_0, C_0)$ on the variance of the error term with the same shape and rate hyperparameter c_0 and C_0 for all p variables y_{it} and set the hyperparameters at $c_0 = 1$ and $C_0 = 0.25$. In our simulations on synthetic data this approach worked well in cases when $p \ll T$, however, with p comparable or bigger than T the posterior distribution of some of the variances happens to be multimodal, with one mode lying at zero (a Bayesian analogue of the Heywood problem).

In view of this problem, we follow Frühwirth-Schnatter and Lopes (2018) in setting the shape and rate hyperparameters in such a way as to bound the prior away from 0. This yields the following prior:

$$\sigma_i^2 \sim \mathcal{G}^{-1}(c_0, (c_0 - 1)/\widehat{\mathbf{\Omega^{-1}}}_{i,i}),$$

where $(\widehat{\mathbf{\Omega^{-1}}})$ is an estimator of the precision matrix of the data and c_0 is a hyperparameter which is set to 2.5. As the sample precision matrix is unstable when p is not small compared to T and not available in case when $p \geq T$, we follow Frühwirth-Schnatter and Lopes (2018) and use the estimator $(\widehat{\mathbf{\Omega^{-1}}}) = (b_0+T/2)(b_0\boldsymbol{S}_0+0.5\sum_{t=1}^{T}\boldsymbol{y}_t\boldsymbol{y}_t^T)^{-1}$, where b_0 is a hyperparameter. This approach has also been applied in Murphy et al. (2020) in the context of infinite factor models.

For standardised data, $\boldsymbol{S}_0$ can be replaced by the identity matrix $\boldsymbol{I}_p$. For the case of unstandardised data, we follow Murphy et al. (2020) and Wang et al. (2015) who suggest to first use the estimator for the inverse correlation matrix and then scale it by the diagonal entries of the sample covariance matrix, thus yielding the following estimator:

$$(\widehat{\mathbf{\Omega^{-1}}}) = \text{diag}(\boldsymbol{S})^{-\frac{1}{2}}\left((b_0 + T/2)(b_0\boldsymbol{R}_0 + 0.5\sum_{t=1}^{T}\boldsymbol{y}_t\boldsymbol{y}_t^T)^{-1}\right)\text{diag}(\boldsymbol{S})^{-\frac{1}{2}},$$

where $\boldsymbol{R}_0$ is the correlation matrix of the data and $\boldsymbol{S}$ is the sample covariance matrix.

This method works well, when the distribution of the original data set is approximately normal. When the data exhibits non-Gaussian features, like in the case when the factors are simulated from a Laplace distribution or the inflation rates data, the above mentioned method cannot be applied. A common practice with datasets of a moderate sample size, exhibiting a bell shape when plotted on a graph, and with unknown variance, is to use the Student's t-distribution, which puts a lower probability to the center and a higher probability to the tails. Following this practice, we suggest to use the multivariate t-distribution to derive an estimator of the inverse of the marginal covariance matrix:

$$\boldsymbol{y}_t \sim t_\nu(\boldsymbol{\mu}, \mathbf{\Omega}),$$

where the degrees of freedom ν is equal to the number of observations minus one.

To obtain an estimator of $\mathbf{\Omega}$, we sample from the posterior distribution of $\mathbf{\Omega}$ given the data. To this aim, we use a Gibbs sampling algorithm based on the representation of the $t_\nu(\boldsymbol{\mu}, \mathbf{\Omega})$ distribution as an infinite scale mixture of multivariate normal distribution, with the scaling parameter having a Gamma distribution:

$$\boldsymbol{y}_t \sim N_p(\boldsymbol{\mu}, \mathbf{\Omega}/\eta_t), \qquad \eta_t \sim \mathcal{G}(\nu/2, \nu/2),$$

where $\boldsymbol{\mu} = \mathbf{0}$ for de-meaned data. Under the Wishart prior $\boldsymbol{\Omega}^{-1} \sim \mathcal{W}_p(b_0, \boldsymbol{B}_0)$, the precision matrix and the scaling parameters $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_T)$ can be sampled in two steps:

Step 1. Sampling the precision matrix from

$$\boldsymbol{\Omega}^{-1}|\boldsymbol{\eta}, \boldsymbol{y} \sim \mathcal{W}_p \left(b_0 + T/2, \boldsymbol{B}_0 + 0.5 \sum_{t=1}^{T} \eta_t \boldsymbol{y}_t \boldsymbol{y}_t^T \right).$$

Step 2. Sampling the scaling parameters η_t independently for each observation $\boldsymbol{y}_t$ from

$$\eta_t|\boldsymbol{\Omega}, \nu, \boldsymbol{y} \sim \mathcal{G}\left((\nu + p)/2, (\nu + d_M(\boldsymbol{y}_t; \boldsymbol{\Omega}))/2\right),$$

where $d_M(\boldsymbol{y}_t; \boldsymbol{\Sigma}) = \boldsymbol{y}_t^T \boldsymbol{\Omega}^{-1} \boldsymbol{y}_t$ is the Mahalanobis distance of $\boldsymbol{y}_t$ from the origin. Hence, observations with a small Mahalanobis distance obtain higher weights than observations with a large Mahalanobis distance.

Like for the Gaussian case, for unstandardised data an estimator based on the correlation matrix can be used:

$$\widehat{(\boldsymbol{\Omega^{-1}})} = \mathrm{diag}(\boldsymbol{S})^{-\frac{1}{2}} \left((b_0 + T/2)(b_0 \boldsymbol{R}_0 + 0.5 \sum_{t=1}^{T} \hat{\eta}_t \boldsymbol{y}_t \boldsymbol{y}_t^T)^{-1} \right) \mathrm{diag}(\boldsymbol{S})^{-\frac{1}{2}},$$

where $\hat{\eta}_t$ is the average of the MCMC draws. Using this estimator to replace the inverse covariance matrix in the prior for idiosyncratic variances significantly decreases the occurrence of multimodality in the posterior of the error term components.

3.2 Adaptive inference on number of factors

Although the number of factors theoretically are allowed to be infinitely large, in reality one should choose a suitable level of truncation k^*, which should be large enough not to miss any important factors, but not overly conservative to waste computational effort.

The sampler is initiated with a conservative guess k_0, which is assumed to be substantially larger than the supposed actual number of factors. At each iteration the posterior samples of the loading matrix contain information about the effective number of factors. Let $m^{(g)}$ be the number of columns of the loading matrix having all elements in a pre-specified small neighbourhood of zero. Then $k^{*(g)} = k^{*(g-1)} - m^{(g)}$ is defined to be the effective number of factors at iteration g. To keep balance between reducing dimensionality and exploring the whole space of possible factors, k^* is adapted with probability $p(g) = \exp(\alpha_0 + \alpha_1 g)$ with the parameters chosen so that the adaptation occurs more often at the beginning of the chain and decreases in frequency exponentially fast (the adaptations are designed to satisfy the diminishing adaptation condition of Roberts and Rosenthal (2007), which is necessary for convergence). When the adaptation occurs, the redundant factors are discarded and the corresponding columns are deleted from the loading matrix. If the number of such columns drops to zero, a factor is added, and the parameters are sampled from the corresponding prior distributions. Adaptation is made to occur after a burn-in period, in order to ensure that the true posterior distribution is being sampled from before truncating the loading matrices.

3.3 Adaptive Gibbs sampler

The adaptive Gibbs sampler of Bhattacharya and Dunson (2011) is easily adjusted to a factor model, where the common factors follow a Laplace distribution. These modifications exploit the scale mixture representation (3) of the Laplace distribution and the conditional posterior of the latent scales given in (4).

Step 1. Sample $\boldsymbol{\lambda}_i$ for i in $(1, \dots, p)$ from

$$\boldsymbol{\lambda}_i|- \sim N_{k^*}\left((\boldsymbol{\Psi}_i^{-1} + \sigma_i^{-2}\boldsymbol{F}\boldsymbol{F}^{\boldsymbol{T}})^{-1}\boldsymbol{F}\sigma_i^{-2}\boldsymbol{y}_i^T, (\boldsymbol{\Psi}_i^{-1} + \sigma_i^{-2}\boldsymbol{F}\boldsymbol{F}^T)^{-1}\right)$$

where $\boldsymbol{\Psi}_i^{-1} = \mathrm{diag}(\phi_{i1}\tau_1, \dots, \phi_{ik^*}\tau_{k^*})$.

Step 2. Sample σ_i^{-2} for i in $(1, \dots, p)$ from

$$\sigma_i^{-2}|- \sim \mathcal{G}\left(c_{0i} + \frac{T}{2}, C_{0i} + \frac{1}{2}\sum_{t=1}^{T}(y_{it} - \boldsymbol{\Lambda}_i'\boldsymbol{f}_t)^2\right).$$

Step 3. Sample $\boldsymbol{f}_t$ for t in $(1, \dots, T)$ from

$$\boldsymbol{f}_t|- \sim N_{k^*}\left((\boldsymbol{\Phi}_{k^*} + \boldsymbol{\Lambda}_{k^*}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_{k^*})^{-1}\boldsymbol{\Lambda}_{k^*}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t, (\boldsymbol{\Phi}_{k^*} + \boldsymbol{\Lambda}_{k^*}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_{k^*})^{-1}\right)$$

where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \dots, \sigma_p^2)$ and $\boldsymbol{\Phi}_{k^*} = \mathrm{diag}(\tilde{w}_{1t}, \dots, \tilde{w}_{k^*t})$ in case of a Laplace prior on factors.

Sample $\tilde{w}_{ht}$ for t in $(1, \dots, T)$ and h in $(1, \dots, k^*)$ from

$$p(\tilde{w}_{ht}|f_{ht}) \sim \mathrm{InvGau}\left(\frac{1}{|f_{ht}|\, c_h}, \frac{1}{c_h^2}\right).$$

In case of a Gaussian prior on factors as in Bhattacharya and Dunson (2011), $\boldsymbol{\Phi}_{k^*}$ is equal to the identity matrix and sampling $\tilde{w}_{ht}$ is skipped.

Step 4. Sample ϕ_{ih} for i in $(1, \dots, n)$ and h in $(1, \dots, k^*)$ from

$$\phi_{ih}|- \sim \mathcal{G}\left(\frac{\nu_1 + 1}{2}, \frac{\nu_2 + \tau_h\lambda_{ih}^2}{2}\right).$$

Step 5. Sample δ_1 from

$$\delta_1|- \sim \mathcal{G}\left(\frac{2a_1 + pk^*}{2}, 1 + \frac{1}{2}\sum_{l=1}^{k^*}\tau_l^{(1)}\sum_{i=1}^{p}\phi_{il}\lambda_{il}^2\right).$$

Sample δ_h for $h \geq 2$ from

$$\delta_h|- \sim \mathcal{G}\left(\frac{2a_2 + p(k^* - h + 1)}{2}, 1 + \frac{1}{2}\sum_{l=h}^{k^*}\tau_l^{(h)}\sum_{i=1}^{p}\phi_{il}\lambda_{il}^2\right)$$

where $\tau_l^{(h)} = \prod_{t=1, t\neq h}^{l}\delta_t$ for h in $(1, \dots, k^*)$.

Step 6. As the conditional posterior densities of $a_1|\delta_1$ and $a_2|\delta_2, \dots, \delta_{k^*}$, given by

$$a_1|\delta_1 \sim \frac{a_1}{\Gamma(a_1)}\delta_1^{a_1-1}e^{-(a_1+\delta_1)}, \quad a_2|\delta_2, \dots, \delta_{k^*} \sim \Gamma(a_2)^{-(k^*-1)}a_2\left(\prod_{l=2}^{k^*}\delta_l\right)^{a_2-1}\exp\left[-\left(a_2 + \sum_{l=2}^{k^*}\delta_l\right)\right],$$

are not straightforward to sample from, we use a simple symmetric random walk Metropolis-within-Gibbs step with $a_1^p \sim N(a_1, s_1^2)$ and $a_2^p \sim N(a_2, s_2^2)$ serving as proposal densities. The acceptance probabilities are:

$$\rho_{a_1} = \frac{\Gamma(a_1)}{\Gamma(a_1^p)}\,\frac{a_1^p}{a_1}\,\delta_1^{a_1^p - a_1}\, e^{a_1 - a_1^p},$$

$$\rho_{a_2} = \left(\frac{\Gamma(a_2)}{\Gamma(a_2^p)}\right)^{-(k^*-1)}\frac{a_2^p}{a_2}\left(\prod_{l=2}^{k^*}\delta_l\right)^{a_2^p - a_2} e^{a_2 - a_2^p}.$$

Step 7. At each iteration we generate a random number u_g from $\mathcal{U}(0, 1)$. If $u_g \leq p(g)$ we then check if some columns of the loading matrix are within the pre-specified neighbourhood of 0, and if this is so, we discard the redundant columns. In the case when the number of such columns is zero, we generate an additional column sampling parameters from the prior distributions.

4. Infinite Factor Models with Cumulative Shrinkage Process Prior

One of the drawbacks of the MGP prior model of Bhattacharya and Dunson (2011) is that the hyperparameters a_1 and a_2 control both the shrinkage rate and the prior for loadings on active factors, which creates a trade off between the need to maintain rather diffuse priors for the active terms and shrinkage for the redundant ones. This leads to a problem when the efficient shrinkage conditions as in Durante (2017) imposed on hyperparameters provide too strong shrinkage in bigger datasets (see Section 5 for more details). In addition, deletion of redundant columns depends on yet another parameter, which sets a threshold for the decision to discard the columns, and which has a substantial influence on the performance of the model. With this in mind, Legramanti et al. (2020) proposed a cumulative shrinkage process (CUSP) prior as an alternative, which largely corrects these drawbacks.

4.1 CUSP prior on factor loadings

The CUSP prior on the factor loadings induces shrinkage via a sequence of spike-and slab distributions that assign growing mass to the spike as the model complexity grows. The shrinkage prior on the factor loadings formalises as follows:

$$\lambda_{ih} \,|\, \theta_h \sim N(0, \theta_h), \quad \text{where } i = 1, \ldots, p \text{ and } h = 1, \ldots, \infty$$

$$\theta_h \,|\, \pi_h \sim (1 - \pi_h)\mathcal{G}^{-1}(a_\theta, b_\theta) + \pi_h \delta_{\theta_\infty}, \qquad \pi_h = \sum_{l=1}^{h} w_l, \qquad w_l = v_l \prod_{m=1}^{l-1}(1 - v_m) \tag{5}$$

where $\pi_h \in (0, 1)$ and the v_h are generated independently from $\mathcal{B}(1, \alpha)$, following the usual stick-breaking representation introduced in Sethuraman (1994). By integrating out θ_h, each loading λ_{ih} has the marginal prior[1]

$$\lambda_{ih} \sim (1 - \pi_h) t_{2a_\theta}(0, b_\theta/a_\theta) + \pi_h N(0, \theta_\infty)$$

where $t_{2a_\theta}(0, b_\theta/a_\theta)$ denotes the Student-*t* distribution with $2a_\theta$ degrees of freedom, location 0 and scale b_θ/a_θ. To facilitate effective shrinkage of redundant factors, θ_∞ should be set close to 0. The authors recommend a small value $\theta_\infty > 0$, following Ishwaran and Rao (2005), as it induces a continuous shrinkage prior on every factor loading, thus improving mixing and identification of inactive factors. The slab parameters a_θ and b_θ should be specified so as to induce a moderately diffuse prior on active loadings. For the implementation, the potentially infinite sequence is truncated at some conservative level H.

Prior for idiosyncratic variances remains the same as in the case of the MGP prior for factor loadings, namely $\sigma_i^2 \sim \mathcal{G}^{-1}(c_0, C_0)$. We apply the same procedure to find the

[1] In the equation (5) the inverse gamma distribution for the slab is chosen for the reasons of conjugacy. In principle, this expression provides a general prior, where a sufficiently diffuse continuous distribution needs to be chosen for the slab.

estimator of the precision matrix to bound the prior away from 0 as in Section 3.1.2. For the factors, we assume either Laplace prior, as in Section 2.3, or a Gaussian prior $\boldsymbol{f}_t \sim N_H(0, \boldsymbol{I}_H)$.

4.2 Adaptive Gibbs sampler

The adaptive Gibbs sampler of Legramanti et al. (2020) is easily adjusted to a factor model, where the common factors follow a Laplace distribution. As before, these modifications exploit the scale mixture representation (3) of the Laplace distribution and the conditional posterior of the latent scales given in (4).

Step 1. Sample $\boldsymbol{\lambda}_i$ for i in $(1, \ldots, p)$ from

$$\boldsymbol{\lambda}_i | - \sim N_H\left((\boldsymbol{\Psi}^{-1} + \sigma_i^{-2}\boldsymbol{F}\boldsymbol{F}^{\boldsymbol{T}})^{-1}\boldsymbol{F}\sigma_i^{-2}\boldsymbol{y}_i^T, (\boldsymbol{\Psi}^{-1} + \sigma_i^{-2}\boldsymbol{F}\boldsymbol{F}^T)^{-1}\right)$$

where $\boldsymbol{\Psi} = \operatorname{diag}(\theta_1, \ldots, \theta_H)$.

Step 2. Sample σ_i^{-2} for i in $(1, \ldots, p)$ from

$$\sigma_i^{-2} | - \sim \mathcal{G}\left(c_0 + \frac{T}{2}, C_0 + \frac{1}{2}\sum_{t=1}^{T}(y_{it} - \boldsymbol{\Lambda}_i' \boldsymbol{f}_t)^2\right) .$$

Step 3. Sample $\boldsymbol{f}_t$ for t in $(1, \ldots, T)$ from

$$\boldsymbol{f}_t | - \sim N_H\left((\boldsymbol{\Phi}_H + \boldsymbol{\Lambda}_H^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_H)^{-1}\boldsymbol{\Lambda}_H^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t, (\boldsymbol{\Phi}_H + \boldsymbol{\Lambda}_H^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}_H)^{-1}\right)$$

where $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ and $\boldsymbol{\Phi}_H = \operatorname{diag}(\tilde{w}_{1t}, \ldots, \tilde{w}_{Ht})$ in case of a Laplace prior on factors.

Sample $\tilde{w}_{ht}$ for t in $(1, \ldots, T)$ and h in $(1, \ldots, H)$ from

$$p(\tilde{w}_{ht} | f_{ht}) \sim \operatorname{InvGau}\left(\frac{1}{|f_{ht}|\, c_h}, \frac{1}{c_h^2}\right) .$$

In case of a Gaussian prior on the factors as in Legramanti et al. (2020), $\boldsymbol{\Phi}_H$ is equal to the identity matrix and sampling of the $\tilde{w}_{ht}$s is skipped.

Step 4. Sampling θ_h requires a data augmentation step. Thus, (5) can be obtained by marginalising out independent latent indicators z_h, with probabilities $p(z_h = l \,|\, w_l) = w_l$ for $l = 1, \ldots, H$, from the equation

$$\theta_h \,|\, z_h \sim \{1 - \mathbf{1}(z_h \leq h)\}\mathcal{G}^{-1}(a_\theta, b_\theta) + \mathbf{1}(z_h \leq h)\delta_{\theta_\infty}.$$

Sample z_h for h in $(1, \ldots, H)$ from a categorical distribution with probabilities as below

$$p(z_h = l \,|\, -) \sim \begin{cases} w_l N_p(\boldsymbol{\lambda}_h; 0, \theta_\infty \boldsymbol{I}_p), & l = 1, \ldots, h, \\ w_l t_{2a_\theta}\left(\boldsymbol{\lambda}_h; 0, (b_\theta/a_\theta)\boldsymbol{I}_p\right), & l = h+1, \ldots, H. \end{cases}$$

Step 5. Sample v_l for l in $(1, \ldots, H-1)$ from

$$v_l \,|\, - \sim \mathcal{B}\left(1 + \sum_{h=1}^{H}\mathbf{1}(z_h = l), \alpha + \sum_{h=1}^{H}\mathbf{1}(z_h > l)\right) .$$

Set $v_H = 1$ and update $w_1, \ldots, w_H$ from $w_l = v_l \prod_{m=1}^{l-1}(1 - v_m)$.

Step 6. For h in $(1, \ldots, H)$:

if $z_h \leq h$ set $\theta_h = \theta_\infty$, otherwise sample θ_h from $\mathcal{G}^{-1}\left(a_\theta + \frac{1}{2}p, b_\theta + \frac{1}{2}\sum_{j=1}^{p}\lambda_{ih}^2\right)$.

Step 7. Adaptation of the number of factors H. With the factor model truncated at H and the Hth factor modelled by a spike at θ_∞ by construction, this leaves at most $H-1$ active factors. As there cannot be more than p factors in the model, this imposes a conservative upper limit of $p+1$ upon H.

After some burn-in period $\tilde{g}$ required for the stabilization of the chain (usually set around 10% of the number of iterations), the truncation index $H^{(g)}$ and the number of active factors $H^{*(g)} = \sum_{h=1}^{H^{(g)}} \mathbf{1}(z_h^{(g)} > h)$ are adapted with probability $p(g) = exp(\alpha_0 + \alpha_1 g)$[2] as follows:

- if $H^{*(g)} < H^{(g-1)} - 1$:

  set $H^{(g)} = H^{*(g)} + 1$, drop inactive columns in $\mathbf{\Lambda}^{(g)}$ along with the associated parameters in $\boldsymbol{F}^{(g)}$, $\boldsymbol{\theta}^{(g)}$ and $\boldsymbol{w}^{(g)}$, and add the final component sampled from the spike to $\mathbf{\Lambda}^{(g)}$, together with the associated parameters in $\boldsymbol{F}^{(g)}$, $\boldsymbol{\theta}^{(g)}$ and $\boldsymbol{w}^{(g)}$ sampled from the corresponding priors

- otherwise:

  set $H^{(g)} = H^{(g-1)} + 1$ and add the final column sampled from the spike to $\mathbf{\Lambda}^{(g)}$, together with the associated parameters in $\boldsymbol{F}^{(g)}$, $\boldsymbol{\theta}^{(g)}$ and $\boldsymbol{w}^{(g)}$ sampled from the corresponding priors.

5. Simulation results

5.1 Model with the MGP prior on factor loadings and its parameter dependence

5.1.1 Gaussian prior on factors

At first, simulations were made for a model with Gaussian factors to provide some benchmark. Following Bhattacharya and Dunson (2011), a synthetic data set was simulated with $T = 100$ and idiosyncratic variances sampled from $\mathcal{G}^{-1}(1, 0.25)$. The number of non-zero elements in each column of $\mathbf{\Lambda}$ were chosen between $2k$ and $k+1$, with zeros allocated randomly and non-zero elements sampled independently from $N(0, 9)$. We generated $\boldsymbol{y}_t$ from $N_p(0, \mathbf{\Omega})$, where $\mathbf{\Omega} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Sigma}$.

We chose six (p, k) combinations, namely (6, 2), (10, 3), (30, 5), (50, 8), (100, 15) and (150, 25) with a conservative initial upper bound of $k_0 = \min(p, 5\log(p))$, and $k_0 = 10\log(p)$ for the case when $p > T$. For each pair we considered between 5 and 10 simulation replicates. The simulation was run for 25 000 iterations with a burn-in of 10 000.[3] Thinning, such as collecting every 5th sample, provided only very slight improvement, if at all, and has not significantly changed the results. After some tuning, we settled at the following tuning parameters: ν_1 and ν_2 both equal 3, $s_1 = 1$ for the scale the of the Metropolis-Hastings step for a_1 and the scale s_2 between 2 and 4 for a_2. For the case $p < T$, α_0 and α_1 in the adaptation probability expression were set as -0.5 and $-3 \times (10)^{-4}$, and the threshold for monitoring the columns to discard as 0.15 with the proportion of elements required to be below the threshold at 80 % of p. For the case $p \geq T$, α_0 and α_1 in the adaptation probability expression were set as -1 and $-5 \times (10)^{-4}$, and the threshold for monitoring the columns to discard as 0.0001.[4]

[2] The coefficients α_0 and α_1 are chosen according to the criteria described in Section 3.2

[3] For the case $p > T$ we ran the Gibbs sampler for 30 000 iterations due to slower convergence.

[4] Setting the threshold for monitoring the redundant columns at 0.0001 in case $p < T$ leads to a significant

Table 2: *Simulated data with $T = 100$. Performance of the adaptive Gibbs sampler based on the MGP prior for various combinations of p and k (Gaussian prior on factors).*

(p, k)	mean k^*	mode k^*	$\hat{a}_1$	$\hat{a}_2$
$(6, 2)$	3.78	3.8	1.29	4.18
$(10, 3)$	4.23	4.7	1.35	3.91
$(30, 5)$	5.37	5.7	2.66	2.43
$(50, 8)$	8.60	9.1	2.31	2.13
$(100, 15)$	15.43	16.0	2.21	2.04
$(150, 25)$	12.01	12.2	4.17	5.24

The simulation results in Table 2 show that the model tends to somewhat overestimate the number of factors, however, remains rather close to truth for the first 5 combinations, especially those where $k \ll p$. In the case with $p > T$ the model severely underestimated the number of the common factors.

The last two columns in Table 2 show the posterior mean of a_1 and a_2. The first efficient shrinkage condition of Durante (2017), $a_2 > b_2 + 1$, holds for all (p, k) combinations considered. For the first two combinations of p and k and the last one with $p > T$, the column shrinkage parameters a_1 and a_2, estimated from the data, are in accordance with the second efficient shrinkage condition of Durante (2017) $a_2 > a_1$. However, with higher p and $p < T$, the condition $a_2 > a_1$ seems to cease holding with $p \geq 30$. This result is of some interest especially in view of the simulation study in Durante (2017), which suggests that as the dimension of the dataset increases, the induced MGP shrinkage prior satisfying $a_2 > a_1$ might provide too strong shrinkage.

5.1.2 Laplace prior on factors

We used the same (p, k) combinations as in Section 5.1.1 to check the performance of the model with the Laplace prior on the factors. The same dataset was generated with the difference that the factors were generated from $\mathcal{L}(0, 1/\sqrt{2})$. The same parameters and the number of iterations were used as in the previous section.

Table 3: *Simulated data with $T = 100$. Performance of the adaptive Gibbs sampler based on the MGP prior for various combinations of p and k (Laplace prior on factors).*

(p, k)	mean k^*	mode k^*	$\hat{a}_1$	$\hat{a}_2$
$(6, 2)$	3.32	3.3	1.65	3.99
$(10, 3)$	3.97	3.9	1.91	3.68
$(30, 5)$	5.17	5.0	6.98	6.73
$(50, 8)$	7.40	7.5	7.86	7.03
$(100, 15)$	13.35	13.9	5.63	5.33
$(150, 25)$	18.84	20.0	3.04	3.51

The simulation results in Table 3 exhibit slightly stronger shrinkage, especially as the dimension of the data set grows. This can be explained by the additional shrinkage coming from the Laplace prior on the factors. The threshold for discarding additional columns in the factor loading matrix had to be set slightly higher to ensure a similar degree of

overestimation of the number of factors, However, tuning the threshold parameters can be tricky working with the real data sets when the true number of factors is not known.

shrinkage as in case of Gaussian factors. In the case with $p > T$, the model with Laplace factors performed better than the Gaussian factor model, which severely underestimated the number of factors.

The data driven values of the column shrinkage parameters a_1 and a_2 in Table 3 demonstrate similar pattern as in the case with the Gaussian prior on factors, namely a_1 becomes higher than a_2 when p increases.

5.2 CUSP prior on factor loadings

We once again simulated data sets the same way as for testing the model with the MGP prior on factor loadings (see Section 5.1.1, using the same (p, k) combinations. The stick breaking parameter α, which represents a prior expectation of the number of active factors in the dataset, was set to 5 (although varying it did not result in any significant difference in the performance of the model). We choose the same parameters of the slab distribution as in Legramanti et al. (2020), namely $\alpha_\theta = \beta_\theta = 2$ and $\theta_\infty = 0.05$ in case of the Gaussian prior on the factors and $\theta_\infty = 0.005$ in case of the Laplace prior on factors. The parameters of the adaptation probability of the sampler α_0 and α_1 were set as -0.5 and $-3 \times (10)^{-4}$ for the cases when p was relatively small compared to T, and as -1 and $-5 \times (10)^{-4}$ otherwise. The simulations were run for 15,000 iterations, with 5,000 discarded as burn-in.

Both models, with Gaussian and Laplace prior on factors, discover the correct number of latent factors (see Table 4 and 5) . The model with Laplace prior requires a somewhat smaller spike parameter due to the additional shrinkage imposed by choosing the Laplace distribution. In addition, similar to the simulations with the MGP prior model, the parameters of the adaptation probabilities had to be adjusted for the case with p relatively small compared to T so that to allow for a slightly higher adaptation probability. However, the number of parameters for tuning is less, and the sampler is more robust than in the case with the MGP prior on factor loadings, which seems to be working best for the models with $k \ll p$ and with $p \leq T$.

Table 4: *Simulated data with $T = 100$. Performance of the adaptive Gibbs sampler based on the CUSP prior for various combinations of p and k (Gaussian prior on factors).*

(p, k)	mean H^*	mode H^*
$(6, 2)$	2.17	2.0
$(10, 3)$	3.0	3.0
$(30, 5)$	5.0	5.0
$(50, 8)$	8.0	8.0
$(100, 15)$	15.00	15.0

Table 5: *Simulated data with $T = 100$. Performance of the adaptive Gibbs sampler based on the CUSP prior for various combinations of p and k (Laplace prior on factors).*

(p, k)	mean H^*	mode H^*
$(6, 2)$	2.23	2.0
$(10, 3)$	3.0	3.0
$(30, 5)$	5.0	5.0
$(50, 8)$	8.0	8.0
$(100, 15)$	15.0	15.0

6. Real data applications

6.1 Application to the Italian olive oil data set

We also tested the models with the CUSP prior on the factor loadings on the benchmark Italian olive oils dataset (see, e.g. Murphy et al. (2020)), available in the R package *FlexDir*. The data describe the composition of 8 fatty acids in 572 Italian olive oils, which originate from three areas: southern and northern Italy and Sardinia. Each area breaks down to several regions: southern Italy comprises north Apulia, Calabria, south Apulia, and Sicily; Sardinia is divided into inland and coastal Sardinia; and northern Italy comprises Umbria and east and west Liguria. Hence, the true number of factors should correspond to either 3 areas or 9 regions.

The results of the two models, with the Gaussian and with the Laplace prior on factors are presented in Table 6. The sampler was run for 10,000 iteration with a burn-in of 5,000.

Table 6: *Italian olive oil data set. Performance of the CUSP model with Gaussian and Laplace priors on the factors.*

Prior on factors	mean H^*	mode H^*	sd H^*
$Gaussian$	3.14	3	0.35
$Laplace$	3	3	0

Both models recovered 3 latent factors, thus satisfying the variance identification condition $k \leq \frac{p-1}{2}$. The three factors correspond to the three areas described above, namely, southern Italy, northern Italy and Sardinia. The trivial rotation problem is addressed by ordering the columns of the loading matrices in the following way. Columns are sorted so that the first column has the least number of near zero elements, the second the next smallest number of near zero elements, and so on. If two columns have the same number of near zero elements, then the one with the higher sum of its elements comes in front in order. To avoid sign switching, the signs of the elements $\lambda_{1,1}$, $\lambda_{2,2}$ and $\lambda_{3,3}$ were fixed, since their posterior distributions seem to be bounded away from 0.

The recovered identified factor loading matrices are presented in Table 7. The models have rather similar performance, while the model with the Laplace factors has the advantage, that it does not require dealing with the rotational invariance in the identification of the factor loading matrix.

Table 7: *Italian olive oil data set. Factor loadings matrix estimated by the CUSP model with, respectively, a Gaussian (left-hand side) and a Laplace (right-hand side) prior on the factors.*

	factor 1	factor 2	factor 3
1	0.74	0.05	0.08
2	0.75	0.39	0.02
3	-0.21	-0.16	0.09
4	-0.74	-0.28	0.10
5	0.51	0.31	-0.17
6	0.36	-0.33	0.40
7	0.36	-0.24	0.33
8	0.50	-0.10	0.19

	factor 1	factor 2	factor 3
1	0.35	-0.56	0.03
2	0.20	-0.63	0.14
3	-0.20	0.07	-0.50
4	-0.31	0.50	-0.06
5	0.15	-0.34	0.21
6	0.69	-0.09	-0.37
7	0.69	0.02	-0.16
8	0.41	-0.47	-0.42

6.2 Application to Eurozone inflation data

6.2.1 MGP prior on factor loadings

To check and compare the Gaussian and Laplace factor models performance on the inflation data, we chose those time series for countries, which densities look unimodal. Thus, 9 countries were chosen, namely France, Austria, Germany, Luxembourg, Italy, Cyprus, Ireland, Belgium and Malta.

Although in this case $T \gg p$ with $T = 273$ and $p = 9$, taking into account that this is an adaptive MCMC algorithm, we ran both models for 20,000 iterations to ensure that the sampler runs long enough for the adaptation frequency to reach 0 according to the diminishing adaptation condition (Roberts and Rosenthal (2007)). This was enough to achieve convergence and 6,000 initial iterations were discarded as burn-in. The variance parameters in the Metropolis-Hastings step for sampling a_1 and a_2 were 1 and 1.8 respectively. The parameters of the adaptation probability of the sampler were chosen at $\alpha_0 = -1$ and $\alpha_1 = -5 \times (10)^{-4}$.

Table 8: *Eurozone inflation data set. Performance of the MGP model with a Gaussian and Laplace prior on factors.*

Prior on factors	mean k^*	mode k^*	sd k^*	$\hat{a}_1$	$\hat{a}_2$
$Gaussian$	4.20	4	0.63	2.19	5.94
$Laplace$	2.50	2	0.90	2.33	4.17

The results for the estimated number of factors are presented in Table 8. The estimated covariance and factor loadings matrices from each of the models are presented in Table 9 to 12. In both models, the number of factors parameter stabilised after approximately 15,000 iterations. As expected, the model with the Laplace prior on factors produced a stronger shrinkage having stabilised at 2 factors, while the model with the Gaussian prior on factors stabilised at 4 factors. However, the near 0 factor loadings in the last 2 columns of the loading matrix in the Gaussian model, reported in Table 9, indicate that 2 is effectively the true number of factors discovered by the model.

Table 9: *Eurozone inflation data set. Factor loadings matrix estimated by the MGP model with Gaussian factors.*

	factor 1	factor 2	factor 3	factor 4
FR	0.91	0.03	0.00	0.00
AT	0.82	0.34	-0.00	-0.00
DE	0.84	0.29	0.00	0.00
LU	0.88	0.06	-0.02	0.00
IT	0.84	-0.28	0.01	-0.00
CY	0.74	-0.40	0.00	-0.00
IE	0.55	-0.38	-0.02	-0.00
BE	0.83	0.19	0.03	0.00
MA	0.50	-0.36	0.09	0.00

In both cases, the number of factors 4 and 2 satisfy the identifiability condition $k \leq \frac{p-1}{2}$. The rotation problem is addressed in the way described in Section 6.1. To avoid sign switching, the signs of the elements $\lambda_{1,1}$ and $\lambda_{2,2}$ were fixed, since their posterior distributions seem to be bounded away from 0.

Table 10: *Eurozone inflation data set. Covariance matrix estimated by the MGP model with Gaussian factors.*

	FR	AT	DE	LU	IT	CY	IE	BE	MA
FR	1.07	0.77	0.79	0.83	0.79	0.69	0.52	0.78	0.46
AT	0.77	1.05	0.82	0.76	0.59	0.47	0.32	0.76	0.26
DE	0.79	0.82	1.06	0.78	0.63	0.51	0.36	0.77	0.31
LU	0.83	0.76	0.78	1.06	0.75	0.66	0.50	0.75	0.39
IT	0.79	0.59	0.63	0.75	1.06	0.80	0.62	0.66	0.58
CY	0.69	0.47	0.51	0.66	0.80	1.04	0.62	0.55	0.57
IE	0.52	0.32	0.36	0.50	0.62	0.62	1.01	0.39	0.43
BE	0.78	0.76	0.77	0.75	0.66	0.55	0.39	1.04	0.41
MA	0.46	0.26	0.31	0.39	0.58	0.57	0.43	0.41	1.01

Table 11: *Eurozone inflation data set. Factor loadings matrix estimated by the MGP model with Laplace factors.*

	factor 1	factor 2
FR	0.94	0.11
AT	0.88	-0.16
DE	0.90	-0.12
LU	0.91	0.09
IT	0.84	0.37
CY	0.73	0.48
IE	0.53	0.47
BE	0.88	-0.02
MA	0.49	0.42

Table 12: *Eurozone inflation data set. Covariance matrix estimated by the MGP model with Laplace factors.*

	FR	AT	DE	LU	IT	CY	IE	BE	MA
FR	0.65	0.13	0.13	0.14	0.13	0.12	0.09	0.13	0.08
AT	0.13	0.68	0.13	0.12	0.11	0.09	0.06	0.12	0.06
DE	0.13	0.13	0.67	0.13	0.11	0.10	0.07	0.13	0.06
LU	0.14	0.12	0.13	0.69	0.13	0.11	0.08	0.13	0.08
IT	0.13	0.11	0.11	0.13	0.74	0.12	0.10	0.12	0.09
CY	0.12	0.09	0.10	0.11	0.12	0.63	0.10	0.10	0.09
IE	0.09	0.06	0.07	0.08	0.10	0.10	0.72	0.07	0.07
BE	0.13	0.12	0.13	0.13	0.12	0.10	0.07	0.67	0.07
MA	0.08	0.06	0.06	0.08	0.09	0.09	0.07	0.07	0.77

6.2.2 CUSP prior on factor loadings

As an alternative to the MGP shrinkage prior, we have also applied the CUSP prior. As real data can be relatively noisy, we decided to use the inverse gamma parameters $a_\theta = 1$ and $b_\theta = 0.1$ to achieve a sufficiently diffuse prior in the slab part. In case of the Laplace prior on the factors, the parameter θ_∞ was chosen to be smaller than in the Gaussian case, at 0.005. The stick-breaking parameter is chosen at $\alpha = 5$ as in Legramanti et al. (2020).

The parameters of the adaptation probability of the sampler we chosen at $\alpha_0 = -1$ and $\alpha_1 = -5 \times (10)^{-4}$, same as for the sampler with the MGP prior.

We ran the sampler for 15,000 iterations with 5,000 of them discarded as burn-in, which was sufficient for the adaptation frequency to reach 0, according to the diminishing adaptation condition. The results are presented in Table 13. The corresponding covariance and factor loading matrices are presented in Table 14 to 17. Identification of factor loadings was performed as described in Section 6.2.1. Columns were sorted so that the first column has the least number of near zero elements, and the signs of the elements $\lambda_{1,1}$ and $\lambda_{2,2}$ were fixed.

Both models discovered 2 factors, which seems to be the ”truth” for this data set, with the model with Laplace prior on factors having stabilised at 2 factors after the burn-in period. The CUSP model thus performed clearly better that the MGP model on the inflation rates data set. The values of factor loadings are, in fact, very similar for all 4 considered models. The first factor could be interpreted as related to a common monetary policy, as it has rather high loadings for all countries. The second factor has significant loadings for Malta, Ireland, Cyprus, and Italy, and loadings with the opposite sign for Austria and Germany. A possible interpretation could be structural productivity issues in these countries, which have a negative impact upon inflation rates.

Table 13: *Eurozone inflation data set. Performance of the CUSP model with Gaussian and Laplace priors on factors applied to the Eurozone inflation dataset.*

Prior on factors	mean H^*	mode H^*	sd H^*
$Gaussian$	2.23	2	0.42
$Laplace$	2	2	0

Table 14: *Eurozone inflation data set. Factor loadings matrix estimated by the CUSP model with Gaussian factors.*

	factor 1	factor 2
FR	0.92	0.03
AT	0.83	0.34
DE	0.84	0.27
LU	0.90	0.05
IT	0.86	-0.27
CY	0.76	-0.38
IE	0.57	-0.37
BE	0.84	0.17
MA	0.52	-0.34

Both factor loading matrices as well as covariance matrices seem to be very similar for the MGP and CUSP models, however, somewhat differ in absolute values for the Gaussian and Laplace prior on factors. The relative values remain very similar, i.e. the covariances which are higher in the case of the Gaussian prior, are also higher in the case of the Laplace prior. Correspondingly, the significant and insignificant factor loadings are the same in the models with both types of prior on factors, although their absolute values may slightly differ.

Table 15: *Eurozone inflation data set. Covariance matrix estimated by the CUSP model with Gaussian factors.*

	FR	AT	DE	LU	IT	CY	IE	BE	MA
FR	1.11	0.77	0.78	0.82	0.78	0.69	0.52	0.77	0.46
AT	0.77	1.08	0.80	0.76	0.60	0.47	0.32	0.75	0.27
DE	0.78	0.80	1.08	0.77	0.63	0.52	0.37	0.77	0.32
LU	0.82	0.76	0.77	1.08	0.75	0.66	0.50	0.75	0.40
IT	0.78	0.60	0.63	0.75	1.09	0.78	0.62	0.66	0.57
CY	0.69	0.47	0.52	0.66	0.78	1.06	0.61	0.56	0.56
IE	0.52	0.32	0.37	0.50	0.62	0.61	1.03	0.39	0.42
BE	0.77	0.75	0.77	0.75	0.66	0.56	0.39	1.06	0.40
MA	0.46	0.27	0.32	0.40	0.57	0.56	0.42	0.40	1.02

Table 16: *Eurozone inflation data set. Factor loadings matrix estimated by the CUSP model with Laplace factors.*

	factor 1	factor 2
FR	0.93	-0.05
AT	0.88	0.16
DE	0.90	0.12
LU	0.90	-0.03
IT	0.82	-0.24
CY	0.71	-0.32
IE	0.52	-0.33
BE	0.88	0.05
MA	0.48	-0.29

Table 17: *Eurozone inflation data set. Covariance matrix estimated by the CUSP model with Laplace factors.*

	FR	AT	DE	LU	IT	CY	IE	BE	MA
FR	0.66	0.13	0.13	0.13	0.13	0.12	0.09	0.13	0.08
AT	0.13	0.69	0.13	0.12	0.11	0.09	0.07	0.12	0.06
DE	0.13	0.13	0.68	0.13	0.11	0.10	0.07	0.13	0.06
LU	0.13	0.12	0.13	0.70	0.13	0.11	0.09	0.13	0.08
IT	0.13	0.11	0.11	0.13	0.76	0.12	0.10	0.12	0.09
CY	0.12	0.09	0.10	0.11	0.12	0.63	0.10	0.10	0.09
IE	0.09	0.07	0.07	0.09	0.10	0.10	0.73	0.08	0.07
BE	0.13	0.12	0.13	0.13	0.12	0.10	0.08	0.68	0.07
MA	0.08	0.06	0.06	0.08	0.09	0.09	0.07	0.07	0.78

7. Conclusion

We have extended two existing models with nonparametric priors on the factor loadings and automatic inference on the number of factors to the case, which allows non-Gaussian factors. More specifically, we have adjusted the proposed adaptive sampling algorithm with the Laplace prior on latent factors and suggested the respective estimator of the precision matrix for the prior on idiosyncratic variances. Apart from the obvious case when the data exhibits non-Gaussian features, allowing factors to have non-Gaussian distribution has the advantage of solving the problem of rotational invariance of the factor loading matrix.

References

Anderson, T.W. and H. Rubin (1956). “Statistical inference in factor analysis”. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* V, pp. 111–150.

Bai, J. and S. Ng (2002). “Determining the number of factors in approximate factor models”. In: *Econometrica*(70), pp. 191–221.

Bhattacharya, A. and D.B. Dunson (2011). “Sparse Bayesian infinite factor models”. In: *Biometrika* 98(2), pp. 291–306.

Chiu, C.J., H. Mumtaz, and G. Pinter (2016). *VAR Models with Non-Gaussian Shocks*. Discussion Papers 1609. Centre for Macroeconomics (CFM).

Durante, D. (2017). “A note on the multiplicative gamma process”. In: *Statistics & Probability Letters*(122), pp. 198–204.

Frühwirth-Schnatter, S. and H. F. Lopes (2018). “Sparse Bayesian Factor Analysis when the number of factors is unknown”. In: *arXiv preprint arXiv:1804.04231*.

Ishwaran, H. and J.S. Rao (2005). “Spike and slab variable selection: Frequentist and Bayesian strategies”. In: *The Annals of Statistics*(33(2)), pp. 730–773.

Kapetanios, G. (2010). “A testing procedure for determining the number of factors in approximate factor models with large datasets”. In: *Journal of Business and Economic Statistics* 3(28), pp. 251–258.

Knowles, D. and Z. Ghahramani (2011). “Nonparametric Bayesian sparse factor models with application to gene expression modeling”. In: *The Annals of Applied Statistics*(5(2B)), pp. 1534–1552.

Legramanti, S., D. Durante, and D.B. Dunson (2020). “Bayesian cumulative shrinkage for infinite factorizations”. In: *Biometrika*(107(3)), pp. 745–752.

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

Murphy, K., C. Viroli, and I.C. Gormley (2020). “Infinite Mixtures of Infinite Factor Analysers”. In: *Bayesian analysis* 15(3), pp. 937–963.

Neely, C.J. and D.E. Rapach (2011). “International comovements in inflation rates and country characteristics”. In: *Journal of International Money and Finance* 30(7), pp. 1471–1490.

Piatek, R. and O. Papaspiliopoulos (2018). *A Bayesian Nonparametric Approach to Factor Analysis*. Working paper.

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

Roberts, G.O. and J.S. Rosenthal (2007). “Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms”. In: *Journal of Applied Probability*(44), pp. 458–475.

Rockova, V. and E.I. George (2016). “Fast Bayesian factor analysis via automatic rotation to sparsity”. In: *Journal of the American Statistical Association*(111(516)), pp. 1608–1622.

Sethuraman, J. (1994). “A constructive definition of Dirichlet priors”. In: *Statistica Sinica* 4, pp. 639–650.

Wang, C. et al. (2015). “Shrinkage estimation of large dimensional precision matrix using random matrix theory”. In: *Statistica Sinica*(25(3)), pp. 993–1008.