Bayesian Analysis (2020) **15**, Number 3, pp. 937–963

Infinite Mixtures of Infinite Factor Analysers

Keefe Murphy*, Cinzia Viroli†, and Isobel Claire Gormley‡,§

Abstract. Factor-analytic Gaussian mixtures are often employed as a model-based approach to clustering high-dimensional data. Typically, the numbers of clusters and latent factors must be fixed in advance of model fitting. The pair which optimises some model selection criterion is then chosen. For computational reasons, having the number of factors differ across clusters is rarely considered.

Here the infinite mixture of infinite factor analysers (IMIFA) model is introduced. IMIFA employs a Pitman-Yor process prior to facilitate automatic inference of the number of clusters using the stick-breaking construction and a slice sampler. Automatic inference of the cluster-specific numbers of factors is achieved using multiplicative gamma process shrinkage priors and an adaptive Gibbs sampler. IMIFA is presented as the flagship of a family of factor-analytic mixtures.

Applications to benchmark data, metabolomic spectral data, and a handwritten digit example illustrate the IMIFA model's advantageous features. These include obviating the need for model selection criteria, reducing the computational burden associated with the search of the model space, improving clustering performance by allowing cluster-specific numbers of factors, and uncertainty quantification.

Keywords: model-based clustering, factor analysis, Pitman-Yor process, multiplicative gamma process, adaptive Markov chain Monte Carlo.

1 Introduction

In cases where the number of variables p is comparable to or greater than the number of observations N, many clustering techniques tend to perform poorly or be intractable. Factor analysis (FA; Knott and Bartholomew, 1999) is a well-known approach to parsimoniously modelling data. Bai and Li (2012) outline some computational difficulties which arise when $N \ll p$. Model-based clustering methods which rely on latent factor models have long been successfully utilised to cluster high-dimensional data. Ghahramani and Hinton (1996) propose a mixture of factor analysers model (MFA) with cluster-specific parsimonious covariance matrices and estimate it via an expectation-maximisation algorithm; McLachlan and Peel (2000) provide a succinct overview. Estimation of MFA models has also been considered in a Bayesian framework (Diebolt and Robert, 1994; Richardson and Green, 1997). McNicholas and Murphy (2008) develop a suite of similar parsimonious Gaussian mixture models. Other related developments in this area include Baek et al. (2010) and Viroli (2010), among others.

*School of Mathematics and Statistics, University College Dublin, Ireland, keefe.murphy@ucd.ie
†Department of Statistical Sciences, University of Bologna, Italy, unibo.it/sitoweb/cinzia.viroli/en
‡School of Mathematics and Statistics, University College Dublin, Ireland, maths.ucd.ie/~cgormley
§Corresponding author: claire.gormley@ucd.ie

 https://doi.org/10.1214/19-BA1179

Clustering using a MFA model typically requires specifying the number of clusters and factors in advance of model fitting. Generally, a range of MFA models with different numbers of clusters and factors are fitted and then compared through the use of information criteria, such as the Bayesian Information Criterion (BIC; Kass and Raftery, 1995) or the Deviance Information Criterion (Spiegelhalter et al., 2002, 2014). Within a Bayesian framework Fokoué and Titterington (2003) use a stochastic model selection approach but do not simultaneously choose the optimal number of clusters and factors. Conducting an exhaustive search of the model space is computationally expensive; the cost is typically reduced by only considering models in which the number of factors is common across clusters. Regardless, even searching the reduced model space can be computationally onerous. The problem of identifying the optimal model is exacerbated by the fraught task of choosing among the range of model selection tools available, which often suggest different optimal models. Moreover, enforcing a common number of factors across clusters may lead to poor clustering performance due to a lack of flexibility.

The infinite mixture of infinite factor analysers (IMIFA) model is introduced here. It theoretically allows infinitely many components and infinitely many factors within each component. The need to select a model selection criterion is obviated and quantification of the uncertainty in the optimal numbers of non-empty clusters and cluster-specific factors is facilitated. IMIFA relies on an infinite mixture model through the use of a nonparametric Pitman-Yor process (PYP) prior (Perman et al., 1992; Pitman and Yor, 1997), of which the well-known Dirichlet process (DP; Ferguson, 1973) is a special case. The infinite mixture model framework allows the number of clusters present to be automatically inferred; here the stick-breaking construction (Pitman, 1996) and an independent slice-efficient sampler (Kalli et al., 2011) are employed to facilitate this.

By allowing infinitely many factors within each cluster, IMIFA addresses the difficulty in choosing the optimal number of factors. This facilitates fitting factor-analytic models which are more flexible, in the sense that the number of factors may be cluster-specific, thereby potentially improving clustering performance. This is achieved by assuming multiplicative gamma process (MGP) shrinkage priors (Bhattacharya and Dunson, 2011; Durante, 2017) on the cluster-specific factor loading matrices, thus generalising the MGP prior to the mixture setting. Such a prior allows the degree of shrinkage of the factor loadings towards zero to increase as the factor number tends towards infinity. The number of factors with non-negligible loadings can be considered as the 'active' number of factors within each cluster. Following Bhattacharya and Dunson (2011), a computationally efficient adaptive Gibbs sampling algorithm is employed for estimation. Thus, the choice of the numbers of active factors in different clusters is automated.

The IMIFA model with its PYP-MGP prior thus offers a single-pass and therefore computationally efficient approach to clustering high-dimensional data. It can be viewed as the most flexible model at the head of a family of Bayesian factor-analytic mixture models. Section 2 develops the hierarchy of the IMIFA model family, beginning with the MFA model and concluding with the flagship IMIFA model. Between these extremes the novel finite mixture of infinite factor analysers model (MIFA) is introduced. Overfitted factor-analytic mixtures (Papastamoulis, 2018) also belong to the IMIFA family; the overfitted mixture of infinite factor analysers (OMIFA) model is also introduced here.

Section 3 considers implementation of the IMIFA family of models. A benchmarking experiment is conducted on the well-known Italian olive oil data set. A real data application follows through the cluster analysis of spectral metabolomic data from an epilepsy study. Finally an illustrative application is provided through clustering United States Postal Service handwritten digit data, a setting for which fitting sub-models of the IMIFA family is practically infeasible. Comparisons against other clustering methods are provided throughout. Simulation studies demonstrating the performance of IMIFA under different scenarios are deferred to the supplementary material (Murphy et al., 2019a). Section 4 concludes the article with a discussion of IMIFA and thoughts on future research directions.

A software implementation for IMIFA and its family of sub-models is provided by the associated R package `IMIFA` (Murphy et al., 2019b), which is freely available from `www.r-project.org` (R Core Team, 2019), with which all results were generated.

2 The IMIFA Model Family

The hierarchy of the IMIFA family of models is delineated herein, including a review of extant methodologies, the introduction of novel sub-models, and concluding with the flagship IMIFA model. Prior specifications, Markov chain Monte Carlo (MCMC) inferential procedures, approaches to posterior predictive model checking, and model-specific implementation issues that arise in practice are addressed.

2.1 Mixtures of Factor Analysers

Mixtures of factor analysers are Gaussian latent variable models used for clustering high-dimensional data. For each of G clusters in these finite mixtures, the cluster-specific FA model in cluster g is given by $\mathbf{x}_i - \boldsymbol{\mu}_g = \boldsymbol{\Lambda}_g\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_{ig}$. The observed feature vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$ is assumed to linearly depend on a q-vector ($q \ll p$) of latent common factor scores $\boldsymbol{\eta}_i$ and additional sources of variation called specific factors $\boldsymbol{\varepsilon}_{ig}$. It is assumed that $\boldsymbol{\eta}_i$ has a q-variate Gaussian distribution $\mathrm{N}_q(\mathbf{0}, \boldsymbol{\mathcal{I}}_q)$, where $\boldsymbol{\mathcal{I}}_q$ denotes the $q \times q$ identity matrix, and that $\boldsymbol{\varepsilon}_{ig} \sim \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Psi}_g)$, where $\boldsymbol{\Psi}_g$ is a diagonal matrix with non-zero elements $\psi_{1g}, \ldots, \psi_{pg}$ known as uniquenesses. Here, $\boldsymbol{\Lambda}_g$ denotes the $p \times q$ factor loadings matrix of cluster g and notably $q = 0$ is permitted.

To facilitate estimation, a latent cluster indicator vector $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})^\top$ is introduced such that $z_{ig} = 1$ if observation i belongs to cluster g and $z_{ig} = 0$ otherwise. Hence, $\mathbf{z}_i$ has a $\mathrm{Mult}(1, \boldsymbol{\pi})$ distribution where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)^\top$ are the cluster mixing proportions which sum to 1. A symmetric uniform Dirichlet prior $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha} = (\alpha, \ldots, \alpha) = \mathbf{1})$ is assumed. Upon marginalising out $\mathbf{z}_i$ and $\boldsymbol{\eta}_i$, MFA yields a parsimonious finite sum covariance structure for the observed data

$$f(\mathbf{x}_i \,|\, \boldsymbol{\theta}) = \sum_{g=1}^{G} \pi_g \mathrm{N}_p\big(\mathbf{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g\boldsymbol{\Lambda}_g^\top + \boldsymbol{\Psi}_g\big), \tag{1}$$

where $\mathrm{N}_p(\mathbf{x}_i;\cdot,\cdot)$ is the density of a p-variate Gaussian distribution evaluated at $\mathbf{x}_i$ and $\boldsymbol{\theta}_g = \{\boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g\}$ are the cluster-specific FA parameters for which inference is straightforward under a Gibbs sampling scheme. Imposing constraints on $\boldsymbol{\Psi}_g$ (McNicholas and Murphy, 2008) and/or fixing $\pi_g = 1/G \;\forall\; g$ may be useful in some settings.

Prior Specification and Practical Issues

The conditionally conjugate nature of the various prior distributions detailed below facilitates MCMC sampling via straightforward Gibbs updates. A multivariate Gaussian prior is assumed for the factor loadings of the variable j across the q factors of cluster g: $\boldsymbol{\Lambda}_{jg} = (\lambda_{j1g}, \ldots, \lambda_{jqg}) \sim \mathrm{N}_q(\mathbf{0}, \boldsymbol{\mathcal{I}}_q)$. Similarly, a diffuse multivariate Gaussian prior is assumed for the component means, $\boldsymbol{\mu}_g \sim \mathrm{N}_p\big(\widetilde{\boldsymbol{\mu}}, \varphi^{-1}\boldsymbol{\mathcal{I}}_p\big)$, where $\widetilde{\boldsymbol{\mu}}$ is the overall sample mean and the scalar φ controls the level of diffusion.

An inverse gamma prior $\psi_{jg} \sim \mathrm{IG}(\alpha, \beta_j)$ is assumed for the uniquenesses of variable j in cluster g. Guided by Frühwirth-Schnatter and Lopes (2010), hyperparameters are chosen to ensure ψ_{jg} is bounded away from 0, thereby avoiding Heywood problems. With sufficiently large shape α, variable-specific scales are derived from the sample precision matrix $\mathbf{S}^\star = \mathbf{S}^{-1}$ via $\beta_j = (\alpha - 1)/S^\star_{jj}$. However, when N/p is close to or less than 1, or when $\mathbf{S}^{-1}$ is otherwise unavailable, $\mathbf{S}^\star$ is replaced by a ridge-type estimator $\widehat{\mathbf{S}^{-1}} = \big(\beta_0 + N/2\big)\big(\beta_0\boldsymbol{\mathcal{I}}_p + 0.5\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^\top\big)^{-1}$, where β_0 is a hyperparameter. For unstandardised data, this estimator is constructed for the inverse correlation matrix and then appropriately scaled using the diagonal entries of $\mathbf{S}$ (Wang et al., 2015). When the variances are roughly balanced, constraining $\boldsymbol{\Psi}_g$ to $\psi_g\boldsymbol{\mathcal{I}}_p$, and/or using $\beta_j = \beta = (\alpha - 1)/\max(\mathrm{diag}(\mathbf{S}^\star))$, provides additional parsimony. Notably, the isotropic constraint provides the link between factor analysis and probabilistic principal component analysis (Tipping and Bishop, 1999).

The rotational invariance property which makes FA models non-identifiable is well known: most covariance matrices $\boldsymbol{\Sigma}$ cannot be uniquely factored as $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}$ when $q > 1$. Though identifiability of $\boldsymbol{\Lambda}$ is not strictly necessary for the purposes of clustering or inferring $\boldsymbol{\Sigma}$, addressing the identifiability problem offline using the parameter expanded approach of Ghosh and Dunson (2008) in tandem with Procrustean methods, as in McParland et al. (2014), yields interpretable posterior summaries. Another practical issue is the label switching phenomenon (Frühwirth-Schnatter, 2010) which is addressed offline using the cost-minimising permutation given by the square assignment algorithm (Carpaneto and Toth, 1980). Finally, optimal FA and MFA models are chosen using the BIC-MCMC criterion (Frühwirth-Schnatter, 2011) where necessary in what follows.

2.2 Mixtures of Infinite Factor Analysers

To overcome the requirement to specify q, infinite factor analysis (IFA) models are employed (Bhattacharya and Dunson, 2011). The IFA model is a factor analysis model which assumes a multiplicative gamma process (MGP) shrinkage prior on the loadings matrix. This prior allows the degree of shrinkage towards zero to increase as the column index $k \to \infty$, mitigating against the factor splitting phenomenon. Here the IFA model

is generalised to the mixture setting, leading to the novel mixture of infinite factor analysers (MIFA) model. Under MIFA, the MGP prior is placed on each parameter expanded $\mathbf{\Lambda}_g$ matrix with no restrictions on its entries, thereby making the induced prior on $\mathbf{\Sigma}_g$ invariant to the ordering of the variables. The MGP prior is conditionally conjugate, facilitating block Gibbs updates of the loadings and hence rapid mixing. Thus, the MGP prior in mixture settings is given by

$$\begin{aligned}
\lambda_{jkg} \,|\, \phi_{jkg}, \tau_{kg}, \sigma_g &\sim \mathrm{N}_1\left(0, \phi_{jkg}^{-1}\tau_{kg}^{-1}\sigma_g^{-1}\right), \qquad & \phi_{jkg} &\sim \mathrm{Ga}(\nu_1, \nu_2)\,, \\
\tau_{kg} &= \prod_{h=1}^{k} \delta_{hg}\,, & \sigma_g &\sim \mathrm{Ga}(\varrho_1, \varrho_2)\,, \\
\delta_{1g} &\sim \mathrm{Ga}(\alpha_1, \beta_1)\,, & \delta_{hg} &\sim \mathrm{Ga}(\alpha_2, \beta_2) \;\; \forall \; h \geq 2,
\end{aligned}$$

where τ_{kg} is a column shrinkage parameter for the k-th column in the g-th cluster's loadings matrix $\mathbf{\Lambda}_g \; \forall \; k = 1, \ldots, \infty$, and $\mathrm{Ga}(\alpha, \beta)$ denotes the gamma distribution with mean α/β. The role of the local shrinkage parameters $\phi_{1kg}, \ldots, \phi_{pkg}$ for the p elements in column k of $\mathbf{\Lambda}_g$ is to favour sparsity while also preserving the signal of non-zero loadings. Lastly, the cluster shrinkage parameter σ_g reflects the belief that the degree of shrinkage is cluster-specific. A schematic illustration of the MGP prior is given in Figure 1; note that loadings can shrink arbitrarily close, but not exactly, to zero.

Bhattacharya and Dunson (2011) fix $\beta_1 = \beta_2 = 1$ and recommend that $\alpha_2 > \beta_2$. However, Durante (2017) elaborates on the cumulative shrinkage properties and roles played by hyperparameters, showing in particular that $\alpha_2 > \beta_2 + 1$ is necessary in order to have variances that decrease in expectation with k. It is also recommended that α_2 be moderately large relative to α_1 and to avoid excessively high values for α_1. While Bhattacharya and Dunson (2011) assume $\mathrm{Ga}(\nu, \nu)$ priors for the local shrinkage parameters, here more general settings are used to allow control over prior non-informativity. In the spirit of Durante (2017), the expectation $\nu_2/(\nu_1 - 1)$ of the induced inverse gamma prior on ϕ_{jkg}^{-1} is suggested to be ≤ 1 to induce sparsity on average. It is generally advisable that MGP hyperparameters are chosen such that the first two moments of the associated hyperprior are defined. In the mixture setting, α_1 and α_2 may need to be higher than the values suggested by Durante (2017) to enforce a greater degree of

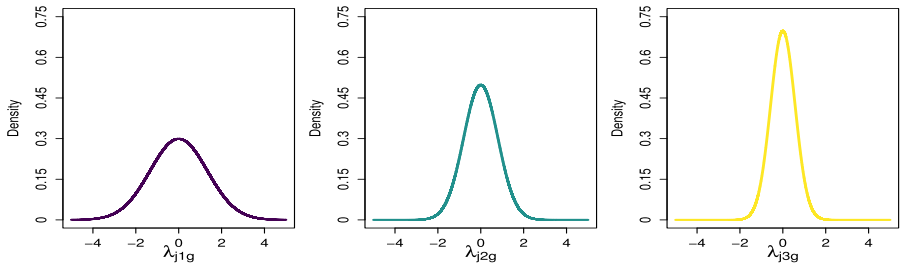

Figure 1: Density of a typical element in the first, second, and third columns of a cluster-specific loadings matrix under the MGP shrinkage prior.

shrinkage in clusters with few units; this aspect is highlighted in simulation studies in Appendix B.

The Adaptive Gibbs Sampler

An adaptive Gibbs sampler (AGS) is employed when performing inference for MIFA. This dynamically shrinks the loadings matrices (and the infinite scores matrix $\boldsymbol{\eta}$) to have finite numbers of columns, by selecting the number of 'active' factors. This practically facilitates posterior computation while closely approximating the IFA model, without requiring specification of $\mathbf{Q} = (q_1, \ldots, q_G)^\top$. However, a strategy is required for choosing appropriate truncation levels, $\widehat{q}_g$, that strike a balance between missing important factors and wasting computational effort. For computational reasons, a conservatively high upper bound is used, such that $q_g^\star = \min\big(\lfloor 3(p) \rceil, N-1, p-1\big)\ \forall\ g$. The number of factors in each $\boldsymbol{\Lambda}_g$ is then adaptively tuned as the MCMC chain progresses. Adaptation can be made to occur only after the burn-in period, in order to ensure the true posterior distribution is being sampled from before truncating the loadings matrices.

At the t-th iteration, adaptation occurs with probability $\mathrm{p}(t) = \exp(-b_0 - b_1 t)$, with b_0 and b_1 chosen so that adaptation occurs often at the beginning but then decreases exponentially fast in frequency. Here $b_0 = 0.1$ and $b_1 = 5 \times 10^{-5}$ are used. With probability $\mathrm{p}(t)$, loadings columns having some pre-specified proportion of elements ς in a small neighbourhood ϵ of zero are monitored. If there are no such columns, an additional column is added by simulation from the MGP prior. Otherwise redundant columns are discarded and the AGS proceeds with all parameters corresponding to non-redundant columns retained. Choice of ς and ϵ can be delicate: here $\varsigma = \lfloor 0.7 \times p \rfloor / p$ and $\epsilon = 0.1$ are found to strike an appropriate balance. The dimension of the matrix $\boldsymbol{\eta}$ of factor scores at a given iteration are set to $p \times \overline{q} = p \times \max(\mathbf{Q}(t))$; rows corresponding to observations currently assigned to a cluster with fewer latent factors than $\overline{q}$ are padded with zeros. Notably, here $\widehat{q}_g$ may shrink to 0 thus allowing diagonal covariance structure within a component. If this occurs, the decision to simulate a new column is based on a binary trial with probability $1 - \varsigma$ as there are no loadings columns to monitor.

The numbers of active factors in each cluster for each retained posterior sample can be used to construct a barchart approximation to the posterior distribution of q_g. The posterior mode is used to estimate each q_g, with credible intervals quantifying uncertainty. The main advantages of MIFA are that different clusters can be modelled by different numbers of factors and that the model search is reduced to one for G only, as q_g is estimated automatically during model fitting. Here, for MIFA models, the optimal G is chosen via the BICM (BIC-Monte (Carlo)) proposed by Raftery et al. (2007), with $\mathrm{BICM} = 2\ln\big(\widetilde{\mathcal{L}}\big) - 2s_l^2 \ln(N)$, where $\widetilde{\mathcal{L}}$ is the largest log-likelihood value calculated for each retained posterior sample and s_l^2 is the sample variance thereof. This criterion is particularly useful in the context of nonparametric models where the number of free parameters is difficult to quantify.

Other Infinite Factor Models

This work offers an extension of the MGP prior and its related AGS routine to the mixture modelling context. Wang et al. (2016) develop a related model employing a multiplicative exponential process prior. Other nonparametric approaches to inferring the number of factors include Knowles and Ghahramani (2007), in which a two-parameter Indian Buffet Process (IBP) prior is assumed on an infinite binary matrix underlying the factor scores, thus selecting features of interest, with associated standard Gaussian weights. A closely related approach using the Beta process (BP) is provided by Paisley and Carin (2009). In Knowles and Ghahramani (2011) and Ročková and George (2016), an IBP prior is instead assumed for sparsifying the loadings. These models assume a single sparse infinite factor model for the whole data set. However, embedding them in a mixture modelling setting, similar to the IMIFA framework, is intuitively feasible.

Indeed, Chen et al. (2010) employ the BP prior, coupled with a Dirichlet process prior, to perform clustering in a manifold learning setting. While the BP and IBP priors achieve exact sparsity, which may be advantageous in certain applications, the MGP prior has a weaker notion of sparsity by virtue of cumulatively shrinking an infinite series arbitrarily close to zero, thereby preserving small signals. The block updates of each row of $\boldsymbol{\Lambda}_g$ facilitated by the MGP prior and parameter expansion mean the AGS approach is a simpler, more computationally efficient alternative to the BP and IBP priors.

2.3 Overfitted Mixtures of (Infinite) Factor Analysers

While MIFA obviates the need to pre-specify $\mathbf{Q}$, the issue of model choice is not yet fully resolved. Overfitted mixtures (Rousseau and Mengersen, 2011; van Havre et al., 2015) are one means of extending MIFA; indeed Papastamoulis (2018) proposes an overfitted mixture of factor analysers (OMFA), albeit with finite factors. Here, the overfitted mixture of infinite factor analysers (OMIFA) model is introduced.

In overfitted mixtures the symmetric Dirichlet prior on $\boldsymbol{\pi}$ plays an important role. Estimation is approached by initially overfitting the number of clusters expected to be present. Small values of the hyperparameter $\boldsymbol{\alpha}$ encourage emptying out excess components in the posterior distribution; the uniform prior with $\boldsymbol{\alpha} = \mathbf{1}$ is rather indifferent in this respect. The sampler is initialised with a conservatively high number of components: $G^\star = \max\big(\lceil 3\ln(N)\rceil, 25, N-1\big)$, though this may be too high if it is close to N. While $\widetilde{G} = G^\star$ remains fixed throughout the MCMC chain, the number of non-empty clusters is recorded at each iteration of the sampler as $G_0 = \widetilde{G} - \sum_{g=1}^{\widetilde{G}} \mathbb{1}\big(\sum_{i=1}^{N} z_{ig} = 0\big)$ where $\mathbb{1}(\cdot)$ is the indicator function. The true G is estimated by $\widehat{G}$, the G_0 value visited most often. Cluster-specific inference is conducted only on samples corresponding to those visits. For the OMIFA model, the AGS is modified to handle empty components: the MGP-related parameters are simulated from the relevant priors and each corresponding $\boldsymbol{\Lambda}_g$ matrix is restricted to having $\overline{q}$ factors, i.e. the same number of columns currently in the matrix of factor scores $\boldsymbol{\eta}$, either by truncation or by padding with zeros, as required.

2.4 Infinite Mixtures of (Infinite) Factor Analysers

Embedding MFA and MIFA in an infinite mixture setting leads, respectively, to the infinite mixture of finite factor analysers model (IMFA) and the flagship infinite mixture of infinite factor analysers model (IMIFA). These models employ a nonparametric Pitman-Yor process (PYP) prior which is easily incorporated into the MCMC sampling scheme.

The PYP is a stochastic process whose draws are discrete probability measures: $H \sim \text{PYP}(\alpha, d, H_0)$ denotes a PYP probability distribution H, with base distribution H_0 interpreted as the mean of the PYP, discount parameter $d \in [0, 1)$, and concentration parameter $\alpha > -d$. For the PYP mixture model IMFA and the PYP-MGP mixture model IMIFA H_0 comes from the factor-analytic mixture (1), hence

$$f(\mathbf{x}_i \,|\, \boldsymbol{\theta}) = \sum_{g=1}^{\infty} \pi_g \text{N}_p\big(\mathbf{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g^\top + \boldsymbol{\Psi}_g\big). \tag{2}$$

The stick-breaking representation of the PYP (Pitman, 1996) is used as a prior process for generating the mixing proportions in (2). This construction views $\{\pi_1, \pi_2, \ldots\}$ as pieces of a unit-length stick that is sequentially broken in an infinite process, with stick-breaking proportions $\boldsymbol{\Upsilon} = \{\upsilon_1, \upsilon_2, \ldots\}$, summarised as

$$\begin{aligned} \upsilon_g &\sim \text{Beta}(1 - d, \alpha + gd)\,, \qquad & \boldsymbol{\theta}_g &\sim H_0\,, \\ \pi_g &= \upsilon_g \prod_{l=1}^{g-1} (1 - \upsilon_l)\,, & H &= \sum_{g=1}^{\infty} \pi_g \delta_{\boldsymbol{\theta}_g} \sim \text{PYP}(\alpha, d, H_0)\,, \end{aligned}$$

where $\delta_{\boldsymbol{\theta}_g}$ is the Dirac delta centred at $\boldsymbol{\theta}_g$, such that draws are composed of a sum of infinitely many point masses. The PYP reduces to the DP when $d = 0$, in which case mass shifts to the right with increasing dispersion as α increases, implying an *a priori* larger number of components. However, some important distributional features fundamentally differ when $d \neq 0$ (De Blasi et al., 2015). The PYP exhibits heavier tail behaviour and allows the stick-breaking distribution to vary according to the component index g, without sacrificing much in the way of tractability. In particular, increasing d values have the effect of flattening the prior, controlling its degree of non-informativity.

Slice sampling (Walker, 2007; Kalli et al., 2011) is used here to yield samples from the PYP by adaptively truncating the number of components needed to be sampled at each iteration. By introducing an auxiliary variable $u_i > 0$ which preserves the marginal distribution of the data, and denoting by $\boldsymbol{\xi} = \{\xi_1, \xi_2, \ldots\}$ a positive sequence of infinite quantities which sum to 1, the joint density of $(\mathbf{x}, \mathbf{u})$ is given by $f(\mathbf{x}, \mathbf{u} \,|\, \boldsymbol{\theta}, \boldsymbol{\xi}) = \sum_{g=1}^{\infty} \pi_g \text{Unif}(\mathbf{u}; 0, \xi_g)\, f(\mathbf{x} \,|\, \boldsymbol{\theta}_g)$. Since only a finite number of ξ_g are greater than $\mathbf{u}$, the conditional density of $\mathbf{x} \,|\, \mathbf{u}$ can be written as a finite mixture with $\widetilde{G} = \max_{1 \leq i \leq N} |\boldsymbol{\mathcal{A}}_{\boldsymbol{\xi}}(u_i)|$ 'active' components at each iteration, where $|\cdot|$ denotes cardinality and $\boldsymbol{\mathcal{A}}_{\boldsymbol{\xi}}(\mathbf{u}) = \{g \colon \mathbf{u} < \xi_g\}$. Though G is infinite in theory, $\widetilde{G}$ can be at most equal to N. Thus, the infinite mixture of (infinite) factor analysers models can be sampled from. Typical implementations of the slice sampler arise when $\xi_g = \pi_g$ (Walker, 2007) but independent slice-efficient sampling (Kalli et al., 2011) allows for a deterministic

decreasing sequence, e.g. geometric decay, given by $\xi_g = (1-\rho)\,\rho^{g-1}$ where $\rho \in [0,1)$ is a fixed value to be chosen with care. Higher values generally lead to better mixing but longer run-times, as the average cardinality of $\mathcal{A}_{\boldsymbol{\xi}}(\mathbf{u})$ increases, and *vice versa*. Setting $\rho = 0.75$ appears to strike an appropriate balance in the applications considered here.

Inference for Infinite Mixtures of Factor Analysers Models

For clarity, what follows focuses on the IMIFA model where inference proceeds via the independent slice-efficient sampler with geometric decay. Inference for other models in the IMIFA family is closely related. The joint density of the IMIFA model is

$$\begin{aligned} f(\mathbf{X}, \boldsymbol{\eta}, \mathbf{Z}, \mathbf{u}, \boldsymbol{\Upsilon}, \boldsymbol{\theta}) &\propto f(\mathbf{X} \,|\, \boldsymbol{\eta}, \mathbf{Z}, \mathbf{u}, \boldsymbol{\Upsilon}, \boldsymbol{\theta})\, f(\boldsymbol{\eta})\, f(\mathbf{Z}, \mathbf{u} \,|\, \boldsymbol{\Upsilon}, \boldsymbol{\pi})\, f(\boldsymbol{\Upsilon} \,|\, \alpha, d)\, f(\boldsymbol{\theta}) \\ &= \left\{ \prod_{i=1}^{N} \prod_{g \in \mathcal{A}_{\boldsymbol{\xi}}(u_i)} \mathrm{N}_p\big(\mathbf{x}_i; \boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g \boldsymbol{\eta}_i, \boldsymbol{\Psi}_g\big)^{z_{ig}} \right\} \left\{ \prod_{i=1}^{N} \mathrm{N}_q\big(\boldsymbol{\eta}_i; \mathbf{0}, \boldsymbol{\mathcal{I}}_q\big) \right\} \\ &\quad \left\{ \prod_{i=1}^{N} \prod_{g=1}^{\infty} \left(\frac{\pi_g}{\xi_g} \mathbb{1}\big(u_i < \xi_g\big) \right)^{z_{ig}} \right\} \left\{ \prod_{g=1}^{\infty} \frac{(1-\upsilon_g)^{\alpha + gd - 1}}{\upsilon_g^d\, \mathrm{B}(1-d, \alpha + gd)} \right\} f(\boldsymbol{\theta})\,, \end{aligned}$$

where $\mathrm{B}(\cdot)$ is the Beta function and $f(\boldsymbol{\theta})$ is the product of the previously defined collection of conditionally conjugate priors with additional layers for hyperparameters. Only the parameters of the $\widetilde{G}$ active components are sampled at each iteration. The algorithm is initialised with the same $G^\star$ value detailed in Section 2.3, typically above the anticipated number to which the algorithm will converge, in the spirit of Hastie et al. (2014). Here, however, $\widetilde{G}$ can theoretically exceed this value. For computational reasons, a finite upper limit is placed on $\widetilde{G}$ with $\max(G^\star, \min(N-1, 50))$ found to be sufficiently large. However, $\widetilde{G}$ is only regarded as a set of proposals as to where to allocate observations; as in Section 2.3, it is the subset of non-empty clusters G_0 that is of inferential interest.

Bayesian approaches to clustering are known to be sensitive to initial cluster allocations. While starting values for $\mathbf{z}_i$ can be obtained by any means, model-based agglomerative hierarchical clustering (Scrucca et al., 2016) is used here. Though this is fast and intuitive given that IMIFA models are initialised at a conservatively high number of components, which are then merged as the sampler proceeds, heavily imbalanced initial cluster sizes are cautioned against. By extension, initial cluster means and mixing proportions are computed empirically. Other parameter starting values are simulated from their relevant prior distributions. The adaptive inferential algorithm for IMIFA then proceeds mostly via Gibbs updates (see Appendix A). For those which are multivariate Gaussian, using the Cholesky factor of the covariance matrices and employing block updates speeds up the algorithm (Rue and Held, 2005). The allocations $\mathbf{z}_i$ are sampled in a fast, numerically stable fashion (see Appendix A), using the Gumbel-Max trick (Yellott, 1977). Finally, state spaces for applications of IMIFA to real data can be highly multimodal with well-separated regions of high posterior probability coexisting, corresponding to clusterings with different numbers of components. Thus, label switching moves (Papaspiliopoulos and Roberts, 2008) are incorporated in order to improve mixing.

Assessing Model Fit and Mixing

As is good statistical practice, posterior predictive model checking (Gelman et al., 2004) is employed. Sampled model parameters from the MCMC chain are used to generate replicate data from the posterior predictive distribution. Valid posterior samples, after conditioning on $\widehat{G}$, are those for which $\max\{\mathbf{Q}(t)\} \geq \max\{\widehat{q}_1, \ldots, \widehat{q}_{\widehat{G}}\}$ such that the dimension of the estimated scores matrix $\widehat{\boldsymbol{\eta}}$ is preserved. To assess model fit, histograms of the modelled data $\mathbf{X}$ are compared to histograms of the replicate data in a global sense using the Posterior Predictive Reconstruction Error (PPRE), calculated as follows:

1. Gather the histogram bin counts of each variable in $\mathbf{X}$ into the $h \times p$ matrix $\boldsymbol{\mathcal{H}}$, where h is the maximum number of bins across all variables and $\boldsymbol{\mathcal{H}}$ is padded with zeros as required.
2. Generate $r \in \{1, \ldots, R\}$ data sets $\boldsymbol{\mathcal{X}}^{(r)}$ from the posterior predictive distribution.
3. Create a similar matrix of histogram bin counts $\boldsymbol{\mathcal{H}}^{(r)}$ for each $\boldsymbol{\mathcal{X}}^{(r)}$ using the same break-points with which $\boldsymbol{\mathcal{H}}$ was constructed (with endpoint bins extended to $\pm\infty$).
4. Compute the Frobenius norm $\|\cdot\|_{\mathcal{F}}$ between $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{\mathcal{H}}^{(r)}$, standardising to the 0-1 scale using the triangle inequality $\left|\|\boldsymbol{\mathcal{H}}\|_{\mathcal{F}} - \|\boldsymbol{\mathcal{H}}^{(r)}\|_{\mathcal{F}}\right| \leq \|\boldsymbol{\mathcal{H}} - \boldsymbol{\mathcal{H}}^{(r)}\|_{\mathcal{F}} \leq \|\boldsymbol{\mathcal{H}}\|_{\mathcal{F}} + \|\boldsymbol{\mathcal{H}}^{(r)}\|_{\mathcal{F}}$.

The distribution of PPRE values can be visualised using boxplots and summarised by the median, with credible intervals. This discrepancy measure is well-suited to assessing model adequacy for mixtures of multivariate data: it accounts for inherent multimodality and gives a global quantitative measure of agreement between the distributions of the observed variables and their posterior predictive counterparts.

Convergence of the MCMC chains is assessed using the potential scale reduction factor (PSRF; Brooks and Gelman, 1998; Plummer et al., 2006). Random allocations of the initial cluster labels, resulting in different draws from the relevant priors for parameter initialisation, are used to construct the multiple overdispersed chains required. The MAP labels of each chain are matched to the main chain prior to computing the diagnostics; $\boldsymbol{\Lambda}_g$ matrices are also rotated to a common template for each cluster. Good convergence is indicated by upper PSRF 95% confidence interval limits close to 1; this is a stricter requirement than the PSRF values themselves being near 1.

Comparing the IMIFA Family Models

Though IMIFA and OMIFA come with the computational complexities inherent in non-parametric methods, diminishing adaptation, and extra tuning parameters, their advantages over other models in the IMIFA family are numerous: i) flexibility, in the sense that models where $q_g \neq q'_g$ can be fitted, ii) computational efficiency, in the sense that the burden is reduced relative to searching over a range of fitted MFA or MIFA models, iii) removing the need for model selection criteria, and iv) the ability to quantify the uncertainty in $\widehat{G}$ and $\widehat{q}_g$. Both methods offer simpler alternatives to reversible jump MCMC (Richardson and Green, 1997) and birth-death MCMC (Stephens, 2000). Hence, among the IMIFA family, the infinite factor models are recommended over the finite

factor models and the infinite and overfitted mixtures are recommended over the finite mixtures. However, the MIFA model is appropriate if one wishes to fix G but infer q_g.

While infinite mixtures are often used for density estimation, they are also employed to infer the number of components in cluster analyses (e.g. Kim et al. 2006; Xing et al. 2006; Yerebakan et al. 2014). However, Miller and Harrison (2013, 2014) raise concerns about the guarantee of posterior consistency for the number of non-empty clusters, showing the number uncovered is typically greater than or equal to the truth, often with several vanishingly small clusters inferred. These concerns highlight the need for practitioners to pay due consideration to the uncertainty in the number of clusters offered by IMIFA models. Relatedly, Frühwirth-Schnatter and Malsiner-Walli (2019) compare infinite mixtures to overfitted ('sparse finite') mixtures. They highlight that overfitted mixtures are useful for applications in which the data arise from a moderate number of clusters, even as the sample size increases, whereas infinite mixtures are suited to cases where the number of clusters also increases. However, they show that clustering results are driven less by the assumption of whether the data arose from a finite or infinite mixture, but by the hyperprior on the DP parameters or the sparseness of the Dirichlet prior in the overfitted setting. Indeed, they show that overfitted and infinite mixtures yield comparable clustering performance on the observed data when these hyperpriors are matched. This matching leads to 'sparse' infinite mixtures that avoid overfitting the number of clusters. Similar behaviour is observed in the applications in Section 3, where the IMIFA and OMIFA models, with matched hyperpriors, give comparable results.

The issue of choosing α can make implementing overfitted models challenging. With fixed $\alpha = \gamma/G^\star$, the prior approximates a DP with concentration parameter γ as $G^\star$ tends to infinity (Green and Richardson, 2001). Here, following Frühwirth-Schnatter and Malsiner-Walli (2019), a $\text{Ga}(a, bG^\star)$ hyperprior is assumed for α. This favours small values and allows α to be updated via Metropolis-Hastings. In the infinite mixture setting, learning the PYP parameters (which also requires Metropolis-Hastings steps) and adopting the label-switching moves enables accurate inference on G_0. A joint hyperprior $\text{p}(\alpha, d) = \text{p}(\alpha \,|\, d)\,\text{p}(d)$ is assumed (Carmona et al., 2019) where $\text{p}(\alpha \,|\, d) = \text{Ga}(\alpha + d; a, b)$; choosing a large b encourages clustering (Müller and Mitra, 2013). A spike-and-slab hyperprior $d \sim \kappa\delta_0 + (1-\kappa)\,\text{Beta}(a', b')$ is assumed. The estimated proportion $\widehat{\kappa}$ can then be used to assess whether the data arose from a DP or a PYP at little extra computational cost. See Appendix A for further details.

3 Illustrative Applications

The flexibility and performance of the IMIFA model and its related model family are demonstrated below through application to benchmark and real data sets. All results are obtained through the `IMIFA` R package; code to reproduce many of the results is available in the associated vignette[1]. Appendix B reports on simulation studies demonstrating the performance of IMIFA under different scenarios, including effects of the N/p ratio, the PYP parameters, imbalanced cluster sizes, uncommon q_g and the degree of loadings sparsity, while Appendix C explores the robustness of IMIFA.

[1] https://cran.r-project.org/web/packages/IMIFA/vignettes/IMIFA.html

Parameter(s)	Hyperparameter(s)	Value(s)
$\boldsymbol{\mu}_g$	φ	0.01
$\boldsymbol{\Psi}_g$	(α, β_0)	$(2.5, 3)$
ϕ_{jkg}	(ν_1, ν_2)	$(3, 2)$
δ_{1g}	(α_1, β_1)	$(2.1, 1)$
δ_{kg}	(α_2, β_2)	$(3.1, 1)$
σ_g	(ϱ_1, ϱ_2)	$(3, 2)$
α	(a, b)	$(2, 4)$
d	(a', b', κ)	$(1, 1, 0.5)$

Table 1: Hyperparameter specifications for the IMIFA model. Note that the specification of the beta distribution in the prior for d amounts to a standard uniform.

MCMC chains were run for 50,000 iterations, except for Section 3.3 in which 20,000 were run. Every 2$^{\text{nd}}$ sample was thinned and the first 20% of iterations were discarded as burn-in. All computations were performed on a Dell Latitude 5491 laptop, equipped with a 2.60 GHz Intel Core i7-8850H processor and 16 GB of RAM. Where necessary, the optimal finite and infinite factor models are chosen by the BIC-MCMC and BICM criteria, respectively. Throughout, $\hat{\cdot}$ denotes the posterior mode, posterior mean, or relevant optimal value. Unless otherwise stated data were mean-centred and unit-scaled and no constraints were imposed on the uniquenesses. Hyperprior specifications are detailed in Table 1. While there are many hyperparameters to select, the choices are all reasonably standard. However, poor settings may introduce additional factors or clusters to maintain flexibility and so care in specifying hyperparameters is advised.

3.1 Benchmark Data: Italian Olive Oils

The Italian olive oil data (Forina et al., 1983) is often clustered using factor-analytic models, e.g. McNicholas (2010). The data detail the percentage composition of 8 fatty acids in 572 Italian olive oils, known to originate from three areas: southern and northern Italy and Sardinia. Each area is composed of different regions: southern Italy comprises north Apulia, Calabria, south Apulia, and Sicily; Sardinia is divided into inland and coastal Sardinia; and northern Italy comprises Umbria and east and west Liguria. Hence, the true number of clusters is hypothesised to correspond to either 3 areas or 9 regions.

The full family of IMIFA models is fitted to the olive oil data with results detailed in Table 2. Models relying on pre-specification of finite ranges of G and/or q are based on $G = 1, \ldots, 9$ and $q = 0, \ldots, 6$. Clustering performance is evaluated using the adjusted Rand index (ARI; Hubert and Arabie, 1985) and the misclassification rate, compared to the 3 area labels. The α parameter is reported as its fixed value or posterior mean, as appropriate. Table 2 shows the flexibility and accuracy of the developed model family, and of the IMIFA model in particular which has the best clustering performance. Additionally, IMIFA is the most computationally efficient model considered, among those in the IMIFA family achieving clustering, as it requires only one run. This speed improvement would be exacerbated with larger data sets. However, methods requiring fitting of multiple models were run here in series; parallel implementations would reduce run-

Model	# Models	Relative Time	α	d	G	**Q**	ARI	Error (%)
IMIFA	1	1.00	0.48	0.01	4	6, 3, 6, 2	0.94	8.39
IMFA	7	4.14	0.62	0.01	5	6, 6, 6, 6, 6	0.91	14.86
OMIFA	1	1.19	0.02	–	4	6, 3, 6, 4	0.93	9.97
OMFA	7	5.11	0.03	–	5	6, 6, 6, 6, 6	0.85	15.56
MIFA	9	3.41	1	–	5	6, 3, 6, 6, 4	0.92	10.31
MFA	63	13.86	1	–	2	5, 5	0.82	17.13
IFA	1	0.11	–	–	1	6	–	–
FA	7	0.37	–	–	1	6	–	–
`mclust`	115	0.01	–	–	6	–	0.56	38.64
MFMA	1,350	4.68	–	–	4	5, 5, 5, 5	0.68	20.28
`pgmm`	588	4.46	–	–	5	6, 6, 6, 6, 6	0.53	35.84

Table 2: Results of fitting a range of models, including the full IMIFA family, to the Italian olive oil data, detailing the number of candidate models explored, the run-time relative to the IMIFA run, the posterior mean or fixed value of α, the posterior mean of d, modal estimates of G and $\mathbf{Q}$, and the ARI and misclassification rate as evaluated against the known area labels, under the optimal or modal model as appropriate.

times. Finally, models with different numbers of cluster-specific factors show improved clustering performance compared to the corresponding finite factor model in every case.

The IMIFA model's performance also compares favourably to the best parsimonious Gaussian mixture model, fit via the `pgmm` R package (McNicholas et al., 2018) and the best mixture of factor mixture analysers (MFMA) model (Viroli, 2010), evaluated with $1, \ldots, 5$ components in both layers. Models with zero factors were not considered in either case. IMIFA also outperforms the best constrained Gaussian mixture model fitted using `mclust` (Scrucca et al., 2016). These finite mixtures are fit via maximum likelihood and use the BIC for model selection after fitting a large number of candidate models.

It is also notable that within the set of IMIFA models relying on information criteria, those deemed optimal were not necessarily optimal in a clustering sense. For instance, the 4-cluster MIFA model yields an ARI of 0.94 and a misclassification rate of 6.99%, with respect to the 3 area labels, despite its sub-optimal BICM. Similarly, the BICM and BIC-MCMC criteria suggest different optimal MFA models. For the IMIFA model $\widehat{\kappa} \approx 0.89$, suggesting similar inference would have resulted under a DP prior. Indeed, the results obtained by the OMIFA and OMFA models are similar to those of their infinite mixture counterparts, though the latter provide a better fit to the data (see Figure 5).

Figure 2 shows a barchart approximation to the posterior distribution of G under the IMIFA model. The modal value of 4, visited in $\approx 90\%$ of posterior samples, is used as the estimate of the true number of clusters (with 95% credible interval $[4, 5]$). Table 3a tabulates the MAP clustering against the 3 area labels and suggests this solution makes geographic sense, in that northern oils are cleanly split into two sub-clusters. Cluster 1 contains all of the 323 southern Italy oils: this large cluster requires the largest number of factors ($\widehat{q}_1 = 6$ $[5, 6]$, with 95% credible intervals in brackets). Some of the other clusters require notably fewer ($\widehat{q}_2 = 3$ $[1, 6]$, $\widehat{q}_3 = 6$ $[3, 6]$, and $\widehat{q}_4 = 2$ $[1, 4]$). Table 3b gives the

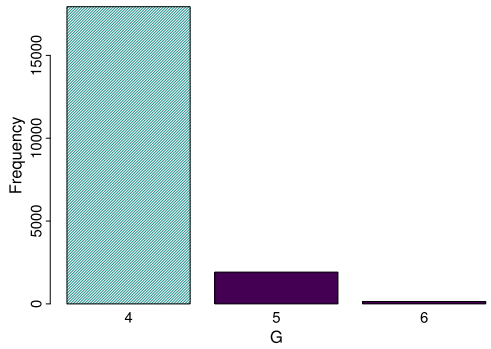

Figure 2: Posterior distribution of G under the IMIFA model for the olive oil data. The number of clusters is estimated by the modal value, $\widehat{G} = 4$.

(a) 3 area cross tabulation

	1	2	3	4
Southern Italy	323	0	0	0
Sardinia	0	98	0	0
Northern Italy	0	0	103	48

(b) 4 area cross tabulation

	1	2	3	4
Southern Italy	323	0	0	0
Sardinia	0	98	0	0
East Liguria & Umbria	0	0	100	0
West Liguria	0	0	3	48

Table 3: Confusion matrices of the MAP IMIFA clustering of the Italian olive oils against (a) the known 3 area labels and (b) the new labelling in which northern Italy is split into its constituent sub-regions.

confusion matrix with oils from the north labelled by their associated region(s), yielding an ARI of 0.994 and a misclassification rate of 0.52%. Figure 3 shows the uncertainty in the allocations to these clusters. Only three oils have large probability of belonging to a cluster other than the one to which they were assigned by the IMIFA model.

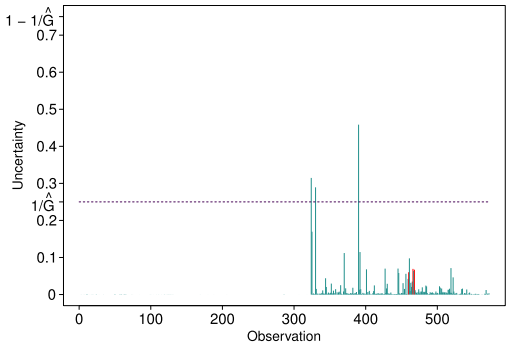

Figure 3: Clustering uncertainties for the IMIFA model for the olive oil data. Oils misclassified according to the labels in Table 3b are highlighted in red.

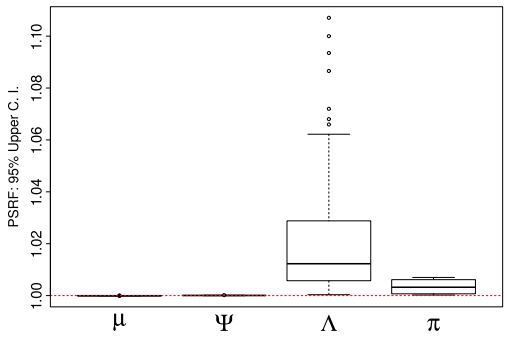

Figure 4: Boxplots of the upper PSRF limits for all cluster means, uniquenesses, loadings, and mixing proportions in the overdispersed IMIFA chains fit to the olive oil data, with red reference line at 1.

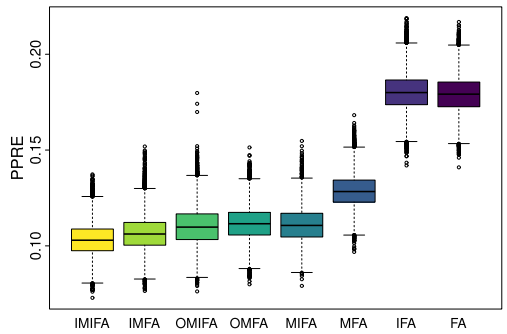

Figure 5: Boxplots of the PPRE values for the full family of IMIFA models fit to the olive oil data. Values close to zero indicate good model fit.

To assess sensitivity to starting values, the IMIFA model was re-fitted using multiple random initial allocations, implying also different random draws from the priors for parameter starting values. These runs led to identical inference about $\widehat{G}$ and $\widehat{\mathbf{Q}}$ and equivalent clustering performance. These overdispersed chains were used to compute the upper 95% PSRF confidence limits depicted in Figure 4, which indicate good convergence. The PPRE boxplots in Figure 5 demonstrate the superior fit of the IMIFA model (with a median PPRE of 0.10) to the olive oil data, compared to the other IMIFA family models. Histograms comparing the bin counts between the modelled and replicate data sets for each variable, under the IMIFA model, are given in Appendix D.

3.2 Spectral Metabolomic Data

IMIFA is employed to cluster spectral metabolomic data for which $N \ll p$ (Figure 6). The data are nuclear magnetic resonance spectra consisting of $p = 189$ spectral peaks from urine samples of $N = 18$ participants, half of which are known to have epilepsy

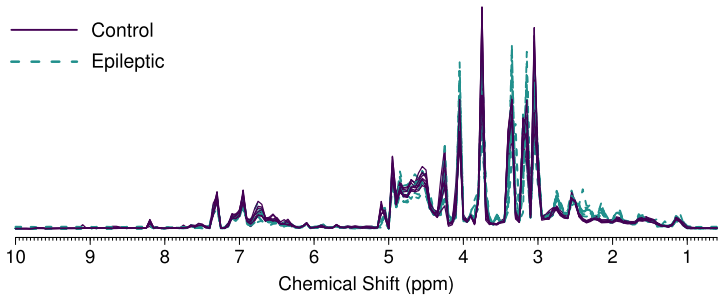

Figure 6: Raw spectral metabolomic data.

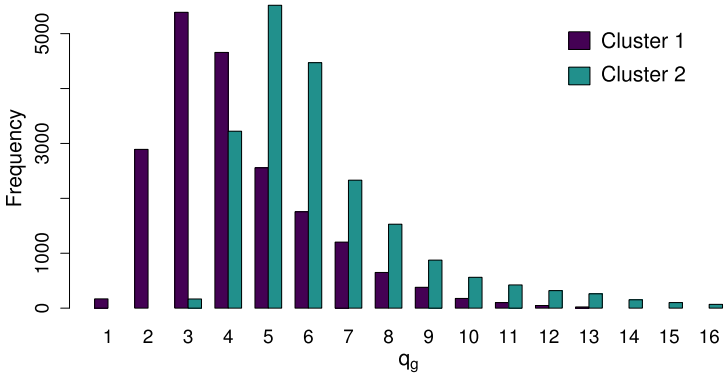

Figure 7: Posterior distribution of q_g under the IMIFA model fit to the metabolomic data.

(Carmody and Brennan, 2010; Nyamundanda et al., 2010). Interest lies in uncovering any underlying clustering structure given the $N \ll p$ setting.

Data were mean-centred and Pareto scaled (van den Berg et al., 2006). Although $N \ll p$, no restrictions are imposed on the uniquenesses as the sample variances are quite imbalanced. Fitting MIFA models for $G = 1, \ldots, 5$ is feasible as N is small. The BICM criterion chooses $\widehat{G} = 2$ as optimal and one participant is misclassified. IMIFA, however, unanimously visits a 2-cluster model and perfectly uncovers the group structure.

The modal estimates of the number of factors in each IMIFA cluster are $\widehat{q}_1 = 3$ $[2, 9]$ and $\widehat{q}_2 = 5$ $[4, 13]$ (see Figure 7). Cluster 1 corresponds to the control group and Cluster 2 to the epileptic participants. Figure 8 illustrates the $p \times \widehat{q}_g$ posterior mean loadings matrices, based on retained samples with $\widehat{q}_g$ or more factors, after Procrustes rotation to a common template for both clusters. The sparsity and shrinkage induced by the MGP prior is apparent, as is the greater complexity in Cluster 2, given the greater variation in colour and larger number of factors. For instance, many elevated loadings are visible for chemical shift values between 8 and 10 for the first two factors in Cluster 2; this activity is not present for other factors in either cluster. In general, the distributions of the

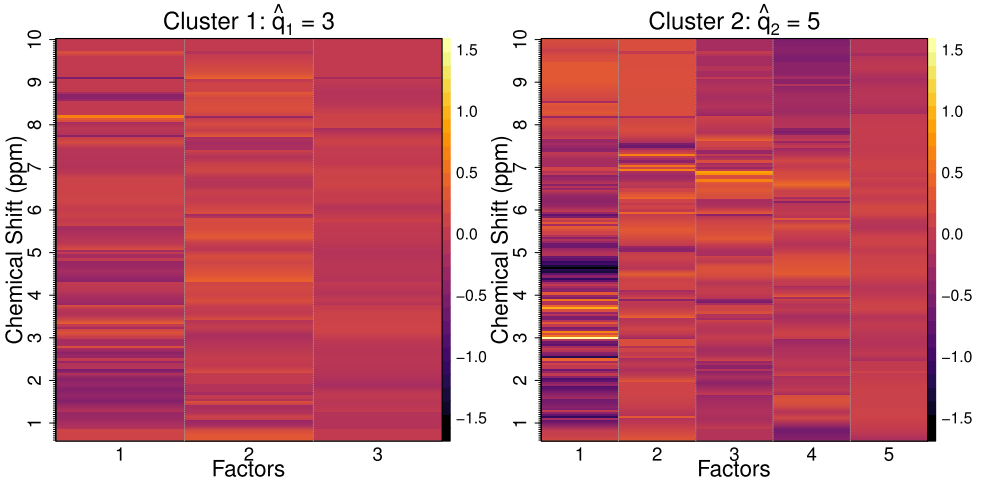

Figure 8: Heat maps, calibrated to a common colour scale, of posterior mean loadings matrices in the clusters uncovered by fitting IMIFA to the spectral metabolomic data.

loadings within a factor exhibit narrow spread around zero, particularly for the cluster of control participants, with the exception of the regions of the spectrum corresponding to the large peaks between chemical shifts of 3 and 5 in Figure 6.

IMIFA outperforms the optimal $\widehat{G} = 3$ `mclust` model and the optimal $\widehat{G} = 2$, $\widehat{q} = 5$ `pgmm` model, with respective ARI values of 0.73 and 0.27. The clustering performance of the optimal MFMA model is identical to the optimal MIFA model described above. Given the $N \ll p$ nature of the data, spectral clustering with the Gaussian kernel (Ng et al., 2001) is also considered. The eigengap heuristic suggests $\widehat{G} = 2$ and a perfect clustering is achieved almost instantaneously. However, the approach does not characterise the uncovered clusters in an interpretable manner, nor provide estimates of cluster membership uncertainty as given by model-based clustering approaches such as IMIFA.

The median PPRE for the IMIFA model of 0.21 $[0.18, 0.24]$ shows good model fit, given the size and dimensionality of the data. The median PSRF upper 95% confidence limits, using three randomly initialised auxiliary chains, for the cluster means, uniquenesses, loadings, and mixing proportions of $1.01\,(0.01)$, $1.00\,(< 0.01)$, $1.01\,(0.08)$, and $1.00\,(< 0.01)$ respectively, show good mixing also (standard deviations in parentheses). Notably, all chains yield the same inference about $\widehat{G}$ and $\widehat{\mathbf{Q}}$. So too, again, does the OMIFA model, although its model fit is inferior (median PPRE=0.26).

3.3 Handwritten Digit Data

A final illustration of IMIFA is given through its application to handwritten digit data from the United States Postal Service (USPS; Hastie et al., 2001). Here $N = 7,291$ images of the digits $0, \ldots, 9$ are considered, taken from handwritten zip codes. The

	0	1	2	3	4	5	6	7	8	9	$\widehat{\pi}_g$	$\widehat{q}_g$
1	359										0.05	4 [2, 8]
2	58		12			3	2				0.01	3 [2, 7]
3	108										0.01	2 [1, 4]
4	9										0.00	16 [3, 16]
5	95										0.01	4 [1, 8]
6	308					3					0.04	7 [4, 10]
7		844			2						0.12	2 [0, 4]
8		133							1		0.02	1 [0, 4]
9		2	392	10		1					0.05	7 [5, 12]
10	59		121	93	19	91	13	2	25	4	0.06	12 [9, 16]
11				136		64					0.03	5 [2, 9]
12					38	1		1			0.01	2 [0, 8]
13	25		3	7	98	51	2	36	59	28	0.04	8 [5, 12]
14	48		73	61	62	135	32	1	16	6	0.06	8 [6, 12]
15	1						83				0.01	3 [1, 7]
16	1						74				0.01	2 [1, 5]
17		2			4	19	381		2		0.06	2 [1, 6]
18								207			0.03	4 [1, 8]
19	123	8	129	348	247	184	77	26	420	84	0.23	6 [3, 9]
20		16	1	3	120	1		338	19	451	0.13	2 [1, 6]
21					62	3		34		71	0.02	3 [1, 6]

Table 4: Cross tabulation of the IMIFA model's MAP clustering (rows) against true digit labels (columns) for the USPS data. Cells that are 0 are blank for clarity. Posterior means $\widehat{\pi}_g$ and modal estimates $\widehat{q}_g$, with associated 95% credible intervals, are also given.

data are not balanced in terms of digit labels. Each image is a 16×16 grayscale grid concatenated into a $p = 256$-dimensional vector; data were mean-centred but not scaled. Such data are often considered in the context of manifold learning, positing that the data dimensionality is artificially high.

Given N and p, fitting a range of MFA or MIFA models is practically infeasible. Results of a single IMIFA run are presented here. For these data, it is reasonable to expect the number of components to grow as the sample size grows. It is anticipated that the flexibility afforded by having cluster-specific numbers of factors will help characterise digits with different geometric features.

The IMIFA model visited a $\widehat{G} = 21$ cluster solution in all posterior samples; Table 4 cross-tabulates the MAP clustering against the known digit labels and achieves an ARI of 0.33. The median PPRE of 0.05 $[0.04, 0.06]$ indicates good model fit. The overdispersed chains used to compute the PSRF diagnostics lead to identical inference about the number of clusters but slightly different inference about the modal numbers of cluster-specific factors. The ARI values between each resulting pair of MAP partitions were all in excess of 0.93. As before, good mixing is indicated by median PSRF upper 95% confidence limits for the cluster means, uniquenesses, and mixing proportions of 1.01 (0.01), 1.01 (0.01), and 1.01 (< 0.01), respectively. In computing the diagnostic

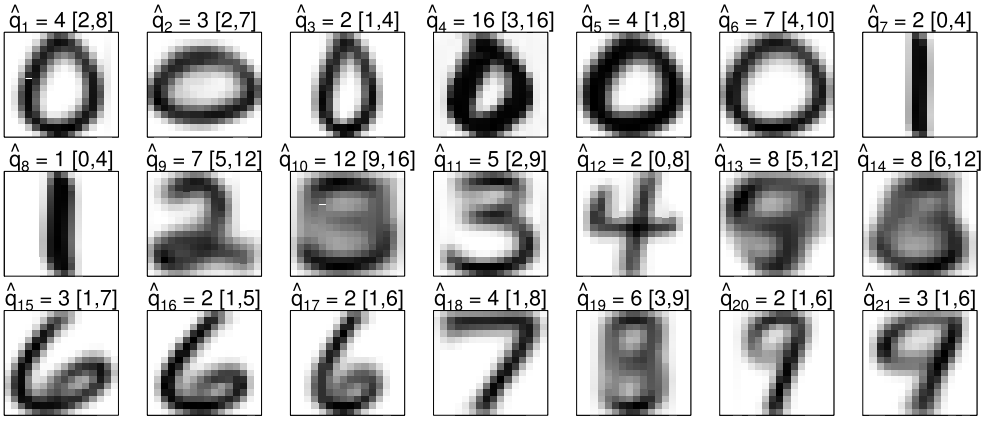

Figure 9: Posterior mean images for clusters uncovered by fitting IMIFA to the USPS data. Plots are ordered according to Table 4 and labelled with the modal $\widehat{q}_g$.

for the loadings – 1.14 (0.35) – only the first factor, common to all loadings matrices across all clusters in all chains, was considered for reasons of fairness and computational resource constraints.

Generally, IMIFA assigns images of the same digit, albeit written differently, to different clusters. Posterior mean images for each cluster are shown in Figure 9, ordered, as is Table 4, from 0 to 9 according to the digit most frequently assigned to the related cluster. Cluster 7 and the smaller cluster 8 capture the digit 1 written in a straight and slanted fashion, respectively. Clusters 15, 16, and 17 represent the digit 6 written with extended, medium, and compact loop curvature, respectively. Notably, cluster 15 requires more factors than clusters 16 and 17. A similar interpretation follows for clusters 20 and 21 ($\widehat{q}_{20} = 2, \widehat{q}_{21} = 3$), capturing the digit 9 with a small and large loop, respectively. Cluster 19 appears to represent the digit 8 and has a large number of factors ($\widehat{q}_{19} = 6$) in comparison, say, to clusters 7 and 8 ($\widehat{q}_7 = 2, \widehat{q}_8 = 1$) which capture the digit 1. This is intuitive, as 8 is a more geometrically complex than 1. Many clusters capture the digit 0, with differing degrees of elongation and border thickness. Of concern here is cluster 4, containing just 9 observations; that $\widehat{q}_4 = 16$, the upper AGS limit, suggests the model struggles to shrink the number of factors in poorly populated clusters. This difficulty is highlighted further in the simulation studies in Appendix B. Finally, Table 4 indicates that clusters 10, 13, and 14 also capture several other digits, all of which are reflected in the blurriness of the resulting posterior mean images and in $\widehat{q}_{10}$, $\widehat{q}_{13}$, and $\widehat{q}_{14}$ being quite large. The cluster-membership uncertainties are visualised in Appendix D.

It is computationally infeasible to run `mclust`, `pgmm`, or MFMA on these large data, as an exhaustive model search would be too vast. For comparative purposes, a DP-BP model (Chen et al., 2010) is fitted; this approach also simultaneously assumes infinitely many components and factors. It finds 43 clusters, each with around 14 factors, and achieves an ARI of 0.32. Cross tabulating this clustering against the 21 clusters of the IMIFA model shows that some of the DP-BP clusters are encapsulated by the larger

IMIFA clusters. IMIFA is thus the more parsimonious approach and affords greater cluster-specific factor flexibility. Additionally, a finite mixture of matrix-normal distributions (Viroli, 2011) is also fitted. This approach accounts for the grid nature of the data, but is computationally infeasible for $G > 15$ and requires a model selection strategy. The optimal model according to BIC yields $\widehat{G} = 12$ and ARI $= 0.38$. While neither IMIFA nor the DP-BP model account for the spatial structure in the data, they demonstrate comparative performance without the need for a computationally expensive model search.

4 Discussion

The IMIFA model is a Bayesian nonparametric approach to clustering high-dimensional data using factor-analytic mixture models. By extending the MGP prior (Bhattacharya and Dunson, 2011) to the PYP-MGP setting, the model sidesteps the fraught and computationally intensive task of determining the optimal number of clusters and factors using model selection criteria. Thus, the IMIFA model is recommended when fitting factor-analytic mixtures in settings where an exhaustive model search is computationally infeasible. Though IMIFA is not entirely choice-free, it achieves improved clustering results by allowing factor-analytic models of different dimensions in different clusters. If small clusters are inferred, one may wish to prune or merge small clusters with the larger clusters (West et al., 1994) or assess whether the small clusters are in fact of domain-specific interest. While comparative performance can be achieved by the IMIFA and OMIFA models, one may wish to fit a MIFA or OMIFA model when the expectation is that the number of clusters is fixed or unlikely to grow with N, respectively.

Future research directions are varied and plentiful. Incorporating covariates, in the spirit of Bayesian factor regression models (West, 2003; Carvalho et al., 2008), would allow for direct inclusion of the weight and urine pH covariates available with the metabolomic data, for example. Furthermore, the models could be extended to the (semi-)supervised model-based classification setting where all (or some) of the data are labelled. While constraints on the uniquenesses across variables and/or clusters are allowed, there is scope for also constraining the loadings across clusters. Though the number of factors would no longer be cluster-specific, the common number of loadings columns would be estimated in a similarly automatic fashion. However, incorporating covariance matrix constraints in the IMIFA model family problematically reintroduces the need for model selection strategies, in order to choose between them.

As proposed by Bhattacharya and Dunson (2011), the MGP hyperparameters could be learned via Metropolis-Hastings, and thus also be made cluster-specific. This could help combat some difficulties identified in the simulation studies in Appendix B. For example, learning those related to local shrinkage may help when loadings are notably dense. Learning those related to column shrinkage may help in settings with many small clusters, where IMIFA struggles to adaptively truncate loadings columns. In principle, a further global shrinkage parameter ϖ could be added to the MGP prior to borrow information across clusters, i.e. $\lambda_{jkg} \mid \ldots \sim \mathrm{N}_1\big(0, \phi_{jkg}^{-1}\tau_{kg}^{-1}\sigma_g^{-1}\varpi^{-1}\big)$. Alternatively, the infinite factor prior of Legramanti et al. (2019) could be employed, which decouples control

over the shrinkage rate and the active loadings terms. Finally, the IMIFA family can in fact be considered as wider than the range of models presented here. For example, the IBP prior (Knowles and Ghahramani, 2007, 2011; Ročková and George, 2016) could be extended to the infinite mixture setting, as per the DP-BP model of Chen et al. (2010).

For applied problems, a mismatch between the assumed model and the data distribution will impact inference. Miller and Harrison (2013, 2014) highlight that posterior consistency for the number of non-empty clusters in infinite mixtures is contingent on correct specification of the component distributions. While they do not discourage the use of infinite mixtures for clustering, they show that a few tiny extra clusters are typically fitted and suggest robustifying inference. If the data distribution is close to but not exactly a finite mixture of Gaussians, an infinite Gaussian mixture will introduce more components as the amount of data increases. Potential avenues of exploration thus include considering the IMIFA model with the heavy tailed multivariate t-distribution (Peel and McLachlan, 2000). Similarly, modelling of complex component distributions can be achieved by considering the MFMA approach in the context of infinite factor models. Defining robust inference functions as in Lee and MacEachern (2014) or using nonparametric unimodal component distributions as in Rodriguez and Walker (2014) may also prove fruitful. Another means of robustifying inference is to explicitly include a noise component with zero factors to capture outliers which depart from the component multivariate normality assumption. Finally, a 'coarsened' posterior (Miller and Dunson, 2018) could be used for addressing misspecification, by conditioning on the event that the model generates data close to the observed data in a distributional sense.

Supplementary Material

Supplementary material: infinite mixtures of infinite factor analysers.
(DOI: 10.1214/19-BA1179SUPP; .pdf).

References

Baek, J., McLachlan, G. J., and Flack, L. K. (2010). "Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualization of high-dimensional data." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7): 1298–1309. 937

Bai, J. and Li, K. (2012). "Statistical analysis of factor models of high dimension." *The Annals of Statistics*, 40(1): 436–465. MR3014313. doi: https://doi.org/10.1214/11-AOS966. 937

Bhattacharya, A. and Dunson, D. B. (2011). "Sparse Bayesian infinite factor models." *Biometrika*, 98(2): 291–306. MR2806429. doi: https://doi.org/10.1093/biomet/asr013. 938, 940, 941, 956

Brooks, S. P. and Gelman, A. (1998). "Generative methods for monitoring convergence of iterative simulations." *Journal of Computational and Graphical Statistics*, 7(4): 434–455. MR1665662. doi: https://doi.org/10.2307/1390675. 946

Carmody, S. and Brennan, L. (2010). "Effects of pentylenetetrazole-induced seizures on metabolomic profiles of rat brain." *Neurochemistry International*, 56(2): 340–344. 952

Carmona, C., Nieto-barajas, L., and Canale, A. (2019). "Model based approach for household clustering with mixed scale variables." *Advances in Data Analysis and Classification*, 13(2): 559–583. MR3954522. doi: https://doi.org/10.1007/s11634-018-0313-6. 947

Carpaneto, G. and Toth, P. (1980). "Solution of the assignment problem." *ACM Transactions on Mathematical Software*, 6(1): 104–111. MR0551750. doi: https://doi.org/10.1145/355853.355872. 940

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008). "High-dimensional sparse factor modeling: applications in gene expression genomics." *Journal of the American Statistical Association*, 103(484): 1438–1456. MR2655722. doi: https://doi.org/10.1198/016214508000000869. 956

Chen, M., Silva, J., Paisley, J., Wang, C., Dunson, D. B., and Carin, L. (2010). "Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: algorithm and performance bounds." *IEEE Transactions on Signal Processing*, 58(12): 6140–6155. MR2790088. doi: https://doi.org/10.1109/TSP.2010.2070796. 943, 955, 957

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., 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(2): 212–229. 944

Diebolt, J. and Robert, C. P. (1994). "Estimation of finite mixture distributions through Bayesian sampling." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 56(2): 363–375. MR1281940. 937

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. 938, 941

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 1(2): 209–230. MR0350949. 938

Fokoué, E. and Titterington, D. M. (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. 938

Forina, M., Armanino, C., Lanteri, S., and Tiscornia, E. (1983). "Classification of olive oils from their fatty acid composition." In Martens, H. and Russrum Jr., H. (eds.), *Food Research and Data Analysis*, 189–214. Applied Science Publishers, London. 948

Frühwirth-Schnatter, S. (2010). *Finite mixture and Markov switching models*. Series in Statistics. New York: Springer. MR2265601. 940

Frühwirth-Schnatter, S. (2011). "Dealing with label switching under model uncertainty." In Mengersen, K. L., Robert, C. P., and Titterington, D. M. (eds.), *Mixtures: Estima-*

tion and Applications, Wiley Series in Probability and Statistics, 193–218. Chichester: John Wiley & Sons. MR2867716. doi: https://doi.org/10.1002/9781119995678. 940

Frühwirth-Schnatter, S. and Lopes, H. F. (2010). "Parsimonious Bayesian factor analysis when the number of factors is unknown." Technical report, The University of Chicago Booth School of Business. 940

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(1): 33–63. MR3935190. doi: https://doi.org/10.1007/s11634-018-0329-y. 947

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2004). *Bayesian data analysis*. Chapman and Hall/CRC Press, third edition. MR2027492. 946

Ghahramani, Z. and Hinton, G. E. (1996). "The EM algorithm for mixtures of factor analyzers." Technical report, Department of Computer Science, University of Toronto. 937

Ghosh, J. and Dunson, D. B. (2008). "Default prior distributions and efficient posterior computation in Bayesian factor analysis." *Journal of Computational and Graphical Statistics*, 18(2): 306–320. MR2749834. doi: https://doi.org/10.1198/jcgs.2009.07145. 940

Green, P. J. and Richardson, S. (2001). "Modelling heterogeneity with and without the Dirichlet process." *Scandinavian Journal of Statistics*, 28(2): 355–375. MR1842255. doi: https://doi.org/10.1111/1467-9469.00242. 947

Hastie, D. I., Liverani, S., and Richardson, S. (2014). "Sampling from Dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations." *Statistics and Computing*, 25(5): 1023–1037. MR3375633. doi: https://doi.org/10.1007/s11222-014-9471-3. 945

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning*. Springer Series in Statistics. New York: Springer, second edition. MR2722294. doi: https://doi.org/10.1007/978-0-387-84858-7. 953

Hubert, L. and Arabie, P. (1985). "Comparing partitions." *Journal of Classification*, 2(1): 193–218. 948

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). "Slice sampling mixture models." *Statistics and Computing*, 21(1): 93–105. MR2746606. doi: https://doi.org/10.1007/s11222-009-9150-y. 938, 944

Kass, R. E. and Raftery, A. E. (1995). "Bayes factors." *Journal of the American Statistical Association*, 90(430): 773–795. MR3363402. doi: https://doi.org/10.1080/01621459.1995.10476572. 938

Kim, S., Tadesse, M. G., and Vannucci, M. (2006). "Variable selection in clustering via Dirichlet process mixture models." *Biometrika*, 93(4): 877–893. MR2285077. doi: https://doi.org/10.1093/biomet/93.4.877. 947

Knott, M. and Bartholomew, D. J. (1999). *Latent variable models and factor analysis*. Number 7 in Kendall's library of statistics. London: Edward Arnold, second edition. MR1711686. 937

Knowles, D. and Ghahramani, Z. (2007). "Infinite sparse factor analysis and infinite independent components analysis." In Davies, M. E., James, C. J., Abdallah, S. A., and Plumbley, M. D. (eds.), *Independent component analysis and signal separation*, 381–388. Berlin, Heidelberg: Springer. 943, 957

Knowles, D. and Ghahramani, Z. (2011). "Nonparametric Bayesian sparse factor models with application to gene expression modeling." *The Annals of Applied Statistics*, 5(2B): 1534–1552. MR2849785. doi: https://doi.org/10.1214/10-AOAS435. 943, 957

Lee, J. and MacEachern, S. N. (2014). "Inference functions in high dimensional Bayesian inference." *Statistics and Its Interface*, 7(4): 477–486. MR3302376. doi: https://doi.org/10.4310/SII.2014.v7.n4.a5. 957

Legramanti, S., Durante, D., and Dunson, D. B. (2019). "Bayesian cumulative shrinkage for infinite factorizations." *arXiv*:1902.04349. 956

McLachlan, G. J. and Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons. MR1789474. doi: https://doi.org/10.1002/0471721182. 937

McNicholas, P. D. (2010). "Model-based classification using latent Gaussian mixture models." *Journal of Statistical Planning and Inference*, 140(5): 1175–1181. MR2581120. doi: https://doi.org/10.1016/j.jspi.2009.11.006. 948

McNicholas, P. D., ElSherbiny, A., McDaid, A. F., and Murphy, T. B. (2018). `pgmm`*: parsimonious Gaussian mixture models*. R package version 1.2.3. URL https://cran.r-project.org/package=pgmm. 949

McNicholas, P. D. and Murphy, T. B. (2008). "Parsimonious Gaussian mixture models." *Statistics and Computing*, 18(3): 285–296. MR2413385. doi: https://doi.org/10.1007/s11222-008-9056-0. 937, 940

McParland, D., Gormley, I. C., McCormick, T. H., Clark, S. J., Kabudula, C. W., and Collinson, M. A. (2014). "Clustering South African households based on their asset status using latent variable models." *The Annals of Applied Statistics*, 8(2): 747–767. MR3262533. doi: https://doi.org/10.1214/14-AOAS726. 940

Miller, J. W. and Dunson, D. B. (2018). "Robust Bayesian inference via coarsening." *Journal of the American Statistical Association*, 114(527): 1113–1125. MR4011766. doi: https://doi.org/10.1080/01621459.2018.1469995. 957

Miller, J. W. and Harrison, M. T. (2013). "A simple example of Dirichlet process mixture inconsistency for the number of components." *Advances in Neural Information Processing Systems*, 26: 199–206. 947, 957

Miller, J. W. and Harrison, M. T. (2014). "Inconsistency of Pitman-Yor process mixtures for the number of components." *The Journal of Machine Learning Research*, 15(1): 3333–3370. MR3277163. 947, 957

Müller, P. and Mitra, R. (2013). "Bayesian nonparametric inference – why and how." *Bayesian Analysis*, 8(2): 269–360. MR3066939. doi: https://doi.org/10.1214/13-BA811. 947

Murphy, K., Viroli, C., and Gormley, I. C. (2019a). "Supplementary material: infinite mixtures of infinite factor analysers." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1179SUPP. 939

Murphy, K., Viroli, C., and Gormley, I. C. (2019b). `IMIFA`*: infinite mixtures of infinite factor analysers and related models*. R package version 2.1.0. URL https://cran.r-project.org/package=IMIFA. 939

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2001). "On spectral clustering: analysis and an algorithm." In *Advances in neural information processing systems*, 849–856. Cambridge, MA, USA: MIT Press. 953

Nyamundanda, G., Brennan, L., and Gormley, I. C. (2010). "Probabilistic principle component analysis for metabolomic data." *BMC Bioinformatics*, 11(571): 1–11. 952

Paisley, J. and Carin, L. (2009). "Nonparametric factor analysis with Beta process priors." In *Proceedings of the 26th annual international conference on machine learning*, ICML '09, 777–784. New York, NY, USA: ACM. 943

Papaspiliopoulos, O. and Roberts, G. O. (2008). "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models." *Biometrika*, 95(1): 169–186. MR2409721. doi: https://doi.org/10.1093/biomet/asm086. 945

Papastamoulis, P. (2018). "Overfitting Bayesian mixtures of factor analyzers with an unknown number of components." *Computational Statistics & Data Analysis*, 124: 220–234. MR3787623. doi: https://doi.org/10.1016/j.csda.2018.03.007. 938, 943

Peel, D. and McLachlan, G. J. (2000). "Robust mixture modelling using the t distribution." *Statistics and Computing*, 10: 339–348. 957

Perman, M., Pitman, J., and Yor, M. (1992). "Size-biased sampling of Poisson point processes and excursions." *Probability Theory and Related Fields*, 92(1): 21–39. MR1156448. doi: https://doi.org/10.1007/BF01205234. 938

Pitman, J. (1996). "Random discrete distributions invariant under size-biased permutation." *Advances in Applied Probability*, 28(2): 525–539. MR1387889. doi: https://doi.org/10.2307/1428070. 938, 944

Pitman, J. and Yor, M. (1997). "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator." *The Annals of Probability*, 25(2): 855–900. MR1434129. doi: https://doi.org/10.1214/aop/1024404422. 938

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). "CODA: convergence diagnosis and output analysis for MCMC." *R News*, 6(1): 7–11. 946

Raftery, A. E., Newton, M., Satagopan, J., and Krivitsky, P. (2007). "Estimating the integrated likelihood via posterior simulation using the harmonic mean identity." In *Bayesian statistics 8*, 1–45. MR2433201. 942

R Core Team (2019). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 939

Richardson, S. and Green, P. J. (1997). "On Bayesian analysis of mixtures with an unknown number of components (with discussion)." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4): 731–792. MR1483213. doi: https://doi.org/10.1111/1467-9868.00095. 937, 946

Ročková, V. and George, E. I. (2016). "Fast Bayesian factor analysis via automatic rotations to sparsity." *Journal of the American Statistical Association*, 111(516): 1608–1622. MR3601721. doi: https://doi.org/10.1080/01621459.2015.1100620. 943, 957

Rodriguez, C. E. and Walker, S. G. (2014). "Univariate Bayesian nonparametric mixture modeling with unimodal kernels." *Statistics and Computing*, 24(1): 35–49. MR3147696. doi: https://doi.org/10.1007/s11222-012-9351-7. 957

Rousseau, J. and Mengersen, K. (2011). "Asymptotic behaviour of the posterior distribution in overfitted mixture models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5): 689–710. MR2867454. doi: https://doi.org/10.1111/j.1467-9868.2011.00781.x. 943

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*, volume 104 of *Monographs on statistics and applied probability*. London: Chapman and Hall/CRC Press. MR2130347. doi: https://doi.org/10.1201/9780203492024. 945

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). "mclust 5: clustering, classification and density estimation using Gaussian finite mixture models." *The R Journal*, 8(1): 289–317. 945, 949

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4): 583–639. MR1979380. doi: https://doi.org/10.1111/1467-9868.00353. 938

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2014). "The deviance information criterion: 12 years on." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3): 485–493. MR3210727. doi: https://doi.org/10.1111/rssb.12062. 938

Stephens, M. (2000). "Bayesian analysis of mixture models with an unknown number of components – an alternative to reversible jump methods." *The Annals of Statistics*, 28(1): 40–74. MR1762903. doi: https://doi.org/10.1214/aos/1016120364. 946

Tipping, M. E. and Bishop, C. M. (1999). "Mixtures of probabilistic principal component analyzers." *Neural Computation*, 11(2): 443–482. 940

van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., and van der Werf, M. J. (2006). "Centering, scaling, and transformations: improving the biological information content of metabolomics data." *BMC Genomics*, 7(1): 142. 952

van Havre, Z., White, N., Rousseau, J., and Mengersen, K. (2015). "Overfitting Bayesian mixture models with an unknown number of components." *PloS one*, 10(7): e0131739. 943

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. 937, 949

Viroli, C. (2011). "Finite mixtures of matrix normal distributions for classifying three-way data." *Statistics and Computing*, 21(4): 511–522. MR2826689. doi: https://doi.org/10.1007/s11222-010-9188-x. 956

Walker, S. G. (2007). "Sampling the Dirichlet mixture model with slices." *Communications in Statistics – Simulation and Computation*, 36(1): 45–54. MR2370888. doi: https://doi.org/10.1080/03610910601096262. 944

Wang, C., Pan, G., Tong, T., and L, Z. (2015). "Shrinkage estimation of large dimensional precision matrix using random matrix theory." *Statistica Sinica*, 25(3): 993–1008. MR3409734. 940

Wang, Y., Canale, A., and Dunson, D. B. (2016). "Scalable geometric density estimation." In Gretton, A. and Robert, C. P. (eds.), *Proceedings of the 19th international conference on artificial intelligence and statistics*, volume 51 of *Proceedings of Machine Learning Research*, 857–865. Cadiz, Spain: PMLR. 943

West, M. (2003). "Bayesian factor regression models in the "large p, small n" paradigm." In *Bayesian statistics 7*, 723–732. Oxford University Press. MR2003537. 956

West, M., Müller, P., and Escobar, M. D. (1994). "Hierarchical priors and mixture models, with applications in regression and density estimation." In Smith, A. F. M. and Freeman, P. R. (eds.), *Aspects of uncertainty: a tribute to D. V. Lindley*, 363–386. New York: John Wiley & Sons. MR1309702. 956

Xing, E. P., Sohn, K. A., Jordan, M. I., and Teh, Y. W. (2006). "Bayesian multi-population haplotype inference via a hierarchical Dirichlet process mixture." In *Proceedings of the 23rd International Conference on Machine Learning*, 1049–1056. ACM. 947

Yellott, J. I., Jr. (1977). "The relationship between Luce's choice axiom, Thurstone's theory of comparative judgment, and the double exponential distribution." *Journal of Mathematical Psychology*, 15(2): 109–144. MR0449795. doi: https://doi.org/10.1016/0022-2496(77)90026-8. 945

Yerebakan, H. Z., Rajwa, B., and Dundar, M. (2014). "The infinite mixture of infinite Gaussian mixtures." In *Advances in Neural Information Processing Systems*, 28–36. 947

Acknowledgments

This research was supported by the Science Foundation Ireland funded Insight Centre for Data Analytics in University College Dublin under grant number SFI/12/RC/2289_P2. The authors thank the members of the UCD Working Group in Statistical Learning and Prof. Adrian Raftery's Working Group in Model-based Clustering and Prof. David Dunson for helpful discussions. The authors also thank Prof. Lorraine Brennan (UCD), for the metabolomic data, and the anonymous reviewers for constructive feedback from which this work greatly benefited.