Bayesian Analysis (2025) **TBA**, Number TBA, pp. 1–30

# Dynamic Mixture of Finite Mixtures of Factor Analysers

Margarita Grushanina* and Sylvia Frühwirth-Schnatter†

**Abstract.** Mixtures of factor analysers represent a popular tool for finding structure in data. While in many applications the number of clusters and latent factors within clusters is held constant, some recent models automatically infer cluster and/or factor dimensionalities. This is done by employing nonparametric priors and allowing the number of clusters and factors to potentially be infinite. Markov chain Monte Carlo (MCMC) estimation is performed via adaptive algorithms, where parameters associated with the redundant factors are discarded. The current work contributes to the literature by allowing automatic inference on the number of clusters and cluster-specific factors while keeping both dimensions finite. For automatic inference on the cluster structure we employ the dynamic mixture of finite mixtures model with a prior on the number of mixture components. Automatic inference on cluster-specific factors is performed by assigning an exchangeable shrinkage process prior, which can be interpreted as a generalized cumulative shrinkage process prior for the columns of the factor loading matrices. Extensive simulation studies and applications to benchmark as well as real data sets demonstrate that our model outperforms competing alternatives, in particular those based on the multiplicative gamma process prior, with respect to recovering the correct number of cluster-specific factors.

**Keywords:** spike-and-slab prior, exchangeable shrinkage process prior, Indian buffet process, Pitman-Yor process prior, finite mixture models, telescoping sampling.

**MSC2020 subject classifications:** Primary 62H25, 62H30; secondary 62F15, 62G05.

## 1 Introduction

Mixtures of factor analysers (MFA) models combine clustering with local dimensionality reduction performed separately in each cluster and are particularly useful for modelling data with complex and nonhomogeneous structure. First works that describe MFA models appeared already in the 1990s when Ghahramani and Hinton (1996) developed an expectation-maximization (EM) algorithm for inference on the parameters of such a model. This algorithm was further developed for the purpose of clustering high-dimensional data in McLachlan et al. (2003); see also McLachlan and Peel (2000, Chapter 9). A Bayesian treatment of MFA via a variational approximation was first considered by Ghahramani and Beal (2000). At the same time, Fokoue (2000) discussed

*Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom, m.grushanina@imperial.ac.uk

†Institute for Statistics and Mathematics, Vienna University of Economics and Business, Vienna, Austria, sylvia.fruehwirth-schnatter@wu.ac.at

© 2025 International Society for Bayesian Analysis https://doi.org/10.1214/25-BA1543

an algorithm for Markov chain Monte Carlo (MCMC) inference of an MFA model, which was further ameliorated in Fokoue and Titterington (2003). The following years have seen a fair amount of literature on various versions of MFAs. The most notable ones include McNicholas and Murphy (2008), who assessed an MFA model in the context of parsimonious Gaussian mixture models, and Viroli (2010), who introduced a mixture of factor mixture analysers. The key feature of this model is that it assumes that the data are generated according to several factor models with a certain prior probability (thus performing a local dimension reduction at the first level), and that in each factor model the factors are described by a multivariate mixture of Gaussians (thus performing a global dimension reduction at the second level).

Determining the number of clusters and the number of cluster-specific factors has always been a challenging issue. Many authors treat both quantities as known while others, like McNicholas and Murphy (2008) and Viroli (2010), run their models for various numbers of components in the mixture and factors in the factor-analytical part and use model selection criteria to choose the best fitting model. In an early attempt to find a way to learn model dimensions from data, Fokoue and Titterington (2003) developed a Birth-and-Death MCMC algorithm, which uses the fact that the posterior distributions of both factor models and finite mixture models are invariant to permutations of the order of their parameters, and thus the collection of their parameters can be viewed as a point process. More recently, Papastamoulis (2018) introduced an overfitting Bayesian MFA model, which estimates the unknown number of mixture components assuming a fixed number of factors. The optimal number of factors is then determined using information criteria.

The most flexible Bayesian MFA model to date is the infinite mixture of infinite factor analysers (IMIFA) model introduced in Murphy et al. (2020) that allows automatic inference on both the number of clusters and cluster-specific factors by assigning nonparametric priors to both cluster weights and cluster-specific factor loadings. For automatic inference on the number of clusters, the IMIFA model employs a Pitman-Yor process (PYP) prior, using its stick-breaking representation and a slice sampler for MCMC estimation. Automatic inference on the number of cluster-specific factors is achieved with the multiplicative gamma process (MGP) prior of Bhattacharya and Dunson (2011) and an adaptive Gibbs sampler is used to facilitate estimation with varying dimensions. In fact, Murphy et al. (2020) offer an entire family of Bayesian MFA models which encompass various approaches to model selection regarding both the number of factors and the number of clusters. These models are briefly reviewed in Section 2.

The automatic inference on the number of clusters and factors achieved with Bayesian nonparametric priors has the invaluable advantage of solving the model selection problem, but staying within an infinite-dimensional framework comes at certain costs, both regarding methodological and computational aspects. In many applied sciences, the focus lies on data sets in which the number of mixture components $K$ is expected to be finite, even if unknown. Recent insights (Alamichel et al., 2024; Miller, 2023) indicate that finite mixtures with a prior on $K$ provide an attractive alternative to infinite mixtures in this case. Computational issues are, for instance, subjective truncation criteria on which models based on the MGP prior are highly dependent; see Schiavon and Canale (2020) for a thorough investigation of this issue.

In the present paper, we complement the IMIFA family and suggest an innovative Bayesian MFA model with automatic inference on the number of clusters and cluster-specific factors, all of which are assumed to be finite but unknown. This is achieved by replacing the infinite Bayesian nonparametric priors (i.e. PYP and MGP priors) employed by Murphy et al. (2020) by finite counterparts. This also leads to computational advantages.

Regarding the mixture part, we employ a mixture of finite mixtures (MFM) model that assumes that the number of mixture components $K$ is finite but unknown, and puts a prior on $K$ (Miller and Harrison, 2018). This model class allows inference with respect to the number of filled components, which can be regarded as the number of clusters in the data. More specifically, we employ the dynamic MFM model introduced in Frühwirth-Schnatter et al. (2021) where the hyperparameter of the multinomial Dirichlet prior on the component weights is inversely proportional to $K$, thus facilitating more shrinkage with increasing number of components, which leads to the model being "dynamic" with respect to $K$. This framework has the computational advantage that the telescoping sampler of Frühwirth-Schnatter et al. (2021) can be employed, allowing inference to be performed almost entirely within a Gibbs sampler, with only one hyperparameter requiring a Metropolis-Hastings step, despite $K$ being unknown.

Regarding the factor models in each cluster, we employ a cluster-specific exchangeable shrinkage process (ESP) prior. Under this prior, the dimensionality of the cluster-specific factor model is fixed while the number of active factors is unknown and estimated from the data. As shown by Frühwirth-Schnatter (2023), this prior generalizes the cumulative shrinkage process (CUSP) prior introduced by Legramanti et al. (2020). Such (generalized) CUSP priors are two-component mixture priors where the spike shrinks factor loadings toward zero and the spike probability increases as the column index in the factor loading matrix increases. This allows inference on the number of active columns in the loading matrix, which can be regarded as the number of cluster-specific factors. The amount of shrinkage of the prior is controlled by a hyperparameter which is not fixed as in Legramanti et al. (2020), but learned from the data. However, plugging a CUSP prior instead of a MGP prior into a Bayesian factor model – suggested as a future research direction in Murphy et al. (2020) – is not trivial from a computational viewpoint. The ESP prior has the computational advantage of solely involving Gibbs steps, even for the unknown hyperparameter in control of shrinkage. It is based on an intrinsically binary classification problem with a binary latent indicator, improving the sampler of Legramanti et al. (2020) that involves a potentially high-dimensional latent categorical variable for classification.

The rest of the paper is organized as follows. Section 2 introduces the general notion of a Bayesian MFA model, reviews the IMIFA model family, and discusses how to extend this model class through alternative prior choices. Section 3 outlines the specific prior choices of our Bayesian MFA model, with details on, respectively, the dynamic MFM model given in Section 3.1 and the ESP prior given in Section 3.2. Section 4 presents our four-block MCMC algorithm based on the telescoping sampler. The performance of our method is illustrated in an extensive simulation study in Section 5 and through a benchmark data application in Section 6. The paper concludes in Section 7. Techni-

cal details, additional simulation studies and real data examples are presented in the Supplementary material (Grushanina and Frühwirth-Schnatter, 2025).

## 2 Bayesian MFA

In this section, we provide a brief review of Bayesian MFA models and discuss how this model class can be extended. Given $N$ observations $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ of a multivariate $p$-dimensional random variable, a Bayesian MFA model is formulated as follows. It is assumed that these observations can be grouped into $K$ groups (clusters) and within each of these $K$ clusters, labelled by $k = 1, \ldots, K$, the variability of the observations can be described by $H_k$ latent factors the number of which can vary across clusters. The model can be formalised in the following way. The variation of the observations within each cluster $k$ is described by a cluster-specific factor-analytical model:

$$\mathbf{y}_n - \boldsymbol{\mu}_k = \boldsymbol{\Lambda}_k \mathbf{f}_n^k + \boldsymbol{\epsilon}_n,$$

where $\boldsymbol{\mu}_k$ is a $p$-dimensional vector of cluster-specific means and $\boldsymbol{\Lambda}_k$ is a $p \times H_k$-dimensional cluster-specific factor loading matrix with elements $\lambda_{ihk}$ in row $i = 1, \ldots, p$ and column $h = 1, \ldots, H_k$. $\mathbf{f}_n^k$ is a $H_k$-dimensional vector of latent factors, and $\boldsymbol{\epsilon}_n$ is a $p$-dimensional vector of idiosyncratic errors. It is usually assumed that the latent factors are orthogonal, namely:

$$\mathbf{f}_n^k \sim \mathcal{N}_{H_k}(\mathbf{0}, \mathbf{I}_{H_k}), \tag{2.1}$$

and that $\mathbf{f}_n^k$, $\mathbf{f}_s^k$, $\boldsymbol{\epsilon}_n$ and $\boldsymbol{\epsilon}_s$ are pairwise independent for all $n \neq s$. The idiosyncratic errors are assumed to follow a normal distribution with cluster-specific diagonal covariance matrices:

$$\boldsymbol{\epsilon}_n \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Xi}_k), \qquad \boldsymbol{\Xi}_k = \operatorname{diag}(\xi_{1k}^2, \ldots, \xi_{pk}^2). \tag{2.2}$$

This suggests that conditional on $\mathbf{f}_n^k$ all $p$ elements of $\mathbf{y}_n$ are independent, so all dependencies between the variables are explained by the common factors $\mathbf{f}_n^k$. The assumptions (2.1) and (2.2) imply that the data within each cluster arise from a multivariate normal distribution. Hence, we can formulate the following mixture model independently for each observation $\mathbf{y}_n, n = 1, \ldots, N$:

$$f(\mathbf{y}_n \,|\, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_K, \boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_K) = \sum_{k=1}^{K} \eta_k \mathcal{N}_p\left(\mathbf{y}_n; \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k\right), \tag{2.3}$$

where $\boldsymbol{\Omega}_k = \boldsymbol{\Lambda}_k \boldsymbol{\Lambda}_k^T + \boldsymbol{\Xi}_k$ denotes the cluster-specific covariance matrix of the data and $\eta_k (k = 1, \ldots, K)$ are cluster weights with $\sum_{k=1}^{K} \eta_k = 1$. Note that this decomposition of $\boldsymbol{\Omega}_k$ into the sum of the cross-covariance matrix $\boldsymbol{\Lambda}_k \boldsymbol{\Lambda}_k^T$ and the idiosyncratic error covariance matrix $\boldsymbol{\Xi}_k$ is only possible under assumptions (2.1) and (2.2).

The main challenge lies in establishing the values of $K$ and $H_1, \ldots, H_K$. Most of the frequentist MFA literature requires the model dimension to be pre-specified, and usually information criteria are used for model choice. Recently, Murphy et al. (2020) have demonstrated that it is possible to recover both the number of clusters and cluster-specific factors from the data in a one sweep algorithm within a Bayesian framework. To this aim, two classes of prior families, both well-known in Bayesian non-parametric

inference, are introduced within the MFA framework, namely a Pitman-Yor process prior to learn the number of clusters and a multiplicative gamma process prior to learn the number of cluster-specific factors. Below, these prior choices are briefly reviewed and alternative choices are discussed.

## 2.1 Learning the number of clusters

To identify the number of clusters, Murphy et al. (2020) assign the PYP prior to the mixture weights with the stick-breaking representation:

$$\eta_k = v_k \prod_{l=1}^{k-1}(1-v_l), \quad v_k \sim \mathcal{B}(1-\sigma, d+k\sigma), \quad k = 1, 2, \ldots \tag{2.4}$$

where the reinforcement parameter $\sigma$ and the concentration parameter $d$ are required to satisfy the conditions $\sigma \in [0, 1)$ and $d > -\sigma$. The PYP a priori assumes $K = \infty$ components in (2.3) and defines an infinite MFA model. For a given sample size $N$, only finitely many components are occupied and the number of clusters, denoted by $K_+$, is defined as the number of non-empty components. Choosing $\sigma = 0$ yields an infinite MFA based on the Dirichlet process (DP) prior with a concentration parameter equal to $d$. Murphy et al. (2020) impose a joint (spike-and-slab) prior on both $d$ and $\sigma$ to learn these hyperparameters from the data. This is important for learning the number of clusters, as recently demonstrated by Ascolani et al. (2023) for DP priors that lack posterior consistency for the number of clusters if $d$ is fixed (Miller and Harrison, 2013).

It is well-known that both DP and PYP mixtures with $\sigma > 0$ imply that the number of clusters $K_+$ is increasing and diverges as $N$ goes to infinity, with a rate depending on $\sigma$. In the present paper, we are interested in applications in which the data arise from a moderate number of clusters, even if the sample size increases. As alternatives to infinite MFAs, the overfitting MFA (OMFA) model suggested by Papastamoulis (2018) can be applied. In this specification of model (2.3), a large but fixed $K$ is combined with the symmetric Dirichlet prior $\eta_1, \ldots, \eta_K \sim Dir_K(\alpha_\eta/K)$ on the mixture weights. For appropriate choices of the hyperparameter $\alpha_\eta$, this yields a sparse finite mixture where $K_+ < K$ with high probability (Frühwirth-Schnatter and Malsiner-Walli, 2019).

Different prior choices on the mixture weights can be characterized and compared for given hyperparameters by the induced exchangeable partition probability function (EPPF), defined as the prior $p(\mathcal{C}|N)$ over all random partitions $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{K_+}\}$ of the $N$ observations into $K_+$ clusters, with each cluster $\mathcal{C}_j$ containing $N_j > 0$ observations. An OMFA specification can be regarded as an "approximation" to an infinite MFA based on the DP prior, since the EPPF $p(\mathcal{C}|N, K, \alpha_\eta)$ of such a model converges to the EPPF of a DP mixture with concentration parameter $\alpha_\eta$, given by the Ewens distribution $p_{\mathrm{DP}}(\mathcal{C}|N, \alpha_\eta)$, as $K$ increases. On the other hand, the EPPF of an OMFA model with a given value of $K$ can be matched to the EPPF of a second family of PYP priors, where the reinforcement parameter $\sigma < 0$ and the concentration parameter is equal to $d = K|\sigma|$. In the corresponding stick-breaking representation, stick $v_K = 1$ and this PYP prior yields a mixture with infinitely many components, of which only $K$ have non-zero weights, with the symmetric Dirichlet distribution $\mathcal{D}_K(|\sigma|) = \mathcal{D}_K(d/K)$ acting as a prior on the weight distribution (Gnedin, 2010; De Blasi et al., 2015). Based on

Gnedin and Pitman (2006), we find that the OMFA model of Papastamoulis (2018) with concentration parameter $\alpha_\eta > 0$ and $K$ fixed and a Bayesian MFA model with the PYP prior on the mixtures weights with concentration parameter $d = \alpha_\eta$ and reinforcement parameter $\sigma = -\alpha_\eta/K$ imply the same EPPF and are equivalent regarding the implied prior on the partitions and the number of clusters.

In Section 3.1, we generalize the OMFA specification by putting a prior on the number of components $K$, while keeping the symmetric Dirichlet prior $Dir_K(\alpha_\eta/K)$ as a conditional prior on the mixture weights, given $K$. This yields a so-called dynamic mixture of finite MFA, abbreviated as MFMFA, which is a special case of the class of generalized MFM introduced by Frühwirth-Schnatter et al. (2021). The latter paper derives many interesting properties of this model class, among them the EPPF $p(\mathcal{C}|N, \alpha_\eta)$ for any given prior $p(K)$. Based on these results, the EPPF of a dynamic MFMFA reads:

$$p(\mathcal{C}|N, \alpha_\eta) = p_{\text{DP}}(\mathcal{C}|N, \alpha_\eta) \times \sum_{K=K_+}^{\infty} p(K) R(K, \alpha_\eta, K_+), \tag{2.5}$$

$$R(K, \alpha_\eta, K_+) = \prod_{j=1}^{K_+} \frac{\Gamma(N_j + \frac{\alpha_\eta}{K})(K - j + 1)}{\Gamma(1 + \frac{\alpha_\eta}{K})\Gamma(N_j)K},$$

where $p_{\text{DP}}(\mathcal{C}|N, \alpha_\eta)$ is the probability mass function (pmf) of the Ewens distribution. The paper by Greve et al. (2022) derives the induced prior on the number of clusters $p(K_+|N, \alpha_\eta)$ and other functionals from this EPPF which are helpful for a deeper understanding of the clustering behaviour of this specific Bayesian MFA model.

## 2.2 Learning the number of cluster-specific factors

While the OMFA model of Papastamoulis (2018) is able to learn the number of clusters, it pursues the traditional approach regarding the unknown factor dimensions $H_1, \ldots, H_K$. They assume that $H_1 = \ldots = H_K = H$ is fixed and work with the prior $\lambda_{ihk}|\phi_h \sim \mathcal{N}(0, \theta_h)$, $\theta_h \sim \mathcal{G}^{-1}(g_0, h_0)$, $h = 1, \ldots, H$, with a column specific scale parameter $\theta_h$ that is identical for all clusters. Traditional model selection criteria such as Akaike, Bayesian or deviance information criteria are then used for selecting $H$.

The work of Murphy et al. (2020) provides a major improvement over this approach. They introduce a well-known shrinkage prior on the factor loadings to the framework of a Bayesian MFA, namely the MGP prior (Bhattacharya and Dunson, 2011) which allows the factor loading matrix $\mathbf{\Lambda}_k$ in each cluster to have potentially infinitely many columns. More specifically, with $h = 1, \ldots, \infty$ denoting the column index, the prior on the factor loadings $\lambda_{ihk}$ in the $i$th row is formulated for $i = 1, \ldots, p$ as follows:

$$\lambda_{ihk}|\phi_{ihk}, \tau_{hk}, \sigma_k \sim \mathcal{N}(0, \phi_{ihk}^{-1}\tau_{hk}^{-1}\sigma_k^{-1}), \quad \phi_{ihk} \sim \mathcal{G}(\nu_1/2, \nu_2/2), \quad \tau_{hk} = \prod_{l=1}^{h} \delta_{lk},$$
$$\delta_{1k} \sim \mathcal{G}(a_1, b_1), \quad \delta_{lk} \sim \mathcal{G}(a_2, b_2), \quad l \geq 2, \quad \sigma_k \sim \mathcal{G}(\rho_1, \rho_2), \tag{2.6}$$

where $\delta_{lk}$ $(l = 1, \ldots, \infty)$ are independent and $\tau_{hk}$ is a shrinkage parameter for the $h$-th column of the $k$-th cluster-specific loading matrix $\mathbf{\Lambda}_k$. The $\tau_{hk}$s are stochastically

increasing under the restriction $a_2 > 1$, which favours growing shrinkage as the column index $h$ increases. At the same time, $\phi_{ihk}$ are local shrinkage parameters for the factor loadings $\lambda_{ihk}$ in $\mathbf{\Lambda}_k$ and are designed to favour sparsity while also preserving non-zero loadings. Finally, $\sigma_k$ is a global cluster-specific shrinkage parameter reflecting the belief that the overall degree of shrinkage varies across the clusters.

This MGP prior defines a mixture of infinite factor analysers (MIFA) model and, in combination with the PYP prior on the mixture weights, yields the IMIFA model, while in combination with the overfitting mixture it produces the OMIFA model (Murphy et al., 2020). While this model class represents an important benchmark in nonparametric factor models, its implementation comes with certain challenges. First of all, to learn the cluster-specific factor dimensions $H_k$ in the various components $k = 1, \ldots, \infty$, a clear-cut decision which factors are active and which ones are inactive is needed. However, as for any continuous shrinkage prior, only soft thresholding is feasible for a MGP prior. Following Bhattacharya and Dunson (2011), factor $h$ in cluster $k$ is discarded during the MCMC sampling if a certain proportion of elements of the $h$-th column of the $k$-th cluster loading matrix $\mathbf{\Lambda}_k$ lies within an $\epsilon_0$ neighbourhood of zero. While this allows to learn possibly different cluster-specific factor dimensions $H_k$, the choice of the truncation criteria can be rather influential, as shown in Schiavon and Canale (2020). In practice, this makes it rather hard to set the truncation level, in particular, when working with data sets where there is no clear indication of the probable number of latent factors. Second, the hyperparameters $a_1$ and $a_2$ in (2.6) control at the same time both the shrinkage rate and the prior for loadings on active factors, which creates a trade-off between the need to maintain a rather diffuse prior for the active terms and to introduce shrinkage for the redundant ones. As illustrated in Durante (2017), this leads to the problem that the efficient shrinkage conditions imposed on the hyperparameters provide too strong shrinkage in larger data sets.

For this reason, we consider alternative shrinkage priors for the factor loading matrices in a Bayesian MFA model in the present paper. One such family is the CUSP prior of Legramanti et al. (2020) which defines an alternative factor model with a priori infinitely many columns in $\mathbf{\Lambda}_k$ and largely corrects the drawbacks of the MGP prior. This prior has not yet been implemented for Bayesian MFA models, but can easily be adjusted to define an alternative MIFA model (as also mentioned as a possible future research direction in Murphy et al. (2020)). The CUSP prior is a nonparametric prior on the variances of the elements of the factor loading matrix and represents a sequence of spike-and-slab distributions which assign growing mass to the spike as the model complexity grows. Active loadings are controlled by the slab parameters, while inactive loadings are controlled by the spike parameters. For any component $k$ of a MIFA model, regardless whether the number of components $K$ is finite or infinite, the prior on the factor loadings $\lambda_{ihk}$ in the $i$th row of the $h$-th column of $\mathbf{\Lambda}_k$ is formulated for $i = 1, \ldots, p$ as follows:

$$\begin{aligned} \lambda_{ihk} \,|\, \theta_{hk} &\sim \mathcal{N}(0, \theta_{hk}), \\ \theta_{hk} \,|\, \pi_{hk} &\sim (1 - \pi_{hk}) p_{slab}(\theta_{hk}|\phi_\theta) + \pi_{hk}\delta_{\theta_\infty}, \qquad (2.7) \\ \pi_{hk} &= \sum_{l=1}^{h} w_{lk}, \quad w_{lk} = v_{lk} \prod_{m=1}^{l-1} (1 - v_{mk}), \qquad (2.8) \end{aligned}$$

where $\pi_{hk} \in (0,1)$ is the probability of the spike. Note that the definition of $\pi_{hk}$ in (2.8) as a cumulative sum over $w_{1k}, \ldots, w_{hk}$ implies that the sequence $\pi_{1k}, \pi_{2k}, \pi_{3k}, \ldots$ is increasing, pulling in this way all factor loadings a priori toward zero as the column index $h$ increases, similar in spirit to the MGP prior. Moreover, $w_{lk}$ exhibits the usual stick-breaking representation of a DP prior (Sethuraman, 1994) where the $v_{hk}$ are generated independently from a $\mathcal{B}(1, \alpha_\pi)$-distribution. Finally, $\theta_{hk}$ is a column shrinkage parameter for the $h$th column of the cluster-specific factor loading matrix $\mathbf{\Lambda}_k$, $\phi_\theta$ denotes the hyperparameters of the slab distribution and $\theta_\infty$ is fixed in Legramanti et al. (2020) at 0.05.

The CUSP prior and its properties were further studied and generalised in Kowal and Canale (2023) and Frühwirth-Schnatter (2023) with applications, respectively, in the context of nonparametric functional bases as well as sparse Bayesian factor analysis. Their insights are exploited in the present paper in the context of Bayesian MFA. First of all, the construction of the cluster-specific spike probabilities $\pi_{1k}, \pi_{2k}, \pi_{3k}, \ldots$ from the stick-breaking representation of a DP prior with concentration parameter $\alpha_\pi$ can be generalized by involving an arbitrary sequence of sticks $v_{1k}, \ldots, v_{Hk}$, where $H$ can be both finite or infinite. Second, the Dirac spike in (2.7) can be replaced by a continuous distribution $p_{spike}(\theta_{hk})$ without affecting the key properties of the prior, provided that the slab distribution is stochastically dominated by the spike distribution around 0 (Frühwirth-Schnatter, 2023, Proposition 2.2). The resulting spike-and-slab prior yields a generalized CUSP prior which can, in principle, be based on the stick-breaking representation of any finite or infinite mixture. The adaptation of the factor dimensionality $H_k$ in each cluster $k$ of an MFA model based on a generalized CUSP prior avoids the ambiguity of the MGP prior, since the "inactive" columns of $\mathbf{\Lambda}_k$ are identified as those which are assigned to the spike and $H_k$ is defined as the number of "active" columns assigned to the slab.

An interesting class of generalized CUSP priors arises from independent sticks $v_{hk} \sim B(a_h, b_h), h = 1, \ldots, H$, where the decreasing slab probabilities $\pi^\star_{hk} = 1 - \pi_{hk}$ have a representation as a multiplicative beta process (Frühwirth-Schnatter, 2023):

$$\pi^\star_{hk} = \prod_{\ell=1}^{h} (1 - v_{\ell k}) = \prod_{\ell=1}^{h} v^\star_{\ell k}, \quad v^\star_{hk} \sim B(b_h, a_h), \quad h = 1, \ldots, H. \tag{2.9}$$

Various generalized CUSP priors can be applied for a Bayesian MFA by specific choices of $H$ and $(a_h, b_h), h = 1, \ldots, H$ in (2.9). Choosing $H = \infty$ in combination with a beta prior on the sticks, where $a_h$ and $b_h$ are constant yields new MIFA models. For instance, $a_h = 1$ and $b_h = \alpha_\tau$ as in Legramanti et al. (2020) or $a_h = \beta$ and $b_h = \beta\alpha_\tau$ as in Kowal and Canale (2023) yields MIFA models based on, respectively, a one-parameter (Teh et al., 2007) or a two-parameter (Ghahramani et al., 2007) Indian buffet process (IBP) prior. The MCMC procedure introduced by Legramanti et al. (2020) for the CUSP prior could, in principle, be extended to MIFA, however, such a sampling strategy quickly becomes computationally intensive. To perform MCMC for the CUSP prior, Legramanti et al. (2020) truncate the infinite representation (2.7) at a finite upper limit $H < \infty$ for the number of factors (which can be chosen adaptively during MCMC) by defining $v_{Hk} = 1$. Despite being an intrinsically binary classification problem, the assignment of the columns of the factor loading matrices to spike or slab requires in each cluster data augmentation based on $H$ categorical variables $z_{hk}$, each with $H$ realisations in

$\{1, 2, \ldots, H\}$, based on the discrete prior $\Pr(z_{hk} = \ell \,|\, w_{\ell k}) = w_{\ell k})$, $\ell = 1, \ldots, H$. The number of active columns in the cluster-specific loading matrix $\mathbf{\Lambda}_k$ is then defined as $H_k = \sum_{h=1}^{H} I\{z_{hk} > h\}$. As a result, sampling the categorical indicators $z_{1k}, \ldots, z_{Hk}$ to perform classifications of all $H$ columns of $\mathbf{\Lambda}_k$ into active or inactive ones, operates on a separate $H \times H$ grid for each cluster and requires $H^2$ density evaluations.

Frühwirth-Schnatter (2023) discusses an alternative (generalized) CUSP prior with finite $H$ which can be used to define the (new) class of mixture of (finite) overfitting factor analysers (MOFA). In the context of Bayesian MFA, this prior involves the following sticks:

$$v_{hk} \sim B(a_h, b_h), \quad h = 1, \ldots, H, \tag{2.10}$$

where $a_h = 1$ and $b_h = \alpha_\tau(H - h + 1)/H$. For a wide range of hyperparameters $\alpha_\tau$ (which can be learned from the data), this defines a MOFA specification where $H_k < H$ with high probability. As will be discussed in Section 3.2, prior (2.10) emerges as the CUSP representation of an ESP prior on the slab probabilities, which is invariant to permutations of the column index within each cluster. Working with the ESP prior, which is commonly applied in finite sparse Bayesian factor analysis (see e.g. Frühwirth-Schnatter et al. (2025)), is advantageous from a computational points of view, in particular in the context of Bayesian MFA. Due to the exchangeability of the weights (rather than the sticks), classification of the columns into spike and slab involves data augmentation based on $H$ binary (rather than categorical) variables $\gamma_{1k}, \ldots, \gamma_{Hk}$ and requires only $2H$ density evaluations in each cluster. Finally, it is evident that prior (2.10) converges to the CUSP prior of Legramanti et al. (2020) with the concentration parameter $\alpha_\tau$ approaching the hyperparameter $\alpha_\pi$ as $H$ increases. Hence, a MOFA model based on the finite generalized CUSP prior (2.10) offers an attractive alternative to the truncated version of the CUSP prior of Legramanti et al. (2020), in particularly from a computational viewpoint.

# 3 Extending the IMIFA model family

The clustering properties of any prior on the mixture weights hold, in general, independently of the chosen family of component distributions and, therefore, independently of the prior distribution on the component-specific parameters. Hence, for any Bayesian MFA, the prior on the factor loadings within each cluster can be chosen independently from the prior on the mixture weights and a range of new Bayesian MFA models can be constructed to expand the IMIFA model family introduced by Murphy et al. (2020) by combining the priors discussed, respectively, in Section 2.1 and Section 2.2. As mentioned earlier, any MIFA model can be based on the CUSP prior of Legramanti et al. (2020) instead of the MGP prior to learn the cluster specific factor dimensions $H_k$. In combination with the PYP prior for the number of clusters, this yields an alternative to the IMIFA model of Murphy et al. (2020) and avoids the drawbacks of the MGP prior. Efficient computation for this model class can be based on an infinite mixture of overfitted factor analysers (IMOFA) model, where the PYP prior is combined with prior (2.10) with a finite, fixed upper limit $H$ for learning the factor dimension $H_k$ in

each cluster. As discussed, such a model can be regarded as a finite approximation of an IMIFA model based on the CUSP prior.

In the present paper, we focus on Bayesian MFA models with automatic inference on the dimensionality of both cluster and factor structure of the data, which, at the same time, allows to keep both $K$ and $H_1, \dots, H_K$ finite, while being random variables that are inferred from the data. To this aim, we employ the dynamic mixture of finite mixtures model to identify the cluster structure of the data. For the cluster-specific factor-analytical part of the model, we rely on the finite generalized CUSP prior introduced in (2.10) and exploit its relationship to the class of ESP priors (Frühwirth-Schnatter, 2023). We call the resulting novel MFA model a dynamic mixture of finite mixtures of factor analysers, or dynamic MFMFA for short. In addition to the reasons outlined in Section 2.1, this choice is also guided by computational considerations. Learning the number of clusters in a dynamic MFMFA model gives the additional advantage of making the MCMC estimation possible almost entirely within Gibbs sampler steps relying on telescoping sampling and eliminating the need to refer to additional methods such as slice sampling.

## 3.1 Dynamic mixture of finite mixtures of factor analysers

Let $k = 1, \dots, K$ denote the cluster index and $h = 1, \dots, H$ denote the factor indices within a cluster with $H$ being the maximum possible number of factors in any cluster. The dynamic MFM model is a mixture model with a prior $p(K)$ on the number of mixture components $K$ and can be written in the following hierarchical way (Frühwirth-Schnatter et al., 2021):

$$\begin{aligned}
K &\sim p(K), \\
\eta_1, \dots, \eta_K \,|\, K, \alpha_\eta &\sim Dir_K\left(\frac{\alpha_\eta}{K}\right), \\
\boldsymbol{\mu}_k \,|\, \mathbf{b}_0, \mathbf{B}_0 &\sim N_p(\mathbf{b}_0, \mathbf{B}_0), && \text{cluster means for } k = 1, \dots, K, \\
S_n \,|\eta_1, \dots, \eta_K &\sim \mathcal{M}(1; \eta_1, \dots, \eta_K), && \text{latent allocation variables for } n = 1, \dots, N, \\
\mathbf{y}_n \,|S_n = k, \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k &\sim \mathcal{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Omega}_k), && \text{for each data point } n = 1, \dots, N, \\
\boldsymbol{\Omega}_k &= \boldsymbol{\Lambda}_k \boldsymbol{\Lambda}_k^T + \boldsymbol{\Xi}_k, && \text{cluster covariance matrices for } k = 1, \dots, K.
\end{aligned} \tag{3.1}$$

The model is called "dynamic" due to the fact that the Dirichlet concentration parameter $\alpha_\eta / K$ is inversely proportional to the number of components $K$, which favours more sparse solutions as the number of components grows. Under this model, the joint distribution of the data $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$ has a representation as a countably infinite MFM with $K$ $p$-variate Gaussian components:

$$p(\mathbf{y}) = \sum_{K=1}^{\infty} p(K) \prod_{n=1}^{N} \sum_{k=1}^{K} \eta_k \mathcal{N}_p(\mathbf{y}; \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k).$$

In this framework, $K$ is defined as the (theoretical) number of components in the mixture, while the number of clusters $K_+$ is defined as the number of filled components that generated the data, namely $K_+ = \sum_{k=1}^{K} \mathcal{I}\{N_k > 0\}$, where $N_k = \#\{n : S_n = k\}$ counts the observations generated by the component $k$. Assigning a prior to $K$ has the

advantage that both $K$ and $K_+$ are random a priori. Depending both on the number of observations $N$ and the choice of hyperparameters, they can be close or rather different, see the detailed investigation in Greve et al. (2022). Having the Dirichlet concentration parameter decrease with increasing $K$ allows a gap between $K_+$ and $K$ and thus ensures randomness in the prior distribution of $K_+$ for a broad range of hyperparameters $\alpha_\eta$.

The EPPF of a dynamic MFMFA has been provided in (2.5). In the limiting case where the prior $p(K)$ increasingly concentrates all prior mass at $K = +\infty$, this model converges to an infinite MFA where clustering is based on a DP mixture. An OMFA model arises as that special case where the prior $p(K)$ puts all prior mass on a fixed value $K = K_f$. For more general priors $p(K)$, the connection of the dynamic MFMFA to infinite MFA based on the PYP prior is of special interest. The EPPF (2.5) arises from mixing a PYP prior over the reinforcement parameter $\sigma_K = -\alpha_\eta/K$, while the concentration parameter $d = \alpha_\eta$ is independent of $K$. Hence, the dynamic MFMFA model represents a model beyond the class of Gibbs-type priors.

It remains to be discussed how to choose $p(K)$. Following the considerations in Frühwirth-Schnatter et al. (2021) and Grün et al. (2022), we chose the suggested translated beta-negative-binomial (BNB) prior $K - 1 \sim BNB(\alpha_\lambda, a_\pi, b_\pi)$, which represents a hierarchical generalisation of the Poisson, the geometric and the negative-binomial distributions. The p.m.f. of the prior on $K$ takes the following form for $k = 1, 2, \ldots$:

$$\Pr(K = k) = \frac{\Gamma(\alpha_\lambda + k - 1) B(\alpha_\lambda + a_\pi, k - 1 + b_\pi)}{\Gamma(\alpha_\lambda)\Gamma(k) B(a_\pi, b_\pi)} \tag{3.2}$$

where $\alpha_\lambda > 0$, $a_\pi > 0$ and $b_\pi > 0$ are hyperparameters. The choice of hyperparameters can be governed by the desired value of the prior mean $E(K) = 1 + \alpha_\lambda b_\pi/(a_\pi - 1)$, which exists as long as $\alpha_\pi > 1$. An important advantage of this prior is that the three parameters $\alpha_\lambda$, $a_\pi$ and $b_\pi$ allow simultaneous control over both the expectation and the tails of the prior on $K$ as well as the implied prior on $K_+$ and its expectation (see Frühwirth-Schnatter et al. (2021) and Greve et al. (2022) for details on the induced prior on $K_+$).

Since the hyperparameter $\alpha_\eta$ in the Dirichlet concentration parameter $\alpha_\eta/K$ plays an important role for the prior distribution induced on the number of filled clusters $K_+$ and the partitions, we learn it from data by assigning it a prior and updating it from the posterior distribution in a random walk Metropolis-Hastings step. We choose the F-distribution prior $\alpha_\eta \sim \mathcal{F}(\nu_l, \nu_r)$ as it is flexible enough to allow various cluster solutions by modeling the behaviour close to zero and in the tail independently (see Frühwirth-Schnatter et al. (2021) for further motivation of this prior choice).

## 3.2 ESP prior on factor loading variances

Frühwirth-Schnatter (2023) introduces the general class of exchangeable shrinkage process priors, which take the form of unordered spike-and-slab priors. Adjusted for the MFA framework, this prior is defined as follows. For each cluster $k = 1, \ldots, K$, assume $\boldsymbol{\tau}_k = \{\tau_{hk} \in (0, 1) : h = 1, \ldots, H\}$ to be a finite sequence of i.i.d. random probabilities. Let $\Theta_k = \{\theta_{hk}\}$, $h = 1, \ldots, H$ be a finite sequence of column-specific shrinkage parameters assumed to be independent conditional on $\boldsymbol{\tau}_k$ and independent of all $\tau_{lk}$, $l \neq h$ for

all $h$. If the prior $\theta_{hk}|\tau_{hk}$ takes the following spike-and-slab form:

$$\theta_{hk}|\tau_{hk} \sim (1-\tau_{hk})p_{spike}(\theta_{hk}|\phi_0) + \tau_{hk}p_{slab}(\theta_{hk}|\phi_\theta), \tag{3.3}$$

where $\phi_0$ and $\phi_\theta$ are the hyperparameters of the spike and the slab distributions, respectively, then $\Theta_k$ follows an ESP prior. The prior is exchangeable in the sense that it is invariant to permutations of the column indices $h$ within each cluster and to permutations of the cluster indices $k$ across clusters.

It is often assumed in the literature that the slab probabilities $\tau_{1k}, \ldots, \tau_{Hk}$ follow a beta distribution, where the first parameter depends on $H$. Here, $H$ is the same in all clusters and can be considered as the maximum possible number of factors allowed by the data. A typical choice of the beta prior for $\tau_{hk}$ would be

$$\tau_{hk}|H, \alpha_\tau \sim \mathcal{B}\left(b_0\frac{\alpha_\tau}{H}, b_0\right), \quad h = 1, \ldots, H.$$

This prior was proposed in Frühwirth-Schnatter et al. (2025) in the context of sparse finite Bayesian factor models. For $H \to \infty$ it converges to the infinite two-parameter beta prior introduced by Ghahramani et al. (2007) in the framework of Bayesian nonparametric latent feature models. With $b_0 = 1$, this prior becomes the one-parameter beta prior employed by Ročková and George (2016), which converges to the IBP prior for $H \to \infty$ (see Teh et al. (2007) for more details):

$$\tau_{hk}|H, \alpha_\tau \sim \mathcal{B}\left(\frac{\alpha_\tau}{H}, 1\right), \quad h = 1, \ldots, H. \tag{3.4}$$

As shown in Frühwirth-Schnatter (2023), any ESP prior admits a finite generalised CUSP representation. In the context of our Bayesian MFA model, it is obtained by permuting the column index $h$ of the parameters $\theta_{1k}, \ldots, \theta_{Hk}$ according to the decreasing slab probabilities $\tau_{(1k)} > \ldots > \tau_{(Hk)}$ for each cluster $k$. Increasing spike probabilities $\pi_{hk}$ for $h = 1, \ldots, H$, as for a CUSP prior, are obtained for each cluster by defining $\pi_{hk} = 1 - \tau_{(hk)}$. Representation (3.3) allows one to choose an upper limit for the number of factors in each cluster, $H$, and keep it fixed throughout the entire inference procedure. By performing the classification between spike and slab independently for each column, we end up with defining an effective number of active factors $H_k$ in each cluster $k$, which is random a priori and a posteriori, typically smaller than $H$, and varies between clusters.

The CUSP prior of Legramanti et al. (2020) defined in (2.7) can be considered as the limiting case, with $H \to \infty$, of the exchangeable spike-and-slab prior (3.3) on $\theta_{hk}|\tau_{hk}$ where $\tau_{hk}$ arises from prior (3.4). For $H \to \infty$, the hyperparameter $\alpha_\tau$ has the same meaning as the hyperparameter $\alpha_\pi$ of the stick distribution of the CUSP prior. However, while Legramanti et al. (2020) assume a fixed value $\alpha_\pi = 5$, we adapt $\alpha_\tau$ to the data under a generalized beta prime prior,

$$\alpha_\tau \sim \beta'\left(a_\alpha, b_\alpha, 1, H\right). \tag{3.5}$$

This assumption allows a conjugate analysis and simple updating of $\alpha_\tau$, where the number of active factors $H_1, \ldots, H_{K_+}$ in all filled clusters serve as sufficient statistics in a partially marginalised Gibbs step within the MCMC scheme developed in Section 4.

The relationship between the combined priors (3.3) and (3.4) and the CUSP prior holds regardless of the distributions of the spike and the slab, both of which are allowed to depend, respectively, on (random) hyperparameters $\phi_0$ and $\phi_\theta$. Following Legramanti et al. (2020) and Kowal and Canale (2023), we combine the spike and slab distributions $p_{spike}(\theta_{hk}|\phi_0)$ and $p_{slab}(\theta_{hk}|\phi_\theta)$ with the conditionally Gaussian prior $\lambda_{ihk} \,|\, \theta_{hk} \sim \mathcal{N}(0, \theta_{hk})$ for the factor loadings in row $i$ and column $h$ of the cluster $k$ loading matrix. This allows to work out the marginal prior for the $h$th column $\boldsymbol{\lambda}_{hk} = (\lambda_{1hk}, \ldots, \lambda_{phk})^\top$ of the $k$th cluster factor loading matrix for specific spike and slab priors. For example, under the slab prior $\theta_{hk}|a_\theta, b_\theta \sim \mathcal{G}^{-1}(a_\theta, b_\theta)$, a Student-$t$ distribution results for $\boldsymbol{\lambda}_{hk}$, i.e. $\boldsymbol{\lambda}_{hk}|a_\theta, b_\theta \sim t_{2a_\theta}(\mathbf{0}, b_\theta/a_\theta \mathbf{I}_p)$, while $\boldsymbol{\lambda}_{hk}|a_0, b_0 \sim t_{2a_0}(\mathbf{0}, b_0/a_0 \mathbf{I}_p)$ under the spike prior $\theta_{hk}|a_0, b_0 \sim \mathcal{G}^{-1}(a_0, b_0)$. Thus, the full specification of the prior on the factor loadings in each cluster of the MFA model can be formalised as follows:

$$\begin{aligned} &\lambda_{ihk} \,|\, \theta_{hk} \sim \mathcal{N}(0, \theta_{hk}), \quad \text{where } i = 1, \ldots, p, \ h = 1, \ldots, H \text{ and } k = 1, \ldots, K, \\ &\theta_{hk} \,|\, \tau_{hk} \sim \tau_{hk} \mathcal{G}^{-1}(a_\theta, b_\theta) + (1 - \tau_{hk}) \mathcal{G}^{-1}(a_0, b_0), \\ &\tau_{hk}|\alpha_\tau, H \sim \mathcal{B}\left(\frac{\alpha_\tau}{H}, 1\right), \quad \alpha_\tau \sim \beta'\left(a_\alpha, b_\alpha, 1, H\right). \end{aligned} \tag{3.6}$$

Finally, the cluster-specific idiosyncratic variances are assigned the hierarchical prior $\xi_{ik}^2 \sim \mathcal{G}^{-1}(a_\xi, b_{\xi i}), b_{\xi i} \sim \mathcal{G}(a_g, b_{gi})$. Further details on prior choices and hyperparameters are given in the Supplementary material, Appendix A.

For the MCMC estimation, we resort to the usual technique of data augmentation for ESP priors. We introduce $KH$ latent binary indicator variables $\gamma_{hk}$, one for each column $h = 1, \ldots, H$ of the loading matrix $\boldsymbol{\Lambda}_k$ in each cluster $k$, to classify the columns into "active" and "inactive" ones. The indicator $\gamma_{hk}$ takes the value of either zero or one for each column $h = 1, \ldots, H$ and follows the Bernoulli prior $P(\gamma_{hk} = 1|\tau_{hk}) = \tau_{hk}$.

# 4 Posterior computations and MCMC algorithm

## 4.1 Nested Gibbs sampler

Despite the relatively complex nature of the model, with separate factor-analytical models nested within a cluster structure, posterior inference can be done mostly within Gibbs sampling steps. The sampler consists of four major blocks presented in the Algorithm 1. In the first block, the partition $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{K_+}\}$ is updated and the number $K_+$ of non-empty clusters is identified. The second block deals with cluster-specific factor models and consists of two separate steps. In Block 2, step (a), the factor model parameters are updated for every filled cluster and the number of active factors $H_k$ in each cluster $k = 1, \ldots, K_+$ is identified via the non-zero elements in the corresponding columns of the binary indicator matrix $\boldsymbol{\gamma} = \{\gamma_{hk}\}$, where by $\boldsymbol{\gamma}_k$ we denote the row of $\boldsymbol{\gamma}$ that corresponds to cluster $k$. In the same way, $\boldsymbol{\theta}_k = \{\theta_{hk}, h = 1, \ldots, H\}$ denotes the column-specific shrinkage parameters for the factor loadings in cluster $k$. In Block 2, step (b), hyperparameters of the factor model are updated.

In the third block, the new number of mixture components $K \geq K_+$ is sampled and the Dirichlet parameter $\alpha_\eta$ is updated via a random walk Metropolis-Hastings (MH)

**Algorithm 1** Telescoping sampling for the dynamic MFMFA model.

**Block 1.**

(a) Update the partition $\mathcal{C}$ of the data by sampling latent allocation indicators $S_n$ for $n = 1, \ldots, N$, from $\Pr(S_n = k \,|\, \boldsymbol{\eta}_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K, K) \propto \eta_k \mathcal{N}_p(\mathbf{y}_n; \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k), k = 1, \ldots, K$.

(b) Compute the number of observation points in each cluster $N_k = \#\{n \,|\, S_n = k\}$, the number of non-empty components $K_+ = \sum_{k=1}^{K} \mathcal{I}\{N_k > 0\}$, and relabel the components so that the first $K_+$ clusters are non-empty.

**Block 2.**

(a) For each of the filled clusters $k = 1, \ldots, K_+$ run the factor analytical procedure with the spike-and-slab prior on factor loadings, sequentially updating $\mathbf{f}_n^k$, $\boldsymbol{\Lambda}_k$, $\boldsymbol{\Xi}_k$, $\boldsymbol{\theta}_k$, $\boldsymbol{\gamma}_k$ and $\boldsymbol{\tau}_k$. Conditional on $\boldsymbol{\Lambda}_k$ and $\boldsymbol{\Xi}_k$ update cluster means $\boldsymbol{\mu}_k$ for the filled clusters $k = 1, \ldots, K_+$ (see Algorithm 2 for further details).

(b) Update the hyperparameters $\mathbf{b}_\xi = (b_{\xi 1}, \ldots, b_{\xi p})$, $b_\theta$, $b_0$ and $\alpha_\tau$ of the factor-analytical model conditional on $K_+$, $\boldsymbol{\theta}_k$, $\boldsymbol{\gamma}_k$, $\boldsymbol{\Xi}_k$ and $\boldsymbol{\tau}_k$ (see Algorithm 3 for further details).

**Block 3.**

(a) Conditional on the partition $\mathcal{C}$, draw a new value of $K \geq K_+$ from

$$p(K \,|\, \mathcal{C}, \alpha_\eta) \propto p(K) \frac{\alpha_\eta^{K_+} K!}{K^{K_+}(K - K_+)!} \prod_{k=1}^{K_+} \frac{\Gamma(N_k + \alpha_\eta/K)}{\Gamma(1 + \alpha_\eta/K)}, \qquad K = K_+, K_+ + 1, \ldots.$$

(b) Using a random walk MH step, sample $\alpha_\eta \,|\, \mathcal{C}, K$ from

$$p(\alpha_\eta \,|\, \mathcal{C}, K) \propto p(\alpha_\eta) \frac{\alpha_\eta^{K_+} \Gamma(\alpha_\eta)}{\Gamma(N + \alpha_\eta)} \prod_{k=1}^{K_+} \frac{\Gamma(N_k + \alpha_\eta/K)}{\Gamma(1 + \alpha_\eta/K)}.$$

**Block 4.**

(a) If $K > K_+$, add $K - K_+$ empty clusters and sample their cluster means $\boldsymbol{\mu}_k$ and the corresponding factor model parameters, i.e. $\boldsymbol{\Lambda}_k$ and $\boldsymbol{\Xi}_k$ from the priors.

(b) Sample cluster weights from $\boldsymbol{\eta}_K \sim \mathcal{D}(\alpha_\eta/K + N_1, \ldots, \alpha_\eta/K + N_K)$.

step. Finally, in the fourth block, we fill the empty clusters including the parameters of the underlying factor models from the corresponding priors. Thus, the first and third blocks are the standard telescoping sampler clustering steps as described in Frühwirth-Schnatter et al. (2021). As for such an algorithm starting values can be rather influential in forming the chain path, they are carefully chosen to ensure that the sampler moves through all areas of high posterior support, see details in the Supplementary material, Appendix B.1.

Note that in Block 4, step (a) of Algorithm 1, $\boldsymbol{\Lambda}_k$, $\boldsymbol{\Xi}_k$, $\boldsymbol{\theta}_k$, and $\boldsymbol{\tau}_k$ for the added empty clusters are sampled using, respectively, the hyperparameters $\mathbf{b}_\xi = (b_{\xi 1}, \ldots, b_{\xi p})$,

$b_\theta$ and $b_0$ learned in Block 2, step (b) of Algorithm 1 from the $K_+$ filled components. This is a specific feature of the telescoping sampler for MFMs developed in Frühwirth-Schnatter et al. (2021), which ensures that the hyperparameters of the filled components inform the hyperparameters of the empty components.

As mentioned above, a separate factor-analytical procedure needs to be run in Block 2, step (a) for each of the filled clusters $1, \ldots, K_+$ (full details are provided in Algorithm 2). The first three steps of Algorithm 2 are standard Gibbs sampler steps for factor models, with the first step used to update the factors $\mathbf{f}_n^k$ for all observations $n \in \mathcal{C}_k$ assigned to cluster $k$. In the following two steps, factor loadings $\boldsymbol{\lambda}_{ik}$ in the $i$th row of $\boldsymbol{\Lambda}_k$ and idiosyncratic variances $\xi_{ik}^2$ are updated for $i = 1, \ldots, p$. Since classification in Block 1, step (a) of Algorithm 1 is carried out marginalised w.r.t. the factors $\mathbf{f}_n^k$, it is important to update the factors in the first step of Algorithm 2. In this way, factors derived for observations currently assigned to a cluster are used in the subsequent steps of updating the cluster-specific factor loadings and idiosyncratic variances. In step 4 of Algorithm 2, cluster-specific means $\boldsymbol{\mu}_k$ are sampled based on the observations assigned to cluster $k$ and the updated parameters of the cluster-specific factor models.

The remaining steps of Algorithm 2 deal with the classification of the columns of the cluster-specific factor loading matrices, $\{\boldsymbol{\lambda}_{hk}\}, h = 1, \ldots, H$, into "active" (assigned to the slab) and "inactive" (assigned to the spike). As already described in Section 3.2, this is done by introducing a latent binary indicator $\gamma_{hk}$ for each column $h = 1, \ldots, H$ of each matrix $\boldsymbol{\Lambda}_k$, which takes the value of 0 if the corresponding column $\boldsymbol{\lambda}_{hk}$ is assigned to the spike and of 1 if the corresponding column is assigned to the slab. The classification itself is performed in step 5 of Algorithm 2, where the values 0 or 1 are assigned to $\gamma_{hk}$ according to the marginal probabilities of $\boldsymbol{\lambda}_{hk}$ arising from either the spike or the slab distribution. In step 6, slab probabilities $\tau_{hk}$ are updated based on the binary indicators $\gamma_{hk}$. Finally, in step 7, the cluster-specific factor loading variances $\theta_{hk}$ are sampled separately for the columns assigned to the spike and for those assigned to the slab.

Finally, Algorithm 3 describes the procedure in Block 2, step (b) of Algorithm 1, where the hyperparameteres of the factor-analytical models are updated based on the information derived from the $K_+$ filled clusters in Block 2, step (a). Note that in step 5 of Algorithm 3 we marginalise over slab probabilities $\tau_{hk}$ and sample the hyperparameter $\alpha_\tau$ of the ESP prior (3.4) using only the information from the classification of the columns of the factor loading matrices into "active" and "inactive" ones and the corresponding number of active factors in each filled cluster, $H_1, \ldots, H_{K_+}$ (see Frühwirth-Schnatter (2023) for the single factor model solution). As shown in the Supplementary material, Appendix B.2, this still allows a conjugate analysis where the posterior $p(\alpha_\tau \,|\, H_1, \ldots, H_{K_+}, H)$ given in (4.1) is a generalized beta prime distribution, based on the prior (3.5), and the numbers of active factors $H_1, \ldots, H_{K_+}$ in all filled clusters serve as sufficient statistics. A draw from (4.1) is simply obtained as $\alpha_\tau = HZ/(1-Z)$, where the auxiliary variable $Z \sim \mathcal{B}(a_\alpha + H^{++}, b_\alpha + H^{\infty})$ is sampled from a beta distribution.

**Algorithm 2** Details of the step (a) in Block 2 of the Algorithm 1.

**for** ($k$ in $1 : K_+$) **do**

1. Sample $\mathbf{f}_n^k$ for $n : n \in \mathcal{C}_k$ from

$$\mathbf{f}_n^k \,|- \sim \mathcal{N}_H\left((\mathbf{\Phi}_H + \mathbf{\Lambda}_k^T \mathbf{\Xi}_k^{-1} \mathbf{\Lambda}_k)^{-1} \mathbf{\Lambda}_k^T \mathbf{\Xi}_k^{-1} (\mathbf{y}_n - \boldsymbol{\mu}_k), (\mathbf{\Phi}_H + \mathbf{\Lambda}_k^T \mathbf{\Xi}_k^{-1} \mathbf{\Lambda}_k)^{-1}\right),$$

   where $\mathbf{\Xi}_k = \text{diag}(\xi_{1k}^2, \ldots, \xi_{pk}^2)$ and $\mathbf{\Phi}_H = \mathbf{I}_H$.

2. Sample the $i$th row $\boldsymbol{\lambda}_{ik}$ of the $k$th cluster loading matrix for $i$ in $(1, \ldots, p)$ from

$$\boldsymbol{\lambda}_{ik}^\top |- \sim \mathcal{N}_H\left((\mathbf{\Psi}_k^{-1} + \xi_{ik}^{-2} \mathbf{F}_k \mathbf{F}_k^T)^{-1} \mathbf{F}_k \xi_{ik}^{-2} (\mathbf{y}_i - \mathbf{1}_{N_k} \mu_{ik})^T, (\mathbf{\Psi}_k^{-1} + \xi_{ik}^{-2} \mathbf{F}_k \mathbf{F}_k^T)^{-1}\right),$$

   where $\mathbf{\Psi}_k = \text{diag}(\theta_{1k}, \ldots, \theta_{Hk})$, $\mathbf{F}_k = \{\mathbf{f}_n^k : n \in \mathcal{C}_k\}$ is a matrix of factors of the cluster $k$, $\mathbf{y}_i$ is a vector of observations of variable $i$, for which $S_n = k$, and $\mathbf{1}_{N_k}$ is a vector of ones.

3. Sample the diagonal elements $\xi_{ik}^2$ of $\mathbf{\Xi}_k$ for $i$ in $(1, \ldots, p)$ from

$$\xi_{ik}^{-2} |- \sim \mathcal{G}\left(a_\xi + \frac{N_k}{2}, b_{\xi i} + \frac{1}{2} \sum_{n : n \in N_k} (y_{in} - \mu_{ik} - \boldsymbol{\lambda}_{ik} \mathbf{f}_n^k)^2\right).$$

4. Update the cluster-specific mean $\boldsymbol{\mu}_k$ from $\boldsymbol{\mu}_k \,|- \sim \mathcal{N}_p(\mathbf{b}_k, \mathbf{B}_k)$, where

$$\mathbf{b}_k = \mathbf{B}_k\Big(\mathbf{B}_0^{-1} \mathbf{b}_0 + \mathbf{\Xi}_k^{-1} \sum_{n : n \in \mathcal{C}_k} (\mathbf{y}_n - \mathbf{\Lambda}_k \mathbf{f}_n)\Big), \qquad \mathbf{B}_k = (\mathbf{B}_0^{-1} + N_k \, \mathbf{\Xi}_k^{-1})^{-1}.$$

5. Sample the binary indicators $\gamma_{hk}$ for each column $\{\boldsymbol{\lambda}_{hk}\}$, $h = 1, \ldots, H$ of the loading matrix as

$$\begin{aligned} P(\gamma_{hk} = 0 | \boldsymbol{\lambda}_{hk}, b_0, \alpha_\tau, H) &\propto \frac{H}{\alpha_\tau + H} t_{2a_0}(\boldsymbol{\lambda}_{hk}; \mathbf{0}, (b_0/a_0)\mathbf{I}_p), \\ P(\gamma_{hk} = 1 | \boldsymbol{\lambda}_{hk}, b_\theta, \alpha_\tau, H) &\propto \frac{\alpha_\tau}{\alpha_\tau + H} t_{2a_\theta}\left(\boldsymbol{\lambda}_{hk}; \mathbf{0}, (b_\theta/a_\theta)\mathbf{I}_p\right). \end{aligned}$$

6. Sample the (unordered) slab probabilities $\tau_{hk}$ for $h$ in $(1, \ldots, H)$:

$$\tau_{hk} | \gamma_{hk} \sim \mathcal{B}\left(\frac{\alpha_\tau}{H} + \gamma_{hk}, 2 - \gamma_{hk}\right).$$

7. Given $\gamma_{hk}$ and the $h$th column $\boldsymbol{\lambda}_{hk}$ of the loading matrix, for each $h$ in $(1, \ldots, H)$, sample $\theta_{hk} | \gamma_{hk}, \boldsymbol{\lambda}_{hk}$ depending on $\gamma_{hk}$:

$$\begin{aligned} \theta_{hk} | \gamma_{hk} = 0, \boldsymbol{\lambda}_{hk} &\sim \mathcal{G}^{-1}\left(a_0 + \frac{1}{2}p, b_0 + \frac{1}{2}\sum_{i=1}^p \lambda_{ihk}^2\right), \\ \theta_{hk} | \gamma_{hk} = 1, \boldsymbol{\lambda}_{hk} &\sim \mathcal{G}^{-1}\left(a_\theta + \frac{1}{2}p, b_\theta + \frac{1}{2}\sum_{i=1}^p \lambda_{ihk}^2\right). \end{aligned}$$

**end for**

**Algorithm 3** Details of the step (b) in Block 2 of the Algorithm 1.

1: Sample $b_{\xi i}$ for $i = 1, \ldots, p$ from

$$b_{\xi i} \,|\, - \sim \mathcal{G}\left(a_g + K_+ a_\xi, b_{gi} + \sum_{k=1}^{K_+} \frac{1}{\xi_{ik}^2}\right).$$

2: Calculate the effective number of "active" columns $H_k$ in cluster $k$ as $H_k = \sum_{h=1}^{H} \gamma_{hk}$. Define $H^{++} = \sum_{k=1}^{K_+} H_k$ as the total number of "active" columns in all filled clusters and $H^{\infty} = HK_+ - H^{++}$ as the total number of "inactive" columns in all filled clusters.

3: Sample $b_0$ from

$$b_0 | - \sim \mathcal{G}\left(a_1 + H^{\infty} a_0, b_1 + \sum_{k=1}^{K_+} \sum_{h:\gamma_{hk}=0} \frac{1}{\theta_{hk}}\right).$$

4: Sample $b_\theta$ from

$$b_\theta | - \sim \mathcal{G}\left(a_2 + H^{++} a_\theta, b_2 + \sum_{k=1}^{K_+} \sum_{h:\gamma_{hk}=1} \frac{1}{\theta_{hk}}\right).$$

5: Sample $\alpha_\tau \,|\, H_1, \ldots, H_{K_+}, H$ from a generalized beta prime distribution:

$$\alpha_\tau \,|\, H_1, \ldots, H_{K_+}, H \propto \beta'\left(a_\alpha + H^{++}, b_\alpha + H^{\infty}, 1, H\right). \tag{4.1}$$

## 4.2 Post-processing and stratification

Before any further inference is performed, the MCMC output should first undergo a post-processing to resolve the label switching problem inherent in Bayesian inference for mixture models, to ensure correct grouping of the model components into clusters. We estimate the number of active components $\widehat{K}_+$ as the mode of the posterior distribution $p(K_+|\boldsymbol{y})$ derived from all MCMC draws and keep only those draws where the realisation of $K_+$ is equal to $\widehat{K}_+$. For these draws, label switching between clusters is accounted for by applying $k$-means clustering to the point process representation of the MCMC draws in the spirit of Frühwirth-Schnatter (2011), see Appendix B.3 in the Supplementary material for details on this procedure.

For inference on the parameters of cluster-specific factor models, these draws are processed further. First, we estimate the number of factors $\widehat{H}_k$ in each of the $K_+$ clusters as the mode of the corresponding posterior distribution. In the second step, we remove all draws in which the number of active factors $H_k$ is not equal to $\widehat{H}_k$ in any cluster. For the remaining draws, the cluster-specific factor loading matrices exhibit $\widehat{H}_k$ active and $H - \widehat{H}_k$ inactive columns (with the respective indicators $\gamma_{hk} = 1$ and $\gamma_{hk} = 0$), and we keep only the active columns for further inference. These sparsified cluster-

| Parameter(s) | Hyperparameters | Hyperparameter values |
|---|---|---|
| $\boldsymbol{\mu}_k$ | $\mathbf{b}_0, \mathbf{B}_0$ | $\text{median}(\mathbf{y}), \text{diag}(R_1^2, \ldots, R_p^2)$ |
| $K$ | $\alpha_\lambda, \alpha_\pi, \beta_\pi$ | $1, 4, 3$ |
| $\alpha_\eta$ | $\nu_l, \nu_r$ | $6, 3$ |
| $\alpha_\tau$ | $a_\alpha, b_\alpha$ | $6, 2H+1$ |
| $\xi_{ik}^2$ | $a_\xi, a_g, b_{gi}$ | $1, 3, 100/R_i^2$ |
| $\theta_{hk}\|\gamma_{hk}=1$ | $a_\theta, a_2, b_2$ | $3, 2, 1$ |
| $\theta_{hk}\|\gamma_{hk}=0$ | $a_0, a_1, b_1$ | $21, 1, 1$ |

Table 1: Hyperparameter specifications for the MFMFA model. In the prior for $\boldsymbol{\mu}_k$, $R_i$ denotes the range of the data in dimension $i$.

specific factor loading matrices $\boldsymbol{\Lambda}_k$ with $\hat{H}_k$ columns are then used for the calculation of cluster-specific covariance matrices $\boldsymbol{\Omega}_k = \boldsymbol{\Lambda}_k \boldsymbol{\Lambda}_k^\top + \boldsymbol{\Xi}_k$, see Section 5.

However, the cluster-specific factor loading matrices $\boldsymbol{\Lambda}_k$ are not uniquely identified, since we did not impose any constraints on these matrices during the MCMC sampling to prevent rotational invariance. Rather, we apply the MatchAlign algorithm introduced in Poworoznek et al. (2021) independently in each cluster $k$ to resolve rotational invariance for the draws of $\boldsymbol{\Lambda}_k$ in a post-processing step. More details of this approach are discussed in Appendix B.3 of the Supplementary material.

## 5 Simulation study

The performance of the dynamic MFMFA model is first demonstrated on simulation studies. We consider several different settings to assess the ability of the model to correctly infer cluster and factor dimensionality of the data. In Section 5.1, we demonstrate the performance of the model on a range of various settings $p$, $N$ and $H_k$ on data sets with balanced cluster sizes and common number of factors. We consider only $N > p$ settings as this is a precondition for a stable performance of a model based on a mixture of Gaussian distributions. In the case with $p > N$, the cluster covariance matrices are not well defined, which leads to difficulties in both the clustering assignment and factor model performance. The simulation study in Section 5.2 is more challenging with the number of cluster-specific factors set to $H_k = 0$ in some clusters and factor-loading matrices exhibiting a distinct structure in others. The hyperparameter specifications are reported in Table 1. The maximum number of factors $H$ is equal to the largest integer that satisfies the variance identification condition of Anderson and Rubin (1956), $H \leq (p-1)/2$. Unless otherwise specified, the sampler is run for $50,000$ iterations, with 20% of them discarded as burn-in. In both sections, we compare the performance of the dynamic MFMFA model with the IMIFA model of Murphy et al. (2020) for all simulated data sets.

Cluster performance is assessed using the adjusted Rand index (ARI; Hubert and Arabie (1985)) and the misclassification rate is estimated as the percentage of mislabelled observations compared to the true cluster labels used to simulate the data.

To assess the accuracy of the model in estimating the true cluster-specific covariance matrices $\boldsymbol{\Omega}_k^0 = \boldsymbol{\Lambda}_k^0(\boldsymbol{\Lambda}_k^0)^\top + \boldsymbol{\Xi}_k^0$ of the data via the estimated covariance matrices

$$\widehat{\boldsymbol{\Omega}}_k = \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\Omega}_k^{(m)}, \quad \boldsymbol{\Omega}_k^{(m)} = \boldsymbol{\Lambda}_k^{(m)}(\boldsymbol{\Lambda}_k^{(m)})^\top + \boldsymbol{\Xi}_k^{(m)},$$

where $\boldsymbol{\Lambda}_k^{(m)}$ and $\boldsymbol{\Xi}_k^{(m)}$ are the $m$-th among the $M$ posterior draws passing the post-processing procedure described in Section 4.2, we compute for each simulation in each of the scenarios a Monte-Carlo estimate of the mean squared error (MSE) defined by

$$MSE_{\Omega_k} = \sum_{i=1}^{p}\sum_{l=i}^{p}\mathbb{E}((\Omega_{k,il} - \Omega_{k,il}^0)^2 \,|\, \mathbf{y})/(p(p+1)/2).$$

Following Murphy et al. (2020), the data are standardised before being fed into both models. More specifically, the data are transformed as $\tilde{\mathbf{y}}_n = \mathbf{S}^{-1}(\mathbf{y}_n - \mathbf{m})$, where $\mathbf{m}$ denotes the vector of means of $\mathbf{y}$ and the scale matrix $\mathbf{S} = \sqrt{\text{diag}(S_{y,1}\dots S_{y,p})}$ is defined from the empirical variances $S_{y,i}$ of the data $y_{in}$ over $n = 1,\dots,N$. Given the true cluster-specific factor loading and covariance matrices $\boldsymbol{\Lambda}_k^0$ and $\boldsymbol{\Omega}_k^0$ of the original data, the corresponding matrices then take the form $\mathbf{S}^{-1}\boldsymbol{\Lambda}_k^0$ and $\mathbf{S}^{-1}\boldsymbol{\Omega}_k^0\mathbf{S}^{-1}$ for the transformed data.

## 5.1 Simulation study 1

The purpose of this simulation study is to evaluate the performance of our model on data sets of various sizes, i.e., under various settings of $p$, $N$ and $H_k$, with the clusters approximately equally sized, but not very well separated from each other, and the same number of factors in each cluster. Three different settings of $(p, N, H_k)$ are considered, namely $(20, 100, 4)$, $(30, 200, 7)$, and $(50, 500, 10)$, to evaluate the performance of the model against an increasing dimensionality of the data set.

The data are simulated with $K_+ = 3$ clusters and cluster weights $\boldsymbol{\eta}_3 = (1/3, 1/3, 1/3)$. The model parameters are simulated as $\xi_{ik}^2 \sim \mathcal{G}^{-1}(2,1)$ and $\boldsymbol{\lambda}_{ik}^\top \sim \mathcal{N}_4(\mathbf{0}, \mathbf{I}_4)$ for all $i$ and $k$. To ensure that clusters are overlapping, the means are generated similarly to Murphy et al. (2020) as $\boldsymbol{\mu}_k \sim \mathcal{N}_p((2k-4)\mathbf{1}, \mathbf{I}_p)$. Each $\mathbf{y}_n$, $n = 1,\dots,N$, is simulated from a conditional mixture model with $\mathbf{f}_n^k \sim \mathcal{N}_4(\mathbf{0}, \mathbf{I}_4)$:

$$p(\mathbf{y}_n \,|\, \{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \mathbf{f}_n^k, \boldsymbol{\Xi}_k\}, \boldsymbol{\eta}_{K_+}) = \sum_{k=1}^{K_+}\eta_k \mathcal{N}_p\left(\mathbf{y}_n; \boldsymbol{\mu}_k + \boldsymbol{\Lambda}_k\mathbf{f}_n^k, \boldsymbol{\Xi}_k\right).$$

To test the robustness of the model performance, each simulation setting was replicated 25 times and each time a new data set was generated. As the cluster-specific parameters $\boldsymbol{\Lambda}_k$ and $\boldsymbol{\Xi}_k$ could induce some degree of separation between clusters, pairwise scatter plots of a randomly chosen data set are shown in Figure 1 to demonstrate the extent of overlap between clusters. For the sake of clear visibility, five randomly chosen variables from a data set with $p = 20$ variables and $N = 100$ observations are depicted. For the

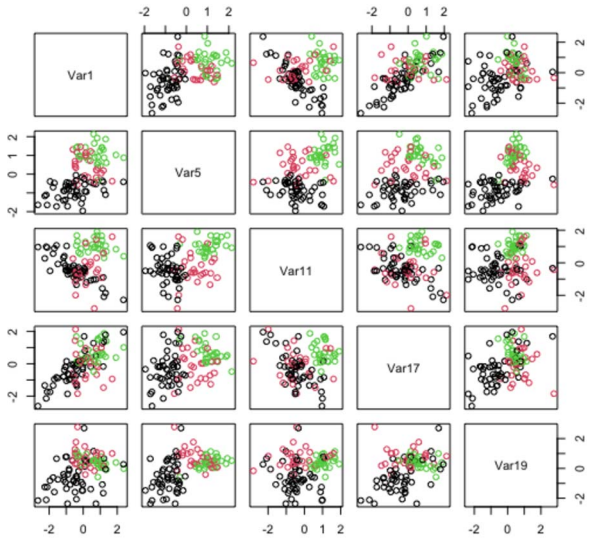



Figure 1: Pairwise scatter plots of five randomly chosen variables from one of the replicate data sets with $p = 20$ and $N = 100$ demonstrating the overlap between the three clusters.

$(50, 500, 10)$ setting we ran the sampler for $75,000$ iterations as convergence was slower to achieve for a larger data set. As in other settings, 20% of the draws were discarded as burn-in.

To provide some benchmark, we compare the performance of our dynamic MFMFA model with the IMIFA model of Murphy et al. (2020) by running the R-package `IMIFA` on exactly the same 25 simulated data sets. As the `IMIFA` R package offers the full family of models, we restricted the choice of model to the most flexible MFA model, namely the IMIFA, to ensure better comparability, and set the PYP parameters of the IMIFA model, namely $\sigma$ and $d$, to be learned from data. The hyperparameters of the MGP prior are left as default, i.e. $a_1 = 2.1$ and $a_2 = 3.1$. We run the IMIFA model for exactly the same number of iterations as the dynamic MFMFA model.

The results provided in Table 2 demonstrate that the dynamic MFMFA model performs well for all three settings of $(p, N, H_k)$, exhibiting the capability to uncover the true structure of the simulated data in most cases. The partition has been identified correctly in all replicates for all settings of $(p, N, H_k)$. The number of cluster-specific factors was occasionally slightly overestimated for the larger settings of $(p, N, H_k)$, however, in most cases it was recovered correctly. In general, the model exhibited stable performance both on small- and medium-sized data sets. The MSEs of the cluster-specific covariance matrices provided in Table 3 are in all three cases very small, with the estimation becoming more exact as the size of the data set grows. This shows that the model generally correctly recovered the data structure even in the very few cases where the number of cluster-specific factors was slightly under- or overestimated.

| $p, N, H_k$ | Model | $\widehat{K}_+$ | $\widehat{H}_1$ | $\widehat{H}_2$ | $\widehat{H}_3$ | ARI |
|---|---|---|---|---|---|---|
| $(20, 100, 4)$ | MFMFA | 3 [3, 3] | 4 [4, 4] | 4 [4, 4] | 4 [4, 4] | 1 [1, 1] |
| | IMIFA | 3 [3, 3.1] | 5 [4.2, 5] | 5 [5, 5] | 5 [5, 5] | 1 [1, 1] |
| $(30, 200, 7)$ | MFMFA | 3 [3, 3] | 7 [7, 7.8] | 7 [7, 8] | 7 [7, 8] | 1 [1, 1] |
| | IMIFA | 3 [3, 3] | 8 [7.6, 8] | 8 [7.6, 8] | 8 [7, 8] | 1 [1, 1] |
| $(50, 500, 10)$ | MFMFA | 3 [3, 3] | 10 [10, 11] | 10 [10, 11.8] | 10 [10, 12] | 1 [1, 1] |
| | IMIFA | 3 [3, 3] | 11 [11, 11] | 11 [11, 11] | 11 [11, 11] | 1 [1, 1] |

Table 2: Simulation results for the dynamic MFMFA and the IMIFA models under different settings for the number of variables $p$, observations $N$ and latent factors in each cluster $H_k$. The true number of clusters is $K_+ = 3$. The performance of $\widehat{K}_+$ and $\widehat{H}_k$ is based on modal estimates. For each performance measure, we report the median, and the 5% and the 95% quantiles (in brackets) across 25 simulated data sets. Clustering performance is assessed via the ARI against the known cluster labels.

| $p, N, H_k$ | $MSE_{\Omega_1}$ | $MSE_{\Omega_2}$ | $MSE_{\Omega_3}$ |
|---|---|---|---|
| $(20, 100, 4)$ | 0.027 [0.017,0.046] | 0.027 [0.020,0.036] | 0.032 [0.021,0.064] |
| $(30, 200, 7)$ | 0.017 [0.011,0.022] | 0.016 [0.011,0.021] | 0.016 [0.012,0.021] |
| $(50, 500, 10)$ | 0.006 [0.006,0.009] | 0.007 [0.006,0.009] | 0.007 [0.006,0.009] |

Table 3: Cluster-specific MSEs of the covariance matrices estimated by the dynamic MFMFA model under three different setting of $p, N, H_k$.

The clustering performance of the IMIFA model is also excellent. Only in one case out of 25, the IMIFA model estimated 5 instead of 3 clusters. In the factor-analytical part, however, it shows a tendency to slightly overestimate the number of cluster-specific factors compared to the dynamic MFMFA model. This can be attributed to the intrinsic inefficiency of the MGP prior described in Section 2.2. Cluster-specific MSEs for the IMIFA model are not reported as the output provided by the `IMIFA` package does not easily facilitate calculation of the MSE as described in Section 5.

## 5.2 Simulation study 2

In this section, we present a simulation study which illustrates settings where the dynamic MFMFA model performs clearly better than the IMIFA model. We consider two different scenarios to assess the comparative performance of the two models in challenging settings, i.e. small data sets with overlapping clusters, structured factor loading matrices, and diagonal covariance matrices in some of the clusters. For each scenario, we generate 25 replicate data sets.

In the first scenario, we generate a data set with $p = 20$ variables, $T = 100$ observations, and 4 equally weighted clusters of which two contain $H_k = 4$ factors each and two contain zero factors (which means that the cluster-specific covariance matrices in these cases are diagonal). The idiosyncratic errors are generated from $\mathcal{N}(0, 1)$. The non-zero

| Scenario | Model | $\widehat{K}_+$ | $\widehat{H}_1$ | $\widehat{H}_2$ | $\widehat{H}_3$ | $\widehat{H}_4$ | ARI | Err. (%) |
|---|---|---|---|---|---|---|---|---|
| Sc. 1 | MFMFA | 4 [4, 4] | 3 [2, 4] | 3 [2, 4] | 0 [0, 0] | 0 [0, 0] | 1 [1, 1] | 0 [0, 0] |
| | IMIFA | 4 [4, 9.8] | 4 [3.9, 5] | 5 [4, 5] | 2 [0, 2] | 2 [0, 2] | 1 [0.79, 1] | 0 [0, 21.2] |
| Sc. 2 | MFMFA | 4 [4, 4.8] | 5 [5, 5] | 5 [5, 5] | 0 [0, 0] | 0 [0, 0] | 1 [0.98, 1] | 0 [0, 2] |
| | IMIFA | 4 [4, 7.4] | 6 [6, 6] | 6 [6, 6] | 1 [0, 2] | 0 [0, 2] | 1 [0.92, 1] | 0 [0, 8.9] |

Table 4: Simulation results for two different scenarios. The true number of clusters is $K_+ = 4$. The true number of factors is $H_1 = H_2 = 4$ and $H_3 = H_4 = 0$ for scenario 1, and $H_1 = H_2 = 5$ and $H_3 = H_4 = 0$ for scenario 2. The performance of $\widehat{K}_+$ and $\widehat{H}_k$ is based on modal estimates. For each performance measure, we report the median, and the 5% and the 95% quantiles (in brackets) across 25 simulated data sets. Clustering performance is assessed via the ARI and the percentage error rate against the known cluster labels.

factors are generated as $\mathbf{f}_n^k \sim \mathcal{N}_4(\mathbf{0}, \mathbf{I}_4)$. The factor loadings are generated in a similar manner to one of the simulation study scenarios in Frühwirth-Schnatter et al. (2023). More specifically, they form two blocks, with the first 10 variables loading on factors 1 and 2 and the remaining 10 variables loading on factors 3 and 4. The non-zero factor loadings are generated as $\lambda_{ihk} = (-1)^{b_{ihk}}(2 + 0.1\mathcal{N}(0,1))$, where the exponent $b_{ihk}$ is a binary variable with $\Pr(b_{ihk} = 1) = 0.3$. The maximum possible number of factors is set as $H \leq (p-1)/2$, which in this case is 9. We run the sampler for $75,000$ iterations and discard 30% of them as burn-in.

The second scenario contains a data set with $p = 20$ variables, $T = 120$ observations, and 4 equally weighted clusters of which two contain each $H_k = 5$ factors and two contain zero factors. As in the previous scenario, the idiosyncratic errors are generated from $\mathcal{N}(0,1)$, and the non-zero latent factors are generated as $\mathbf{f}_n^k \sim \mathcal{N}_5(\mathbf{0}, \mathbf{I}_5)$. The factor loadings are also generated in a similar way to the previous scenario, with a difference that in the clusters with a non-diagonal covariance matrix, one factor (a so-called market factor) has significant loadings on all 20 variables. The remaining four factors have significant loadings assigned in the following way: the first factor has significant loadings on variables 1–6, the second on variables 6–11, the third on variables 11–16, and the fourth on variables 16–20. Again, as in the previous scenario, the maximum possible number of factors is set as $H \leq (p-1)/2$ and the sampler is run for $75,000$ iterations with 30% of them discarded as burn-in.

We run the IMIFA model on exactly the same 25 replicate data sets for each scenario, with the same number of iterations as the dynamic MFMFA model and the model settings as described in Section 5.1. The results provided in Table 4 show a distinct tendency of the IMIFA model to struggle with identifying clusters with diagonal covariance matrices. This is especially evident in the first scenario, where the median number of factors in such clusters is two. Also classification results are better in the case of the dynamic MFMFA model: it delivers perfect classification in all 25 replicate data sets in scenario 1, and overestimated the number of clusters only in two cases out of 25 in scenario 2 estimating five and six clusters, respectively. The IMIFA model overestimated the number of clusters six times in scenario 1 and four times in scenario 2, identifying

up to 13 clusters. This higher frequency of misclassifications of IMIFA is also reflected in the 5% and 95% quantile estimations of the ARI (0.79 and 0.92 for the 5% quantile) and error rate (21.2% and 8.9% for the 95% quantile).

To conclude, the dynamic MFMFA model exhibits more reliable performance than the IMIFA model on data sets which are small in both the number of variables and observations, and contain clusters with zero factors (diagonal covariance matrices). However, when data dimensions increase, the difference in performance between the two models becomes less pronounced, and both models deliver comparable results with IMIFA only slightly tending to overestimate the number of cluster-specific factors. Additional simulation studies in Appendix C of the Supplementary material confirm these results on data with differently sized clusters and varying number of cluster-specific factors.

## 6 Italian olive oils data set

The Italian olive oils data set (Forina et al., 1983) has been widely used in the literature for testing clustering and MFA models (see, e.g. Murphy et al. (2020)) and is available in the R package `FlexDir`. The data describe the composition of eight fatty acids in $N = 572$ Italian olive oils, which originate from three areas: northern Italy, Sardinia, and southern Italy. Each area breaks down into several regions: northern Italy comprises Umbria and east and west Liguria; Sardinia is divided into inland and coastal Sardinia; and southern Italy is composed of north Apulia, Calabria, south Apulia, and Sicily. This induces us to presume that classification should most likely reflect a combination of area and/or region structure with the number of clusters arguably between three and nine, or anyway not exceeding nine by far. Due to a rather small number of $p = 8$ variables in the data set, the initial number of cluster-specific factors in the MFMFA algorithm, which is usually set at $H = \lfloor (p-1)/2 \rfloor$, was replaced by $H = p$.

We compare the performance of our model with the three other MFA models. The first model is the IMIFA model, which is the only one that can be compared with our model in terms of flexibility, as it also provides fully automatic inference on the number of clusters and cluster-specific factors and allows the number of factors to differ between clusters. The other two models, namely, parsimonious Gaussian mixture models (McNicholas and Murphy, 2010), fit via the R package `pgmm` (McNicholas et al., 2018), and overfitting Bayesian mixtures of factor analysers (Papastamoulis, 2020), fit via the R package `fabMix`, only allow the same number of factors in each cluster. The two latter models represent a collection of models with different restrictions on the covariance structure (see details in McNicholas and Murphy (2010) and Papastamoulis (2020)). To ensure the best possible comparability, we chose the fully unrestricted model (denoted as "UUU" model in both papers) which represents an MFA covariance structure equivalent to the one in the dynamic MFMFA model. For both the parsimonious Gaussian mixture model and the IMIFA model, we chose the $k$-means option for initial clustering, as the same way of initialising clustering was employed in the MFMFA model. For the sake of simplicity, we will refer to the models by the names of the R packages which were used to fit them. The data are standardised before being fed into all models, and the MCMC sampler is run for $50,000$ iterations with 20% of them discarded as burn-in. For the

| Model | # clusters | # factors | ARI (areas) | ARI (regions) |
|---|---|---|---|---|
| Dynamic MFMFA | 4 | 5, 3, 6, 6 | 0.89 | 0.55 |
| IMIFA | 6 | 3, 4, 2, 3, 6, 6 | 0.56 | 0.79 |
| *pgmm* | 5 | 5 | 0.59 | 0.76 |
| *fabMix* | 5 | 4 | 0.59 | 0.76 |

Table 5: Results of fitting the dynamic MFMFA model against a range of benchmark MFA models on the Italian olive oils data set. Note that the number of factors is estimated to be same for all clusters in `pgmm` and `fabMix` by model design.

| | Dynamic MFMFA | | | |
|---|---|---|---|---|
| Cluster | 1 | 2 | 3 | 4 |
| Northern Italy | 151 | 0 | 0 | 0 |
| Sardinia | 0 | 98 | 0 | 0 |
| Southern Italy | 0 | 0 | 28 | 295 |

| | IMIFA | | | | | |
|---|---|---|---|---|---|---|
| Cluster | 1 | 2 | 3 | 4 | 5 | 6 |
| Northern Italy | 42 | 61 | 48 | 0 | 0 | 0 |
| Sardinia | 0 | 0 | 0 | 98 | 0 | 0 |
| Southern Italy | 0 | 0 | 0 | 0 | 128 | 195 |

| | *pgmm* | | | | |
|---|---|---|---|---|---|
| Cluster | 1 | 2 | 3 | 4 | 5 |
| Northern Italy | 88 | 63 | 0 | 0 | 0 |
| Sardinia | 0 | 0 | 98 | 0 | 0 |
| Southern Italy | 0 | 0 | 0 | 122 | 201 |

| | *fabMix* | | | | |
|---|---|---|---|---|---|
| Cluster | 1 | 2 | 3 | 4 | 5 |
| Northern Italy | 60 | 91 | 0 | 0 | 0 |
| Sardinia | 0 | 0 | 98 | 0 | 0 |
| Southern Italy | 0 | 0 | 0 | 122 | 201 |

Table 6: Confusion matrices between the estimated and presumed cluster assignments to three areas of the Italian olive oils data set. The estimated cluster assignments are provided by the dynamic MFMFA, `IMIFA`, `pgmm` and `fabMix` models.

dynamic MFMFA model, the same values of hyperparameters were used as reported in Table 1 for the simulated data.

Table 5 presents the results of applying all four MFA models to the Italian olive oils data, while Table 6 and Table 7 present the confusion matrices between the estimated partition and that into areas or regions, respectively. Interestingly, clustering performance exhibited by the dynamic MFMFA model is closer to the partition into three larger areas, while the clustering performance exhibited by the other three models is closer to the partition to the nine smaller regions. With respect to the partition into areas, the dynamic MFMFA delivered a classification with an ARI of 0.89. Only southern Italy was divided into two clusters, with most of the data points corresponding

| | Dynamic MFMFA | | | |
|---|---|---|---|---|
| Cluster | 1 | 2 | 3 | 4 |
| Umbria | 51 | 0 | 0 | 0 |
| East Liguria | 50 | 0 | 0 | 0 |
| West Liguria | 50 | 0 | 0 | 0 |
| Inland Sardinia | 0 | 65 | 0 | 0 |
| Coastal Sardinia | 0 | 33 | 0 | 0 |
| North Apulia | 0 | 0 | 23 | 2 |
| South Apulia | 0 | 0 | 0 | 206 |
| Calabria | 0 | 0 | 0 | 56 |
| Sicily | 0 | 0 | 5 | 31 |

| | IMIFA | | | | | |
|---|---|---|---|---|---|---|
| Cluster | 1 | 2 | 3 | 4 | 5 | 6 |
| Umbria | 39 | 11 | 0 | 0 | 0 | 0 |
| East Liguria | 0 | 50 | 0 | 0 | 0 | 0 |
| West Liguria | 3 | 0 | 48 | 0 | 0 | 0 |
| Inland Sardinia | 0 | 0 | 0 | 65 | 0 | 0 |
| Coastal Sardinia | 0 | 0 | 0 | 33 | 0 | 0 |
| North Apulia | 0 | 0 | 0 | 0 | 25 | 0 |
| South Apulia | 0 | 0 | 0 | 0 | 11 | 195 |
| Calabria | 0 | 0 | 0 | 0 | 56 | 0 |
| Sicily | 0 | 0 | 0 | 0 | 36 | 0 |

| | *pgmm* | | | | |
|---|---|---|---|---|---|
| Cluster | 1 | 2 | 3 | 4 | 5 |
| Umbria | 51 | 0 | 0 | 0 | 0 |
| East Liguria | 37 | 13 | 0 | 0 | 0 |
| West Liguria | 0 | 50 | 0 | 0 | 0 |
| Inland Sardinia | 0 | 0 | 65 | 0 | 0 |
| Coastal Sardinia | 0 | 0 | 33 | 0 | 0 |
| North Apulia | 0 | 0 | 0 | 25 | 0 |
| South Apulia | 0 | 0 | 0 | 12 | 194 |
| Calabria | 0 | 0 | 0 | 56 | 0 |
| Sicily | 0 | 0 | 0 | 36 | 0 |

| | *fabMix* | | | | |
|---|---|---|---|---|---|
| Cluster | 1 | 2 | 3 | 4 | 5 |
| Umbria | 10 | 40 | 0 | 0 | 0 |
| East Liguria | 50 | 0 | 0 | 0 | 0 |
| West Liguria | 0 | 51 | 0 | 0 | 0 |
| Inland Sardinia | 0 | 0 | 65 | 0 | 0 |
| Coastal Sardinia | 0 | 0 | 33 | 0 | 0 |
| North Apulia | 0 | 0 | 0 | 25 | 0 |
| South Apulia | 0 | 0 | 0 | 9 | 197 |
| Calabria | 0 | 0 | 0 | 55 | 1 |
| Sicily | 0 | 0 | 0 | 33 | 3 |

Table 7: Confusion matrices between the estimated and presumed cluster assignments to nine regions of the Italian olive oils data set. The estimated cluster assignments are provided by the dynamic MFMFA, `IMIFA`, `pgmm` and `fabMix` models.

to the north Apulia region being placed in a separate cluster. In terms of partitioning into smaller regions, the closest clustering performance was delivered by IMIFA with an ARI of 0.79. Most data points corresponding to three northern regions were placed into separate clusters, while north Apulia, Calabria and Sicily were all placed in one cluster. The other two models, namely `pgmm` and `fabMix`, also placed north Apulia, Calabria and Sicily mostly in one cluster while northern Italy was split between two clusters. All models put Sardinia into a separate group, which probably indicates that olive oils from that region have distinguishably different chemical characteristics.

Regarding the number of cluster-specific factors, the results of all four models were close. Both the dynamic MFMFA and IMIFA model estimated three factors in the cluster corresponding to Sardinia, and six factors in the two clusters corresponding to southern Italy. For northern Italy, however, the results differ. The dynamic MFMFA

model puts the whole area into one cluster with five cluster-specific factors, while IMIFA divided this area into three clusters with the number of factors ranging between two and four. The `pgmm` and `fabMix` models estimated five and four factors, respectively.

# 7 Conclusion

We proposed a novel model in the MFA framework which allows fully automatic inference on the number of non-empty components in the mixture and the number of latent factors in the cluster-specific factor-analytical models while keeping both dimensions finite. This was done by employing the connection between nonparametric Bayesian priors and their finite-dimensional representations, connecting the MFM framework with the ESP class of priors (Frühwirth-Schnatter, 2023) in the factor-analytical part. This approach allowed to eliminate some of the drawbacks of the popular MGP prior. Posterior inference is performed mostly within Gibbs sampler steps and all influential parameters are learned from the data, which makes it possible to use the dynamic MFMFA model on various data sets with no or little additional tuning. Some hyperparameter tuning may become necessary when working with data sets of essentially different nature. However, the hyperparameter values we provided proved to be rather universal and can be employed for data sets of various sizes and structures.

Future research directions could include, for instance, the introduction of element-wise shrinkage for the columns of cluster-specific factor loading matrices, which could help to achieve more exact identification in sparse factor models. Moreover, the triple gamma prior (Cadonna et al., 2020) could be employed for the spike and the slab distributions instead of the inverse gamma priors (Frühwirth-Schnatter, 2023). This could improve mixing, as well as uncertainty quantification for the number of cluster-specific factors. Another interesting and promising direction of research would be to relax the assumption of independent factors, which is particularly restrictive in the case of time-series data, where observations at different points of time are often correlated.

#### Acknowledgments

The authors thank anonymous reviewers and the associate editor for their valuable comments and suggestions that helped significantly improve this work. The authors also thank Dr. Leonardo Bottolo for his helpful comments on an earlier version of this paper.

# Supplementary Material

Supplementary material for: "Dynamic Mixture of Finite Mixtures of Factor Analysers" (DOI: 10.1214/25-BA1543SUPP; .pdf). This supplement contains additional details on the choice of priors and hyperparameters as well as the initialisations of the MCMC algorithm and the postprocessing procedure. The supplement also contains additional simulation study, applications on real data, and a chapter on identification in cluster-specific factor models. Code for the MCMC algorithm is available at `https://github.com/margotgrush/Dynamic_MFMFA`.

## References


Alamichel, L., Bystrova, D., Arbel, J., and King, G. K. K. (2024). "Bayesian mixture models (in)consistency for the number of clusters." *Scandinavian Journal of Statistics*, 51: 1619–1660. MR4818761. doi: https://doi.org/10.1111/sjos.12739. 2

Anderson, T. and Rubin, H. (1956). "Statistical inference in factor analysis." *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume V: 111–150. MR0084943. 18

Ascolani, F., Lijoi, A., Rebaudo, G., and Zanella, G. (2023). "Clustering consistency with Dirichlet process mixtures." *Biometrika*, 110: 551–558. MR4589070. doi: https://doi.org/10.1093/biomet/asac051. 5

Bhattacharya, A. and Dunson, D. (2011). "Sparse Bayesian infinite factor models." *Biometrika*, 98(2): 291–306. MR2806429. doi: https://doi.org/10.1093/biomet/asr013. 2, 6, 7

Cadonna, A., Frühwirth-Schnatter, S., and Knaus, P. (2020). "Triple the gamma – A unifying shrinkage prior for variance and variable selection in sparse state space and TVP models." *Econometrics*, 8(2): 1–36. 26

De Blasi, P., Favaro, S., Lijoi, A., Mena, R., and Ruggiero, M. (2015). "Are Gibbs-type priors the most natural generalization of the Dirichlet process?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37: 212–229. 5

Durante, D. (2017). "A note on the multiplicative gamma process." *Statistics & Probability Letters*, 122: 198–204. MR3584158. doi: https://doi.org/10.1016/j.spl.2016.11.014. 7

Fokoue, E. (2000). "A Markov chain Monte Carlo (MCMC) approach to the Bayesian analysis of mixtures of factor analysers." In *Proceedings in Computational Statistics 2000, Short Communication and Posters*, 19–30. Statistics The Netherlands. 1

Fokoue, E. and Titterington, D. (2003). "Mixtures of factor analysers. Bayesian estimation and inference by stochastic simulation." *Machine Learning*, 50(1): 73–94. MR2170402. doi: https://doi.org/10.1016/j.jmva.2004.08.004. 2

Forina, M., Armanino, C., Lanteri, S., and Tiscornia, E. (1983). "Classification of olive oils from their fatty acid composition." *Food Research and Data Analysis*, 189–214. 23

Frühwirth-Schnatter, S. (2011). *Dealing with Label Switching under Model Uncertainty*, chapter 10, 213–239. John Wiley & Sons, Ltd. MR2883354. doi: https://doi.org/10.1002/9781119995678.ch10. 17

Frühwirth-Schnatter, S. (2023). "Generalized Cumulative Shrinkage Process Priors with Applications to Sparse Bayesian Factor Analysis." *Philosophical Transactions of the Royal Society A*, (381): 381:20220148. MR4590506. 3, 8, 9, 10, 11, 12, 15, 26

Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. F. (2023). "When It Counts – Econometric Identification of the Basic Factor Model Based on GLT Structures." *Econometrics*, 11(4): 26. 22

Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. F. (2025). "Sparse Bayesian Factor Analysis When the Number of Factors Is Unknown (with Discussion)." *Bayesian Analysis*, 20(1): 213–344. MR4888145. doi: https://doi.org/10.1214/24-ba1423. 9, 12

Frühwirth-Schnatter, S. and Malsiner-Walli, G. (2019). "From here to infinity: Sparse finite versus Dirichlet process mixtures in model-based clustering." *Advances in Data Analysis and Classification*, 13: 33–64. MR3935190. doi: https://doi.org/10.1007/s11634-018-0329-y. 5

Frühwirth-Schnatter, S., Malsiner-Walli, G., and Grün, B. (2021). "Generalised Mixtures of Finite Mixtures and Telescoping Sampling." *Bayesian Analysis*, 16(4): 1279–1307. MR4381135. doi: https://doi.org/10.1214/21-BA1294. 3, 6, 10, 11, 14

Ghahramani, Z. and Beal, M. (2000). "Variational inference for Bayesian mixture of factor analyzers." In *Advanced in Neural Information Processing system*, volume 12, 449–455. 1

Ghahramani, Z., Griffiths, T. L., and Sollich, P. (2007). "Bayesian nonparametric latent feature models (with discussion and rejoinder)." In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics 8*, 1–25. Oxford: Oxford University Press. MR2433194. 8, 12

Ghahramani, Z. and Hinton, G. (1996). "The EM algorithm for mixtures of factor analyzers." Technical report, Department of Computer Science, University of Toronto. 1

Gnedin, A. (2010). "A Species Sampling Model with Finitely Many Types." *Electronic Communications in Probability*, 15: 79–88. MR2606505. doi: https://doi.org/10.1214/ECP.v15-1532. 5

Gnedin, A. and Pitman, J. (2006). "Exchangeable Gibbs partitions and Stirling triangles." *Journal of Mathematical Scienes*, 138: 5674–5684. MR2160320. doi: https://doi.org/10.1007/s10958-006-0335-z. 6

Greve, J., Grün, B., Malsiner-Walli, G., and Frühwirth-Schnatter, S. (2022). "Spying on the prior of the number of data clusters and the partition distribution in Bayesian cluster analysis." *Australian & New Zealand Journal of Statistics*, 64: 205–229. MR4467058. doi: https://doi.org/10.1111/anzs.12350. 6, 11

Grün, B., Malsiner-Walli, G., and Frühwirth-Schnatter, S. (2022). "How many data clusters are in the Galaxy data set? Bayesian cluster analysis in action." *Advances in Data Analysis and Classification*, 16(2): 325–249. MR4440853. doi: https://doi.org/10.1007/s11634-021-00461-8. 11

Grushanina, M. and Frühwirth-Schnatter, S. (2025). "Supplement to "Dynamic Mixture of Finite Mixtures of Factor Analysers"." *Bayesian Analysis*. doi: https://doi.org/10.1214/25-BA1543SUPP. 4

Hubert, L. and Arabie, P. (1985). "Comparing partitions." *Journal of Classification*, 2(1): 193–218. 18

Kowal, D. R. and Canale, A. (2023). "Semiparametric Functional Factor Models with Bayesian Rank Selection." *Bayesian Analysis*, 18: 1161–1189. MR4675036. doi: https://doi.org/10.1214/23-ba1410. 8, 13

Legramanti, S., Durante, D., and Dunson, D. (2020). "Bayesian cumulative shrinkage for infinite factorizations." *Biometrika*, 107(3): 745–752. MR4138988. doi: https://doi.org/10.1093/biomet/asaa008. 3, 7, 8, 9, 12, 13

McLachlan, G., Peel, D., and Bean, R. (2003). "Modelling high-dimensional data by mixtures of factor analyzers." *Computational Statistics & Data Analysis*, 41: 379–388. MR1973720. doi: https://doi.org/10.1016/S0167-9473(02)00183-4. 1

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics. New York: Wiley. MR1789474. doi: https://doi.org/10.1002/0471721182. 1

McNicholas, P., ElSherbiny, A., McDaid, A., and Murphy, T. (2018). "pgmm: parsimonious Gaussian mixture models. R package version 1.2.3." 23

McNicholas, P. and Murphy, T. (2008). "Parsimonious Gaussian mixture models." *Statistics and Computing*, 18(3): 285–296. MR2413385. doi: https://doi.org/10.1007/s11222-008-9056-0. 2

McNicholas, P. D. and Murphy, T. B. (2010). "Model-based clustering of microarray expression data via latent Gaussian mixture models." *Bioinformatics*, 26(21): 2705–2712. MR2413385. doi: https://doi.org/10.1007/s11222-008-9056-0. 23

Miller, J. W. (2023). "Consistency of mixture models with a prior on the number of components." *Dependence Modeling*, 11: 20220150. MR4557090. doi: https://doi.org/10.1515/demo-2022-0150. 2

Miller, J. W. and Harrison, M. T. (2013). "A simple example of Dirichlet process mixture inconsistency for the number of components." In *Advances in Neural Information Processing Systems*, 199–206. 5

Miller, J. W. and Harrison, M. T. (2018). "Mixture models with a prior on the number of components." *Journal of the American Statistical Association*, 113: 340–356. MR3803469. doi: https://doi.org/10.1080/01621459.2016.1255636. 3

Murphy, K., Viroli, C., and Gormley, I. (2020). "Infinite Mixtures of Infinite Factor Analysers." *Bayesian Analysis*, 15(3): 937–963. MR4132655. doi: https://doi.org/10.1214/19-BA1179. 2, 3, 4, 5, 6, 7, 9, 18, 19, 20, 23

Papastamoulis, P. (2018). "Overfitting Bayesian mixtures of factor analyzers with an unknown number of components." *Computational Statistics and Data Analysis*, 124: 220–234. MR3787623. doi: https://doi.org/10.1016/j.csda.2018.03.007. 2, 5, 6

Papastamoulis, P. (2020). "Clustering multivariate data using factor analytic Bayesian mixtures with an unknown number of components." *Statistics and Computing*, (30): 485–506. MR4065215. doi: https://doi.org/10.1007/s11222-019-09891-z. 23

Poworoznek, E., Ferrari, F., and Dunson, D. (2021). "Efficiently resolving rotational ambiguity in Bayesian matrix sampling with matching." *arXiv:2107.13783*. 18

Ročková, V. and George, E. (2016). "Fast Bayesian factor analysis via automatic rotation to sparsity." *Journal of the American Statistical Association*, 111(516): 1608–1622. MR3601721. doi: https://doi.org/10.1080/01621459.2015.1100620. 12

Schiavon, L. and Canale, A. (2020). "On the truncation criteria in infinite factor models." *Stat*, 9(1): e298. MR4156478. doi: https://doi.org/10.1002/sta4.298. 2, 7

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4: 639–650. MR1309433. 8

Teh, Y., Görür, D., and Ghahramani, Z. (2007). "Stick-breaking Construction for the Indian Buffet Process." In Meila, M. and Shen, X. (eds.), *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, 556–563. San Juan, Puerto Rico: PMLR. 8, 12

Viroli, C. (2010). "Dimensionally reduced model-based clustering through mixtures of factor mixture analyzers." *Journal of Classification*, 27(3): 363–388. MR2748989. doi: https://doi.org/10.1007/s00357-010-9063-7. 2