Infinite Mixtures of Infinite Factor Analysers

Notes & Derivations

Keefe Murphy[1, 2], Dr. Claire Gormley[1, 2], and Dr. Cinzia Viroli[3]

[1]School of Mathematics and Statistics, UCD
[2]Insight Centre for Data Analytics, UCD
[3]Department of Statistical Sciences, University of Bologna

Contents

1 Introduction

1.1 Background

Modern clustering problems are increasingly high-dimensional, in the sense that the dimension of the feature vectors may be comparable to or even greater than the number of observations. In such cases, many common clustering techniques are known to perform poorly, and may even be intractable. We introduce a 'choice-free' Bayesian nonparametric approach to fitting mixture models with a factor analytic structure, with particular focus on clustering $n \ll p$ datasets. In particular, we propose a suite of models of varying degress of complexity – (Infinite) Factor Analysis (FA/IFA), Mixtures of (Infinite) Factor Analysers (MFA/MIFA), Overfitted Mixtures of (Infinite) Factor Analysers (OMFA/OMIFA), Infinite Mixtures of (Infinite) Factor Analysers (IMFA/IMIFA).

1.2 Model Set-Up

Let $\underline{x} = (x_1, x_2, \ldots, x_p)^T$ have mean $\underline{\mu}$ and covariance matrix Σ. Orthogonal factor analysis is a Gaussian latent variable model, often used as a dimension reduction technique, under which $\underline{x}$ is linearly dependent upon a few ($q \ll p$) unobservable random variables $\underline{\eta}_i$, called *common factors* and p additional sources of variation $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ called *specific factors*, for $i = 1, \ldots, N$ observations, s.t.

$$\underline{x}_i = \underline{\mu} + \Lambda\underline{\eta}_i + \underline{\varepsilon}_i$$

where	$\underline{x}_i$	$\rightarrow$	$(p \times 1)$	observation vector
	$\underline{\mu}$	$\rightarrow$	$(p \times 1)$	overall mean vector
	Λ	$\rightarrow$	$(p \times q)$	loadings matrix
	$\underline{\eta}_i$	$\rightarrow$	$(q \times 1)$	vector of factor scores for obs i
	$\underline{\varepsilon}_i$	$\rightarrow$	$(p \times 1)$	vector of errors for obs i

Λ_{jk} is the *factor loading* of the j-th variable on the k-th factor of the $(p \times q)$ Λ matrix of factor loadings. If we assume the data have been centred to have column means of 0 then we have

$$\left(\underline{x}_i - \underline{\mu}\right)_{(p\times 1)} = \underline{x}^{\star}_{i_{(p\times 1)}} = \Lambda_{(p\times q)}\underline{\eta}_{i_{(q\times 1)}} + \underline{\varepsilon}_{i_{(p\times 1)}} \tag{1.1}$$

1.3 Assumptions

1. $\underline{\varepsilon}_i$ and $\underline{\eta}_i$ are independent, s.t. $\underline{\eta}_i \perp\!\!\!\perp \underline{\epsilon}_i$ and $\text{Cov}\left(\underline{\eta}_i, \underline{\varepsilon}_i\right) = \text{E}\left(\underline{\eta}_i\underline{\varepsilon}_i^T\right) = 0$

2. $\text{E}\left(\underline{\varepsilon}_i\right) = \underline{0}$ and $\text{Cov}\left(\underline{\varepsilon}_i\right) = \begin{pmatrix} \psi_1 & 0 & \ldots & 0 \\ 0 & \psi_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \psi_p \end{pmatrix} = \Psi$

   $\therefore \underline{\varepsilon}_i \sim \text{MVN}_p\left(\underline{0}, \Psi\right)$, where Ψ is a diagonal matrix whose non-zero elements $\psi_1, \ldots, \psi_p$ are known as *uniquenesses* (1.2)

3. $\text{E}\left(\underline{\eta}_i\right) = \underline{0}$ and $\text{Cov}\left(\underline{\eta}_i\right) = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix} = \mathcal{I}_q$

   $\therefore \underline{\eta}_i \sim \text{MVN}_q\left(\underline{0}, \mathcal{I}_q\right)$ (1.3)

2 Bayesian Framework

2.1 Likelihood

$$\begin{aligned}
\mathrm{E}\left(\underline{x}_i^\star\right) &= \mathrm{E}\left(\Lambda\underline{\eta}_i + \underline{\varepsilon}_i\right) \\
&= \Lambda\mathrm{E}\left(\underline{\eta}_i\right) + \mathrm{E}\left(\underline{\varepsilon}_i\right) \\
&= \underline{0} \\
\therefore \underline{x}_i^\star &\sim \mathrm{MVN}_p\left(\underline{0}, \Sigma\right)
\end{aligned} \tag{2.1}$$

$$\begin{aligned}
\text{Since } \underline{\varepsilon}_i &= \underline{x}_i^\star - \Lambda\underline{\eta}_i, \\
\Sigma &= \mathrm{Cov}\left(x_i\right) \\
&= \mathrm{E}\left[\left(\underline{x}_i - \underline{\mu}_i\right)\left(\underline{x}_i - \underline{\mu}_i\right)^T\right] \\
&= \mathrm{E}\left[\underline{x}_i^\star \underline{x}_i^{\star T}\right] \\
&= \mathrm{E}\left[\left(\Lambda\underline{\eta}_i + \underline{\varepsilon}_i\right)\left(\Lambda\underline{\eta}_i + \underline{\varepsilon}_i\right)^T\right] \\
&= \mathrm{E}\left[\left(\Lambda\underline{\eta}_i\right) + \underline{\varepsilon}_i\left(\Lambda\underline{\eta}_i\right)^T + \left(\Lambda\underline{\eta}_i\right)\underline{\varepsilon}_i^T + \underline{\varepsilon}_i\underline{\varepsilon}_i^T\right] \\
&= \Lambda\mathrm{E}\left(\underline{\eta}_i\underline{\eta}_i^T\right)\Lambda^T + \mathrm{E}\left(\underline{\varepsilon}_i\underline{\eta}_i^T\right)\Lambda^T + \Lambda\mathrm{E}\left(\underline{\eta}_i\underline{\varepsilon}_i^T\right) + \mathrm{E}\left(\underline{\varepsilon}_i\underline{\varepsilon}_i^T\right) \\
&= \Lambda\Lambda^T + \Psi \\
\therefore \underline{x}_i^\star &\sim \mathrm{MVN}_p\left(\underline{0}, \Lambda\Lambda^T + \Psi\right)
\end{aligned} \tag{2.2}$$

$$\begin{aligned}
\mathrm{E}\left(\underline{x}_i^\star \,|\, \underline{\eta}_i\right) &= \mathrm{E}\left(\Lambda\underline{\eta}_i + \underline{\varepsilon}_i \,|\, \underline{\eta}_i\right) \\
&= \Lambda\mathrm{E}\left(\underline{\eta}_i \,|\, \underline{\eta}_i\right) + \mathrm{E}\left(\underline{\varepsilon}_i \,|\, \underline{\eta}_i\right) \\
&= \Lambda\underline{\eta}_i \\
\mathrm{Cov}\left(\underline{x}_i^\star \,|\, \underline{\eta}_i\right) &= \mathrm{E}\left[\left(\underline{x}_i^\star - \Lambda\underline{\eta}_i\right)\left(\underline{x}_i^\star - \Lambda\underline{\eta}_i\right)^T \,|\, \underline{\eta}_i\right] \\
&= \mathrm{E}\left(\underline{\varepsilon}_i\underline{\varepsilon}_i^T \,|\, \underline{\eta}_i\right) \\
&= \Psi \\
\therefore \underline{x}_i^\star \,|\, \underline{\eta}_i &\sim \mathrm{MVN}_p\left(\Lambda\underline{\eta}_i, \Psi\right)
\end{aligned} \tag{2.3}$$

The density of the data is then given by:

$$\mathrm{P}\left(\underline{x}_i^\star \,|\, \underline{\eta}_i, \Lambda, \Psi\right) = (2\pi)^{-\frac{p}{2}} \,|\, \Psi \,|\,^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N}\left(\underline{x}_i^\star - \Lambda\underline{\eta}_i\right)^T \Psi^{-1}\left(\underline{x}_i^\star - \Lambda\underline{\eta}_i\right)\right) \tag{2.4}$$

$$\propto \,|\, \Psi \,|\,^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left[\Psi^{-1}\left(X - \eta\Lambda\right)^T\left(X - \eta\Lambda\right)\right]\right)$$

$$\text{where } \Lambda_{(p\times q)} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1q} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \dots & \lambda_{pq} \end{pmatrix}$$

$$\& \ \eta_{(n\times q)} = \begin{pmatrix} \eta_{11} & \eta_{12} & \dots & \eta_{1q} \\ \eta_{21} & \eta_{22} & \dots & \eta_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_{n1} & \eta_{n2} & \dots & \eta_{nq} \end{pmatrix} \ \& \ \underline{\eta}_i \text{ is a column vector containing the entries of row } i \text{ of } \eta$$

2.2 Posterior Set-Up

$$\begin{aligned} \text{Likelihood} &= \prod_{i=1}^{N} \mathrm{P}\left(\underline{x}_i^\star \,|\, \theta\right) \\ &= \prod_{i=1}^{N} \mathrm{P}\left(\underline{x}_i^\star \,|\, \underline{\eta}_i, \Lambda, \Psi\right) \\ \text{where } \mathrm{P}\left(\underline{x}_i^\star \,|\, \underline{\eta}_i, \Lambda, \Psi\right) &\sim \mathrm{MVN}_p\left(\Lambda\underline{\eta}_i, \Psi\right) \qquad (2.5) \\ \text{Prior} &= \mathrm{P}\left(\theta\right) \\ &= \mathrm{P}\left(\eta\right)\mathrm{P}\left(\Lambda\right)\mathrm{P}\left(\Psi\right) \\ \text{Posterior} &= \text{Likelihood} \times \text{Prior} \\ \therefore \mathrm{P}\left(\eta, \Lambda, \Psi \,|\, X^\star\right) &\propto \mathcal{L}\left(X^\star \,|\, \eta, \Lambda, \Psi\right)\mathrm{P}\left(\eta\right)\mathrm{P}\left(\Lambda\right)\mathrm{P}\left(\Psi\right) \\ &\propto \left[\prod_{i=1}^{N} \mathrm{P}\left(\underline{x}_i^\star \,|\, \underline{\eta}_i, \Lambda, \Psi\right)\right]\left[\prod_{i=1}^{N} \mathrm{P}\left(\underline{\eta}_i\right)\right]\left[\prod_{j=1}^{p} \mathrm{P}\left(\underline{\Lambda}_j\right)\right]\left[\prod_{j=1}^{p} \mathrm{P}\left(\psi_j\right)\right] \qquad (2.6) \end{aligned}$$

Later on, especially as we move into the mixture case, it will be necessary to undo the centering, thereby removing the * on $\underline{x}_i^\star$, and reintroduce $\underline{\mu}$. This will necessitate multiplying the quantity in (2.6) by $\mathrm{P}\left(\underline{\mu}\right)$. However, we will proceed to derive the full conditionals we need for Gibbs Sampling using the centered notation for now as adjusting for $\underline{\mu}$ afterwards will be trivial.

3 Sampling from the Full Conditionals

3.1 Factor Scores - $\underline{\eta}_i$

$$\begin{aligned}\underline{\eta}_i &\sim \text{MVN}_q\left(\underline{0}, \mathcal{I}_q\right) \\ &= (2\pi)^{-\frac{q}{2}} \exp\left(-\frac{1}{2}\underline{\eta}_i^T\underline{\eta}_i\right)\end{aligned} \tag{3.1}$$

To obtain the full conditional for $\underline{\eta}_i$ we can multiply the likelihood by the prior in (3.1) s.t.

$$\begin{aligned}\text{P}\left(\underline{\eta}_i \,|\, , \underline{x}_i^\star, \Lambda, \Psi\right) &\sim \text{P}\left(\underline{x}_i^\star \,|\, \underline{\eta}_i, \Lambda, \Psi\right) \text{P}\left(\underline{\eta}_i\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\left(\underline{x}_i^\star - \Lambda\underline{\eta}_i\right)^T \Psi^{-1}\left(\underline{x}_i^\star - \Lambda\underline{\eta}_i\right) + \underline{\eta}_i^T\underline{\eta}_i\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left[-\underline{x}_i^{\star T}\Psi^{-1}\Lambda\underline{\eta}_i - \left(\Lambda\underline{\eta}_i\right)^T\Psi^{-1}\underline{x}_i^\star + \left(\Lambda\underline{\eta}_i\right)^T\Psi^{-1}\left(\Lambda\underline{\eta}_i\right) + \underline{\eta}_i^T\underline{\eta}_i\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left\{\underline{\eta}_i^T\left[\mathcal{I}_q + \Lambda^T\Psi^{-1}\Lambda\right]\underline{\eta}_i\right\} + \underline{x}_i^{\star T}\Psi^{-1}\Lambda\underline{\eta}_i\right)\end{aligned} \tag{3.2}$$

As this is the product of two MVN distributions we can expect the result to also be MVN. Typically,

$$\begin{aligned}\text{MVN}\left(\mu, \Sigma\right) &\propto \exp\left(-\frac{1}{2}\left(\underline{x} - \underline{\mu}\right)^T\Sigma^{-1}\left(\underline{x} - \underline{\mu}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\underline{x}^T\Sigma^{-1}\underline{x} - 2\underline{\mu}^T\Sigma^{-1}\underline{x} + \underline{\mu}^T\Sigma^{-1}\underline{\mu}\right)\right)\end{aligned}$$

We can identify the μ and Σ^{-1} terms from (3.2) above to yield

$$\text{P}\left(\underline{\eta}_i \,|\, \underline{x}_i^\star, \Lambda, \Psi\right) \sim \text{MVN}_q\left(\left[\mathcal{I}_q + \Lambda^T\Psi^{-1}\Lambda\right]^{-1}\Lambda^T\Psi^{-1}\underline{x}_i^\star, \left[\mathcal{I}_q + \Lambda^T\Psi^{-1}\Lambda\right]^{-1}\right) \tag{3.3}$$

However, we can reintroduce $\underline{\mu}$ and save on computational time if we implement the algorithm of Rue & Held (2005)[1]. In fact, we can extend this to block update the scores, thereby obviating the need to loop over i:

- Calculate $\Omega_\eta = \mathcal{I}_q + \Lambda^T\Psi^{-1}\Lambda$
- Compute the Cholesky Factorization $\Omega_\eta = U^TU$.
- Sample $\underline{z} \sim \text{MVN}_q\left(\underline{0}, \mathcal{I}_q\right)$ N times.
- Backsolve $U\underline{v} = \underline{z}^T$.
- Compute Ω_η^{-1} from U.
- Return $\left(\Omega_\eta^{-1}\Lambda^T\Psi^{-1}\left(C_n\underline{\mu}X\right)^T + \underline{v}\right)^T$ (3.4)
  where $C_n = \mathcal{I}_n - \frac{1}{n}\mathcal{O}$ and $\mathcal{O}$ is an $N \times N$ matrix of all 1's.

[1] To sample $x \sim \text{N}\left(\mu, \Omega^{-1}\right)$, find a matrix U – non-unique, and square or 'tall' – via Cholesky Decomposition s.t. $U^TU = \Omega$, sample from $z \sim \text{N}\left(0, 1\right)$, then backsolve $L^Tv = Uv = z$ s.t $x = \mu + v = \mu + L^{-T}z = \mu + U^{-1}z$. Then:

- $\text{E}\left(x\right) = \mu + U^{-1}\text{E}\left(z\right) = \mu$
- $\text{Cov}\left(x, x\right) = \text{Cov}\left(L^{-T}, z\right) = \left(L^TL\right)^{-1} = \Omega^{-1}$

3.2 Loadings Matrix - Λ

A Gaussian distribution is a conjugate prior for Λ, implying an MVN_q distribution prior for each row $\underline{\Lambda}_j$ of Λ s.t. $\underline{\Lambda}_j \sim \text{MVN}_q\left(\underline{0}, \Sigma_\lambda\right)$ where Σ_λ is a diagonal covariance matrix. As above, we can expect the result of the product of two MVN_q distributions to itself be distributed this way.

$$\begin{aligned}
\text{P}\left(\underline{\Lambda}_j \,|\, X^\star, \eta, \Psi\right) &\sim \text{P}\left(X^\star \,|\, \eta, \underline{\Lambda}_j, \Psi\right) \text{P}\left(\underline{\Lambda}_j \,|\, \Sigma_\lambda\right) \\
&\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N}\left(\underline{x}_i^\star - \underline{\Lambda}_j\underline{\eta}_i\right)^T \psi_j^{-1}\left(\underline{x}_i^\star - \underline{\Lambda}_j\underline{\eta}_i\right)\right)\exp\left(-\frac{1}{2}\left(\underline{\Lambda}_j^T\Sigma_\lambda^{-1}\underline{\Lambda}_j\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N}\left[-2\underline{x}_i^{\star T}\psi_j^{-1}\left(\underline{\Lambda}_j\underline{\eta}_i\right) + \left(\underline{\Lambda}_j\underline{\eta}_i\right)^T\psi_j^{-1}\left(\underline{\Lambda}_j\underline{\eta}_i\right) + \underline{\Lambda}_j^T\Sigma_\lambda^{-1}\underline{\Lambda}_j\right]\right) \\
&\propto \exp\left(\underline{\Lambda}_j\psi_j^{-1}\sum_{i=1}^{N}x_{ij}^{\star T}\underline{\eta}_i - \frac{1}{2}\underline{\Lambda}_j^T\left[\sum_{i=1}^{N}\psi_j^{-1}\underline{\eta}_i^T\underline{\eta}_i\right]\underline{\Lambda}_j - \frac{1}{2}\underline{\Lambda}_j^T\Sigma_\lambda^{-1}\underline{\Lambda}_j\right) \\
&\propto \exp\left(\underline{\Lambda}_j\left[\eta^T\psi_j^{-1}\underline{x}^{j^\star}\right] - \frac{1}{2}\underline{\Lambda}_j^T\left[\Sigma_\lambda^{-1} + \psi_j^{-1}\eta^T\eta\right]\underline{\Lambda}_j\right) \qquad (3.5)
\end{aligned}$$

where $\underline{x}^{j^\star}$ is an N-vector containing the elements of the j-th column of $X^\star$.

$$\begin{aligned}
\therefore \text{P}\left(\underline{\Lambda}_j \,|\, X^\star, \eta, \Psi\right) \sim &\text{MVN}_q\Big(\left[\Sigma_\lambda^{-1} + \psi_j^{-1}\eta^T\eta\right]^{-1}\eta^T\psi_j^{-1}\underline{x}^{j^\star}, \\
&\left[\Sigma_\lambda^{-1} + \psi_j^{-1}\eta^T\eta\right]^{-1}\Big) \qquad (3.6)
\end{aligned}$$

However, we can reintroduce $\underline{\mu}$ and save on computational time, as before, if we:

- Calculate $\Omega_{\lambda_j} = \Sigma_\lambda^{-1} + \psi_j^{-1}\eta^T\eta$.
- Compute the Cholesky Factorization $\Omega_{\lambda_j} = U^TU$.
- Sample $\underline{z} \sim \text{MVN}_q\left(\underline{0}, \mathcal{I}_q\right)$.
- Back-solve $U\underline{v} = \underline{z}$.
- Compute $\Omega_{\lambda_j}^{-1}$ from U.
- Return $\Omega_{\lambda_j}^{-1}\eta^T\psi_j^{-1}\left(\underline{x}^j - \underline{1}\mu_j\right) + \underline{v}$ (3.7)
  where $\underline{1}$ is an N-vector of all 1's.

3.3 Uniquenesses - Ψ

If we suggest an Inverse Wishart prior distribution for Ψ, we have:

$$\text{P}\left(\Psi\right) \propto |\,\Psi^{-1}\,|^{\frac{N+p+1}{2}}\exp\left(-\frac{1}{2}\text{tr}\left(\mathcal{S}^{-1^\star}\Psi\right)\right)$$

Using the fact that $\text{V}^{-1} \sim \text{Wish}_p\left(\nu, \Sigma\right)$ when $\text{V} \sim \text{Wish}_p^{-1}\left(m, \Sigma^{-1}\right)$ with $m = \nu + p + 1$ we get:

$$\text{P}\left(\Psi^{-1}\right) \propto |\,\Psi^{-1}\,|^{\frac{N}{2}}\exp\left(-\frac{1}{2}\text{tr}\left(\mathcal{S}^\star\Psi^{-1}\right)\right)$$

Since Ψ is a diagonal matrix:

$$\text{P}\left(\Psi^{-1}\right) \propto \prod_{j=1}^{p}|\,\psi_j^{-1}\,|^{\frac{N}{2}}\exp\left(-\frac{1}{2}\text{tr}\left(\mathcal{S}_j^\star\psi_j^{-1}\right)\right)$$

This suggests the prior for Ψ^{-1} is a product of p $\text{Ga}(\alpha, \beta)$ distributions. We choose hyperparameters as per Frühwirth-Schnatter & Lopes (2010), by bounding each ψ_j away from zero in such a way that Heywood problems are avoided, with a sufficiently large shape α and variable-specific rate hyperparameters β_j derived from $(\mathbf{S}^{-1})_{jj}$ or, if $p \geq n$, $(\mathbf{S}_{jj})^{-1}$.

$$
\begin{aligned}
\therefore \mathrm{P}\left(\Psi^{-1} \,|\, \alpha, \beta_j\right) &= \prod_{j=1}^{p} \mathrm{P}\left(\psi_j^{-1} \,|\, \alpha, \beta_j\right) \\
&\propto \prod_{j=1}^{p} \left(\psi_j^{-1}\right)^{\alpha-1} \exp\left(-\beta_j \psi_j^{-1}\right) \\
\therefore \mathrm{P}\left(\Psi^{-1} \,|\, X^\star, \eta, \Lambda\right) &\propto \mathrm{P}\left(X^\star \,|\, \eta, \Lambda\right) \mathrm{P}\left(\Psi^{-1} \,|\, \alpha, \beta_j\right) \\
&\propto \prod_{j=1}^{p} \left(\psi_j^{-1}\right)^{\frac{N}{2}} \exp\left(-\frac{\mathcal{S}_j^\star}{2}\psi_j^{-1}\right) \prod_{j=1}^{p} \left(\psi_j^{-1}\right)^{\alpha-1} \exp\left(-\beta_j \psi_j^{-1}\right) \\
&\propto \prod_{j=1}^{p} \left(\psi_j^{-1}\right)^{\frac{N}{2}+\alpha-1} \exp\left(-\left(\frac{\mathcal{S}_j^\star}{2} + \beta_j\right)\psi_j^{-1}\right) \qquad (3.8) \\
\text{where } \mathcal{S}_j^\star &= \sum_{i=1}^{N} \left(x_{ij} - \underline{\Lambda}_j \underline{\eta}_i\right)^2
\end{aligned}
$$

However, we can reintroduce $\underline{\mu}$ at this stage by rewriting:

$$
\mathcal{S}_j = \sum_{i=1}^{N} \left(x_{ij} - \mu_j - \underline{\Lambda}_j \underline{\eta}_i\right)^2
$$

Thus the posterior distribution of each ψ_j^{-1} is given by:

$$
\mathrm{P}\left(\psi_j^{-1} \,|\, X, F, \Lambda\right) \sim \text{Ga}\left(\alpha + \frac{N}{2}, \beta_j + \frac{S_j}{2}\right) \qquad (3.9)
$$

3.4 Reintroducing $\underline{\mu}$

We've already seen from (3.4), (3.7) and (3.9) that reintroducing μ to the other full conditionals is trivial. All that remains is to specify the conjugate Gaussian prior for μ itself, and to derive its full conditional. This implies an MVN_p distribution prior s.t. $\underline{\mu} \sim \text{MVN}_p\left(\underline{\tilde{\mu}}, \Sigma_\mu\right)$ where Σ_μ is a diagonal covariance matrix, typically the diagonal of the data's sample covariance matrix, and $\underline{\tilde{\mu}}$ is a vector of prior mean means, typically the sample mean. In the mixture case, this will be the sample mean of each initialised group. As above, we can expect the result of the product of two MVN_p distributions to itself be distributed this way.

$$
\begin{aligned}
\mathrm{P}\left(\underline{\mu} \,|\, X, \eta, \Psi, \Lambda\right) &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N}\left(\underline{x}_i - \underline{\mu} - \Lambda\underline{\eta}_i\right)^T \Psi^{-1}\left(\underline{x}_i - \underline{\mu} - \Lambda\underline{\eta}_i\right)\right) \exp\left(-\frac{1}{2}\left(\underline{\mu} - \underline{\tilde{\mu}}\right)^T \Sigma_\mu^{-1}\left(\underline{\mu} - \underline{\tilde{\mu}}\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\sum_{i=1}^{N}\left[-2\underline{x}_i^T\Psi^{-1}\underline{\mu} + 2\left(\Lambda\underline{\eta}_i\right)^T\Psi^{-1}\underline{\mu} + \underline{\mu}^T\Psi^{-1}\underline{\mu}\right] + \underline{\mu}^T\Sigma_\mu^{-1}\underline{\mu} - 2\underline{\tilde{\mu}}^T\Sigma_\mu^{-1}\underline{\mu}\right)\right) \\
&\propto \exp\left(\sum_{i=1}^{N}\underline{x}_i^T\Psi^{-1}\underline{\mu} - \sum_{i=1}^{N}\left(\Lambda\underline{\eta}_i\right)^T\Psi^{-1}\underline{\mu} - \frac{1}{2}\left[\underline{\mu}^T\left(\Sigma_\mu^{-1} + N\Psi^{-1}\right)\underline{\mu}\right] + \underline{\tilde{\mu}}^T\Sigma_\mu^{-1}\underline{\mu}\right) \\
\therefore \mathrm{P}\left(\underline{\mu} \,|\, X, \eta, \Psi, \Lambda\right) &\sim \text{MVN}_p\Bigg(\left[\Sigma_\mu^{-1} + N\Psi^{-1}\right]^{-1}\left(\Psi^{-1}(\sum_{i=1}^{N}\underline{x}_i - \sum_{i=1}^{N}\Lambda\underline{\eta}_i) + \Sigma_\mu^{-1}\underline{\tilde{\mu}}\right), \\
&\qquad\qquad \left[\Sigma_\mu^{-1} + N\Psi^{-1}\right]^{-1}\Bigg) \qquad (3.10)
\end{aligned}
$$

However, we can save on computational time, as before, if we:

- Calculate $\Omega_\mu = \Sigma_\mu^{-1} + N\Psi^{-1}$, which is a diagonal $p \times p$ matrix.
- Invert Ω_μ by inverting its diagonal elements.
- $\Omega_\mu^{-1} = U^T U$ can be obtained by taking the square root of Ω_μ since this matrix is diagonal.
- Sample $\underline{z} \sim \text{MVN}_p(\underline{0}, \mathcal{I}_p)$.
- Compute $\underline{v} = U^T \underline{z}$.
- Return $\Omega_\mu^{-1}\left(\Psi^{-1}(\sum_{i=1}^{N} \underline{x}_i - \sum_{i=1}^{N} \Lambda\underline{\eta}_i) + \Sigma_\mu^{-1}\tilde{\underline{\mu}}\right) + \underline{v}$ (3.11).

3.5 Gibbs Sampler Pseudo-Code

The conjugacy of the above priors facilitates MCMC sampling via efficient Gibbs updates:

i) Choose hyperparameters $\Sigma_\mu, \Sigma_\lambda, \alpha,$ and β, select q and initialise $\tilde{\underline{\mu}}$.

ii) Initalise:

$$\begin{aligned}
\underline{\mu}^{(0)} &\sim \text{MVN}_p\left(\tilde{\underline{\mu}}, \Sigma_\mu\right) \\
\underline{\eta}_i^{(0)} &\sim \text{MVN}_q\left(\underline{0}, \mathcal{I}_q\right) \qquad \forall\, i = 1, \ldots, N \\
\underline{\Lambda}_j^{(0)} &\sim \text{MVN}_q\left(\underline{0}, \Sigma_\lambda\right) \qquad \forall\, j = 1, \ldots, p \\
\psi_j^{-1^{(0)}} &\sim \text{Ga}\left(\alpha, \beta_j\right) \qquad \forall\, j = 1, \ldots, p
\end{aligned}$$

iii) For $t = 1, \ldots, T$, using the routines specified in (3.4), (3.7), (3.9) and (3.11):

$$\begin{aligned}
\text{a)} \quad \Omega_\mu^{(t)} &= \Sigma_\mu^{-1} + N\Psi^{-1^{(t-1)}} \\
\underline{\mu}^{(t)} &\sim \text{MVN}_p\left(\Omega_\mu^{-1^{(t)}}\left(\Psi^{-1}(\sum_{i=1}^{N} \underline{x}_i - \sum_{i=1}^{N} \Lambda\underline{\eta}_i) + \Sigma_\mu^{-1}\tilde{\underline{\mu}}\right), \Omega_\mu^{-1^{(t)}}\right) \\
\text{b)} \quad \Omega_\eta^{(t)} &= \mathcal{I}_q + \Lambda^{T^{(t-1)}}\Psi^{-1^{(t-1)}}\Lambda^{(t-1)} \\
\underline{\eta}_i^{(t)} &\sim \text{MVN}_q\left(\Omega_\eta^{-1^{(t)}}\Lambda^{T^{(t-1)}}\Psi^{-1^{(t-1)}}\left(\underline{x}_i - \underline{\mu}^{(t)}\right), \Omega_\eta^{-1^{(t)}}\right) \\
\text{c)} \quad \text{For } j &= 1, \ldots, p \\
\bullet\ \Omega_{\lambda_j}^{(t)} &= \Sigma_\lambda^{-1} + \psi_j^{-1^{(t-1)}}\eta^{T^{(t)}}\eta^{(t)} \\
\underline{\Lambda}_j^{(t)} &\sim \text{MVN}_q\left(\Omega_{\lambda_j}^{-1^{(t)}}\eta^{T^{(t)}}\psi_j^{-1^{(t-1)}}\left(\underline{x}^j - \underline{1}\mu_j^{(t)}\right), \Omega_{\lambda_j}^{-1^{(t)}}\right) \\
\bullet\ \psi_j^{-1^{(t)}} &\sim \text{Ga}\left(\alpha + \frac{N}{2}, \beta_j + \frac{S_j^{(t)}}{2}\right)
\end{aligned}$$

iv) Disregard the first B burn-in iterations and thin every K-th iteration.

v) Calculate the log-likelihood for each remaining sample. Then, using the largest value observed across these draws, BIC-MCMC, as defined by Frühwirth-Schnatter (2011), is determined by $2\ln\hat{\mathcal{L}} - k\ln(N)$, where $k = pq - \frac{q(q-1)}{2} + 2p$ is the effective number of parameters in the model. When choosing between competing models, the one with the highest BIC-MCMC is preferred. Alternatively, AIC-MCMC, or the BICM and AICM of Raftery et al. (2007) can be used. The latter are particularly useful for nonparametric models where k can be hard to quantify.

3.6 Issues Around Identifiability

Most covariance matrices Σ cannot be uniquely factored as $\Lambda\Lambda^T + \Psi$ where $q \ll p$. Let T be any $q \times q$ orthogonal matrix such that $TT^T = \mathcal{I}_q$. Then:

$$\begin{aligned}\underline{x}_i - \underline{\mu} &= \Lambda\underline{\eta}_i + \underline{\varepsilon}_i \\ &= \Lambda TT^T\underline{\eta}_i + \underline{\varepsilon}_i \\ &= \Lambda^\star\underline{\eta}_i^\star + \underline{\varepsilon}_i\end{aligned}$$

where $\Lambda^\star = \Lambda T$ and $\underline{\eta}_i^\star = T^T\underline{\eta}_i$. It follows that $\mathrm{E}\left(\underline{\eta}_i^\star\right) = \underline{0}$ and $\mathrm{Cov}\left(\underline{\eta}_i^\star\right) = \mathcal{I}_q$. Thus it is impossible, given the data X, to distinguish between Λ and $\Lambda^\star$ since they both generate the same covariance matrix Σ:

$$\begin{aligned}\Sigma &= \Lambda\Lambda^T + \Psi \\ &= \Lambda TT^T\Lambda^T + \Psi \\ &= \Lambda^\star{\Lambda^\star}^T + \Psi\end{aligned}$$

However, we can address this identifiability problem, using Procrustean methods, by mapping each iteration's loadings matrix to a common 'template' loadings matrix — which we have taken to be the loadings matrix at the end of the burn-in period. This Procustean map is a rotation only, i.e. translation, scaling, dilation, etc. are not permitted. We then also apply that same rotation matrix at each iteration to each sample of the matrix of factor scores. This amounts to *post-multiplying* the loadings and factor score matrices at each iteration by the Procrustes rotation matrix that maps to that iteration's loadings template and ensures sensible posterior means.

4 Introducing the Shrinkage Prior

4.1 Multiplicative Gamma Process Shrinkage Priors

A limitation of this approach is that a value for the number of factors q must be specified in advance, with selection of the optimal value performed by comparing some model selection criteria across models employing a range of q values. This becomes particularly difficult in the context of mixtures of factor analysers, for which ranges of values for the interdependent numbers of clusters G and factors q must be pre-specified, with the pair which optimise some model selection criterion chosen, because it's computationally intensive to search the space of candidate models. Furthermore, without the introduction of the shrinkage prior, it's generally only reasonable to fit models where q is the same across clusters, as is the case with PGMM models McNicholas & Murphy (2008); Ghahramani & Hinton (1996).

To remove this difficulty, we now propose the multiplicative gamma process shrinkage prior of Bhattacharya & Dunson (2011) on the factor loadings which allows the introduction of infinitely many factors, with the loadings increasingly shrunk towards zero as the column index increases. Their prior is placed on a parameter expanded factor loadings matrix without imposing any restriction on the loading elements, thereby making the induced prior on the covariance matrix invariant to the ordering of the data. The Gibbs sampler can still be used due to the joint conjugacy property of this prior, which allows block updating of the loadings matrix. Furthermore, these authors propose that an adaptive Gibbs sampler be used for automatically truncating the infinite loading matrix, through selection of the number of important factors, to one having finite columns. This facilitates posterior computation while providing a close approximation of the infinite factor model.

The exact specification of this shrinkage-type prior allows the degree of shrinkage to increase across the column index as follows:

$$\begin{aligned}
&\lambda_{jk} \,|\, \phi_{jk}, \tau_k \sim \mathrm{N}\left(0, \phi_{jk}^{-1}\tau_k^{-1}\right) \\
\text{s.t.} \quad &\underline{\lambda}_j \,|\, \underline{\phi}_j, \underline{\tau} \sim \mathrm{MVN}_{q^\star}\left(\underline{0}, \mathrm{D}_j\right) \qquad (4.1) \\
\text{where} \quad &\mathrm{D}_j^{-1} = \mathrm{diag}\left(\phi_{j1}\tau_1, \ldots, \phi_{jq^\star}\tau_{q^\star}\right) \\
&\phi_{jk} \sim \mathrm{Ga}\left(\nu, \nu\right) \qquad (4.2) \\
&\tau_k = \prod_{h=1}^{k} \delta_h \\
&\delta_1 \sim \mathrm{Ga}\left(\alpha_1, \beta_1\right), \quad \delta_h \sim \mathrm{Ga}\left(\alpha_2, \beta_2\right), \quad h \geq 2 \qquad (4.3)
\end{aligned}$$

where $\delta_h \, (h = 1, \ldots, \infty)$ are independent, τ_k is a *global* shrinkage parameter for the k-th column and the ϕ_{jk}s are *local* shrinkage parameters for the elements in the k-th column. The τ_ks are stochastically increasing under the restriction $\alpha_2 > \beta_2 + 1$, which favours more shrinkage as the column index increases according to Durante (2016). Typically $\beta_1 = \beta_2 = 1$. However, we find it useful to consider an alternative reparameterisation of the local shrinkage prior via $\phi_{jk} \sim \mathrm{Ga}\left(\nu + 1, \nu\right)$, s.t. the induced inverse-gamma prior on each ϕ_{jk}^{-1} is non-informative in the sense that it has expectation 1. Though these hyperparameters remain fixed under our consideration, they can also be learned from the data with the introduction of hyperpriors and Metropolis-Hastings steps. In any case, when extending this prior to mixture models, the hyperparameters α_1 and α_2 tend to need to be higher than the values suggested by Bhattacharya & Dunson (2011), in order to enforce stricter priors with a greater degree of shrinkage, as there will be less data in each group to affect the posterior loadings than in the overall dataset.

4.2 Defining new MGP Full Conditionals

We propose a Gibbs sampler for posterior computation, much like the one above, after truncating the loadings matrix to have $q^\star \ll p$ columns. An adaptive strategy for inference on the truncation level $q^\star$ is described in 4.3. For now, let's focus on the new full conditionals for the loadings matrix, global shrinkages, and local shrinkages which need to be derived in order to implement this. Once again, these parameters are initialised according to their priors. The other full conditionals are exactly as before, with just a small adjustment to the factor scores to allow for the truncation to $q^\star$ columns, i.e. $\mathrm{P}\left(\underline{\eta}_i \,|\, —\right) \sim \mathrm{MVN}_{q^\star}\left(\left[\mathcal{I}_{q^\star} + \Lambda_{q^\star}^T \Psi^{-1} \Lambda_{q^\star}\right]^{-1} \Lambda^T \Psi^{-1} \underline{x}_i - \underline{\mu}, \left[\mathcal{I}_{q^\star} + \Lambda_{q^\star}^T \Psi^{-1} \Lambda_{q^\star}\right]^{-1}\right)$

4.2.1 Loadings Matrix - Λ

Incorporating the new prior (4.1), and following the same steps as 3.2 above, it's trivial to show that the Λ_js now have independent conditionally conjugate posteriors given by:

$$\mathrm{P}\left(\Lambda_j \,|\, —\right) \sim \mathrm{MVN}_{q^\star}\left(\left[\mathrm{D}_j^{-1} + \psi_j^{-1}\eta^T\eta\right]^{-1} \eta^T \psi_j^{-1} \underline{x}^{j^\star}, \left[\mathrm{D}_j^{-1} + \psi_j^{-1}\eta^T\eta\right]^{-1}\right) \qquad (4.4)$$

However, we can reintroduce $\underline{\mu}$ and save on computational time, as before, if we follow the routine given in (3.7), with $\Omega_{\lambda_j} = \mathrm{D}_j^{-1} + \psi_j^{-1}\eta^T\eta$.

4.2.2 Local Shrinkage – ϕ_{jk}

Using the conditional prior in (4.1) and the reparameterised version of (4.2), the prior for ϕ_{jk}, we can derive the full conditional for the local shrinkage parameter as follows:

$$\begin{aligned}
\mathrm{P}\left(\phi_{jk} \,|\, —\right) \propto \mathrm{P}\left(\lambda_{jk} \,|\, \phi_{jk}, \tau_k\right) \mathrm{P}\left(\phi_{jk}\right) &\propto \frac{\phi_{jk}^{1/2} \tau_k^{1/2}}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\lambda_{jk}^2 \phi_{jk} \tau_k\right\} \phi_{jk}^{\nu} \exp\left\{-\nu\phi_{jk}\right\} \\
&\propto \phi_{jk}^{1/2} \phi_{jk}^{\nu} \exp\left\{\left(-\frac{1}{2}\lambda_{jk}^2 \tau_k - \nu\right) \phi_{jk}\right\} \\
&\propto \phi_{jk}^{\nu+1/2} \exp\left\{-\frac{1}{2}\left(2\nu + \lambda_{jk}^2 \tau_k\right) \phi_{jk}\right\}
\end{aligned}$$

Thus the full conditional for each ϕ_{jk} is given by:

$$\mathrm{P}\left(\phi_{jk} \,|\, —\right) \sim \mathrm{Ga}\left(\nu + \frac{3}{2}, \nu + \frac{\tau_k \lambda_{jk}^2}{2}\right) \tag{4.5}$$

4.2.3 Global Shrinkage – τ_k

Using the conditional prior in (4.1) and the prior for τ_k in (4.3) we can derive the full conditional for the global shrinkage parameter, in three stages – first by deriving and sampling from $\mathrm{P}\left(\delta_1 \,|\, —\right)$ & $\mathrm{P}\left(\delta_k \,|\, —\right)$ for $k \geq 2$, as follows below — and then obtaining the product $\tau_k = \prod_{h=1}^{k} \delta_h$ thereafter:

$$\begin{aligned}
\mathrm{P}\left(\delta_1 \,|\, —\right) &\propto \prod_{j=1}^{p} \prod_{k=1}^{q^\star} \mathrm{N}\left(\lambda_{jk} \,|\, 0, \phi_{jk}^{-1} \tau_k^{-1}\right) \times \mathrm{Ga}\left(\delta_1 \,|\, \alpha_1, \beta_1\right) \\
&\propto \prod_{j=1}^{p} \mathrm{N}\left(\lambda_{j1} \,|\, 0, \phi_{j1}^{-1} \tau_1^{-1}\right) \times \ldots \times \prod_{j=1}^{p} \mathrm{N}\left(\lambda_{jq^\star} \,|\, 0, \phi_{jq^\star}^{-1} \tau_{q^\star}^{-1}\right) \times \mathrm{Ga}\left(\delta_1 \,|\, \alpha_1, \beta_1\right) \\
&\propto \left(\phi_{j1}\tau_1\right)^{p/2} \exp\left(-\frac{1}{2}\sum_{j=1}^{p} \lambda_{j1}^2 \phi_{j1} \tau_1\right) \times \ldots \times \left(\phi_{jq^\star}\tau_{q^\star}\right)^{p/2} \exp\left(-\frac{1}{2}\sum_{j=1}^{p} \lambda_{jq^\star}^2 \phi_{jq^\star} \tau_{q^\star}\right) \\
&\qquad \times \delta_1^{\alpha_1 - 1} \exp\left(-\beta_1 \delta_1\right) \\
&\propto \left(\phi_{j1}\delta_1\right)^{p/2} \exp\left(-\frac{1}{2}\sum_{j=1}^{p} \lambda_{j1}^2 \phi_{j1} \delta_1\right) \times \ldots \times \left(\phi_{jq^\star}\delta_1\delta_2 \ldots \delta_{q^\star}\right)^{p/2} \exp\left(-\frac{1}{2}\sum_{j=1}^{p} \lambda_{jq^\star}^2 \phi_{jq^\star} \delta_1\delta_2 \ldots \delta_{q^\star}\right) \\
&\qquad \times \delta_1^{\alpha_1 - 1} \exp\left(-\beta_1 \delta_1\right) \\
&\propto \delta_1^{pq^\star/2 + \alpha_1 - 1} \exp\left(-\frac{\delta_1}{2}\left(\sum_{j=1}^{p} \lambda_{j1}^2 \phi_{j1} + \ldots + \lambda_{jq^\star}^2 \phi_{jq^\star} \delta_2 \ldots \delta_{q^\star} + 2\beta_1\right)\right) \\
&\propto \delta_1^{pq^\star/2 + \alpha_1 - 1} \exp\left(-\frac{\delta_1}{2}\left(\sum_{h=1}^{q^\star} \tau_h^{(1)} \sum_{j=1}^{p} \lambda_{jh}^2 \phi_{jh} + 2\beta_1\right)\right)
\end{aligned}$$

$$\text{where } \tau_h^{(k)} = \prod_{t=1}^{h} \frac{\delta_t}{\delta_k} \text{ for } k = 1, \ldots, q^\star \tag{4.6}$$

$$\therefore \mathrm{P}\left(\delta_1 \,|\, —\right) \sim \mathrm{Ga}\left(\alpha_1 + \frac{pq^\star}{2}, \beta_1 + \frac{1}{2}\sum_{h=1}^{q^\star} \tau_h^{(1)} \sum_{j=1}^{p} \lambda_{jh}^2 \phi_{jh}\right) \tag{4.7}$$

$$
\begin{aligned}
\mathrm{P}\left(\delta_k \,|\, —\right) &\propto \prod_{j=1}^{p}\prod_{k=1}^{q^\star} \mathrm{N}\left(\lambda_{jk} \,|\, 0, \phi_{jk}^{-1}\tau_k^{-1}\right) \times \mathrm{Ga}\left(\delta_k \,|\, \alpha_2, \beta_2\right) \\
&\propto \left(\phi_{j1}\delta_1\right)^{p/2} \exp\left(-\frac{1}{2}\sum_{j=1}^{p}\lambda_{j1}^2\phi_{j1}\delta_1\right) \times \ldots \times \left(\phi_{jq^\star}\delta_1\delta_2\ldots\delta_{q^\star}\right)^{p/2} \exp\left(-\frac{1}{2}\sum_{j=1}^{p}\lambda_{jq^\star}^2\phi_{jq^\star}\delta_1\delta_2\ldots\delta_{q^\star}\right) \\
&\qquad \times \delta_k^{\alpha_2-1}\exp\left(-\beta_2\delta_k\right) \\
&\propto \delta_k^{p/2(q^\star-k+1)+\alpha_2-1} \exp\left(-\frac{\delta_k}{2}\left(\sum_{h=k}^{q^\star}\tau_h^{(k)}\sum_{j=1}^{p}\lambda_{jh}^2\phi_{jh} + 2\beta_2\right)\right) \\
\therefore \mathrm{P}\left(\delta_k \,|\, —\right) &\sim \mathrm{Ga}\left(\alpha_2 + \frac{p}{2}\left(q^\star-k+1\right), \beta_2 + \frac{1}{2}\sum_{h=k}^{q^\star}\tau_h^{(k)}\sum_{j=1}^{p}\lambda_{jh}^2\phi_{jh}\right) \qquad (4.8)
\end{aligned}
$$

4.3 Adaptive Step

In practical situations, we expect to have relatively few important factors compared to the dimension p of the data. The most common approach for selecting the number of factors relies on fitting the finite factor model for different choices of $q^\star$, and then using the BIC, BIC-MCMC, or another model selection criterion. This approach can be difficult to implement for $N \ll p$ problems, and the BIC itself isn't well justified for factor models even for small to moderate p. However, the infinite factor model obviates the need for pre-specifying the number of factors, while the sparsity favouring prior on the loadings ensures that the effective number of factors is small when the truth is sparse. However, we need a computational strategy for choosing an appropriate truncation level $q^\star$ that strikes a balance between missing important factors by choosing $q^\star$ too small and wasting computational effort on an overly high truncation level. One can think of $q^\star$ as the effective number of factors, so that the contribution from adding additional factors is negligible. Starting with a conservative guess $\tilde{q}$ of $q^\star$, the posterior samples of $\Lambda_{\tilde{q}}$ from the Gibbs sampler contain information about the effective number of factors. At the t-th iteration, let $m^{(t)}$ denote the number of columns in $\Lambda_{\tilde{q}}$ having all elements in a pre-specified small neighbourhood of zero. Intuitively, $m^{(t)}$ of the factors have a negligible contribution at the t-th iteration. We then define $q^{\star(t)} = \tilde{q} - m^{(t)}$ to be the effective number of factors at iteration t. It's typically necessary to choose a very conservative upper-bound to be assured that $\tilde{q} \geq q^\star$, though this leads to wasted computational effort. Ideally, we would like to discard the redundant factors and continue sampling with a reduced number of loadings columns, to save on computation by discarding unimportant factors. Hence, the algorithm described in 3.5 above is modified to an adaptive Gibbs sampler, which tunes the number of factors as it progresses. We begin with a default $\tilde{q}$ value of $\min\left(\lfloor 3\ln(p)\rfloor, p, N-1\right)$. We adapt only after the burn-in period, in order to ensure we're sampling from the true posterior distribution before truncating the loadings matrix. We adapt with probability $\mathrm{p}\left(t\right) = \exp\left(b_0 + b_1 t\right)$ at the t-th iteration, with b_0, b_1 chosen so that adaptation occurs often at the beginning of the chain but decreases in frequency exponentially fast after burn-in. We fixed b_0 and b_1 at -0.1 and -5×10^{-5} respectively. We generate a sequence u_t of uniform random numbers between 0 and 1. If $u_t \leq \mathrm{p}\left(t\right)$, we monitor the columns in the loadings matrix having 75% of elements less than 10^{-1} in magnitude. If the number of such columns drops to zero, an additional loadings column is added by simulating from the prior distribution. Otherwise redundant columns are discarded and parameters corresponding to non-redundant columns are retained. Other parameters are also modified accordingly. Unlike the PGMM family of models and Bhattacharya & Dunson (2011), we allow $q^\star$ to shrink all the way to 0, thereby allowing a diagonal covariance structure. Letting $\tilde{q}^{(t)}$ denote the truncation level at iteration t and $q^{\star(t)} = \tilde{q}^{(t)} - m^{(t)}$ denote the effective number of factors, we use the posterior mode or median of $q^{\star(t)}$ after burn-in and thinning as an estimate of $q^\star$ with credible intervals quantifying uncertainty. Thus a histogram approximation to the posterior distribution of $q^\star$ is introduced and may be used to address the question about the effective number of latent factors.

5 Extension to Mixture Modelling

5.1 Introducing Mixture Models

Marginally, 2.2 provides a parsimonious covariance matrix, i.e. $\underline{x}_i \,|\, \theta \sim \text{MVN}_p\left(\underline{\mu}, \Lambda\Lambda^T + \Psi\right)$. This allows us to exploit model-based clustering capabilities in high dimensional data settings. We can employ a(n) (in)finite mixture of factor analyis models whereby each of the G clusters is modelled using a cluster specific latent Gaussian model with covariance specified according to the form above. Let's now introduce some basic notation at this stage:

$$\begin{aligned} N &= \sum_{g=1}^{G} n_g && \text{where } n_g \text{ is the size of the } g\text{-th group.} \\ \text{P}\left(X \,|\, \gamma\right) &= \sum_{g=1}^{G} \pi_g \text{P}_g\left(X \,|\, \theta_g\right) && \text{where } \gamma = \left(\theta_1, \ldots, \theta_G, \pi_1, \ldots, \pi_G\right), \\ & && \text{and the p.d.f } \text{P}_g \text{ is parametrized by } \theta_g. \end{aligned} \tag{5.1}$$

The *cluster mixing proportions* - $\pi_1, \ldots, \pi_G$ - have the following properties

$$\begin{aligned} \pi_g &\geq 0 \quad \forall\ g = 1, \ldots, G \\ \sum_{g=1}^{G} \pi_g &= 1 \end{aligned}$$

Introduce an additional latent indicator G-vector of *cluster labels* – $\underline{z}_i$ – s.t.

$$z_{ig} = \begin{cases} 1 \text{ if } i \in g \\ 0 \text{ otherwise} \end{cases}$$

Therefore, if $G = 3$, for instance, and observation i belongs to cluster 2, $\underline{z}_i = (0, 1, 0)$. Hence,

$$\begin{aligned} \underline{x}_i \,|\, z_{ig} = 1 &\sim \text{MVN}_p\left(\underline{\mu}_g, \Lambda_g\Lambda_g^T + \Psi_g\right) \\ \therefore \text{P}\left(\underline{x}_i\right) &= \sum_{g=1}^{G} \pi_g \text{MVN}_p\left(\underline{\mu}_g, \Lambda_g\Lambda_g^T + \Psi_g\right) \end{aligned} \tag{5.2}$$

5.1.1 Decomposable Prior for γ

The posterior distribution of γ is

$$\begin{aligned} \text{P}\left(\gamma \,|\, X\right) &\propto \text{P}\left(\gamma\right) \prod_{i=1}^{N} \text{P}\left(\underline{x}_i \,|\, \gamma\right) \\ &\propto \text{P}\left(\gamma\right) \prod_{i=1}^{N} \left(\sum_{g=1}^{G} \pi_g \text{P}_g\left(\underline{x}_i \,|\, \theta_g\right) \right) \\ \therefore \text{P}\left(\gamma \,|\, X, Z\right) &\propto \text{P}\left(\gamma\right) \prod_{g=1}^{G} \prod_{i:z_{ig}=1} \text{P}_g\left(\underline{x}_i \,|\, \theta_g\right) \end{aligned}$$

If $\text{P}\left(\gamma\right)$ can be decomposed into

$$\begin{aligned} \text{P}\left(\gamma\right) &= \text{P}\left(\pi\right) \prod_{g=1}^{G} \text{P}\left(\theta_g\right), \text{ then} \\ \text{P}\left(\gamma \,|\, X, Z\right) &\propto \text{P}\left(\pi\right) \prod_{g=1}^{G} \prod_{i:z_{ig}=1} \text{P}\left(\theta_g\right) \text{P}_g\left(\underline{x}_i \,|\, \theta_g\right) \end{aligned} \tag{5.3}$$

5.2 Deriving Posterior Distributions

Attention now turns towards deriving full conditional distributions for the new parameter $\underline{\pi}$, as well as the latent variables Z, so that we can sample them for clustering purposes, by incorporating them into the Adaptive Gibbs Sampler framework described variously above.

- Component Parameters – θ_g:
  $\mathrm{P}\left(\theta_g \,|\, \theta_{-g}, X, Z\right) \equiv \mathrm{P}\left(\theta_g \,|\, X, Z\right) \propto \prod_{i:z_{ig}=1} \mathrm{P}\left(\theta_g\right) \mathrm{P}_g\left(\underline{x}_i \,|\, \theta_g\right)$
  where $\theta_{-g} = \left(\theta_1, \ldots, \theta_{g-1}, \theta_{g+1}, \ldots, \theta_G\right)$

- Cluster Mixing Proportions – $\underline{\pi}$:
  $\mathrm{P}\left(\underline{\pi} \,|\, X, Z\right) \equiv \mathrm{P}\left(\underline{\pi} \,|\, Z\right) \propto \mathrm{P}\left(\underline{\pi}\right) \prod_{g=1}^{G} \pi_g^{n_g}$
  where n_g is the number of observations in group g,
  since $\mathrm{P}\left(\underline{z}_i \,|\, \underline{\pi}\right) \sim \mathrm{Mult}\left(1, \underline{\pi}\right)$

- Latent Variables – $\underline{z}_i$:
  $\mathrm{P}\left(\underline{z}_i \,|\, \underline{x}_i, \gamma\right) \propto \mathrm{P}\left(\underline{z}_i\right) \mathrm{P}\left(\underline{x}_i \,|\, \theta_{i:z_{ig}=1}, \underline{z}_i\right)$

5.2.1 Cluster Mixing Proportions – $\underline{\pi}$

Let the prior distribution of $\underline{\pi}$ be Dirichlet with parameter $\underline{\alpha}$ – a multivariate generalisation of the Beta distribution. Typically an exchangeable symmetric uniform prior is chosen, whereby $\alpha_g = 1 \;\forall\; g = 1, \ldots, G$.

$$\begin{aligned}
\mathrm{P}\left(\underline{\pi}\right) &\propto \prod_{g=1}^{G} \pi_g^{\alpha_g - 1} \\
\therefore \mathrm{P}\left(\underline{\pi} \,|\, Z, X\right) &\propto \prod_{g=1}^{G} \pi_g^{\alpha_g - 1} \prod_{g=1}^{G} \pi_g^{n_g} \\
&\propto \prod_{g=1}^{G} \pi_g^{\alpha_g + n_g - 1} \\
\text{i.e. } \mathrm{P}\left(\underline{\pi} \,|\, Z, X\right) &\sim \mathrm{Dir}\left(\underline{\alpha} + \underline{n}\right) \qquad (5.4) \\
\text{where } \underline{n} &= \left(n_1, \ldots, n_G\right)
\end{aligned}$$

5.2.2 Latent Variables – $\underline{z}_i$

$\underline{z}_i \,|\, \underline{x}_i, \gamma \sim \mathrm{Mult}\left(1, \underline{p}\right)$, where
$\underline{p} = \left(p_1, \ldots, p_G\right)$, and

$$\begin{aligned}
p_g = \; \mathrm{P}\left(z_{ig} = 1 \,|\, \underline{x}_i, \gamma\right) &= \frac{\pi_g \mathrm{P}\left(\underline{x}_i \,|\, \theta_g\right)}{\sum_{g=1}^{G} \pi_g \mathrm{P}\left(\underline{x}_i \,|\, \theta_g\right)} = \frac{\pi_g f\left(\underline{x}_i \,|\, \underline{\mu}_g, \Lambda_g \Lambda_g^T + \Psi_g\right)}{\sum_{g=1}^{G} \pi_g f\left(\underline{x}_i \,|\, \underline{\mu}_g, \Lambda_g \Lambda_g^T + \Psi_g\right)} \qquad (5.5) \\
&= \exp\left[\log\left(\pi_g\right) + \log\left(f(\underline{x}_i \,|\, \underline{\mu}_g, \Lambda_g \Lambda_g^T + \Psi_g)\right) - \sum_{g=1}^{G} \left(\log\left(\pi_g\right) + \log\left(f(\underline{x}_i \,|\, \underline{\mu}_g, \Lambda_g \Lambda_g^T + \Psi_g)\right)\right]
\end{aligned}$$

5.2.3 Mixtures of Infinite Factor Analyzers Pseudo-Code

MIFA has the dual advantages of allowing different clusters to be modelled my different numbers of latent factors, and significantly reducing the model search to one for G only, as q_g is estimated during model fitting.

1. Choose hyperparameters as before and initialise cluster labels $Z^{(0)}$: by simulation from the $\text{Mult}(1, \underline{\pi})$ prior, or employ other clustering algorithms, such as K-Means or Mclust. Compute $\underline{n}$, and $\tilde{\underline{\mu}}_g$ for each group.

2. Initialise, $\forall\ g = 1, \ldots, G$:

$$
\begin{aligned}
\underline{\mu}_g^{(0)} &\sim \text{MVN}_p\left(\tilde{\underline{\mu}}_g, \Sigma_\mu\right) \\
\underline{\eta}_i^{(0)} &\sim \text{MVN}_{q_g^\star}\left(\underline{0}, \mathcal{I}_{q_g^\star}\right) && \forall\ i = 1, \ldots, N \\
\underline{\Lambda}_{jg}^{(0)} &\sim \text{MVN}_{q_g^\star}\left(\underline{0}, \Sigma_\lambda\right) && \forall\ j = 1, \ldots, p \\
\psi_{jg}^{-1^{(0)}} &\sim \text{Ga}\left(\alpha, \beta_j\right) && \forall\ j = 1, \ldots, p \\
\phi_{jkg}^{(0)} &\sim \text{Ga}\left(\nu + 1, \nu\right) && \forall\ j = 1, \ldots, p \quad \text{and} \quad k = 1, \ldots, q_g^\star \\
\delta_{1g}^{(0)} &\sim \text{Ga}\left(\alpha_1, \beta_1\right), \quad \delta_{hg}^{(0)} \sim \text{Ga}\left(\alpha_2, \beta_2\right), && h \geq 2 \\
\tau_{kg}^{(0)} &= \prod_{h=1}^{k} \delta_{hg}^{(0)} && \forall\ k = 1, \ldots, q_g^\star
\end{aligned}
$$

3. For $g = 1, \ldots, G$, sample other parameters as before, but this time from their *group specific* full conditionals:

a)
$$
\begin{aligned}
\Omega_{\mu_g}^{(t)} &= \Sigma_\mu^{-1} + n_g \Psi_g^{-1^{(t-1)}} \\
\underline{\mu}_g^{(t)} &\sim \text{MVN}_p\left(\Omega_{\mu_g}^{-1^{(t)}}\left(\Psi_g^{-1^{(t-1)}}\left(\sum_{i:z_{ig}=1} \underline{x}_i - \sum_{i:z_{ig}=1} \Lambda_g^{(t-1)} \underline{\eta}_i^{(t-1)}\right) + \Sigma_\mu^{-1} \tilde{\underline{\mu}}_g\right), \Omega_{\mu_g}^{-1^{(t)}}\right)
\end{aligned}
$$

b)
$$
\begin{aligned}
\Omega_{\eta_g}^{(t)} &= \mathcal{I}_{q_g^\star} + \Lambda_g^{T^{(t-1)}} \Psi_g^{-1^{(t-1)}} \Lambda_g^{(t-1)} \\
\underline{\eta}_{i:z_{ig}=1}^{(t)} &\sim \text{MVN}_q\left(\Omega_{\eta_g}^{-1^{(t)}} \Lambda_g^{T^{(t-1)}} \Psi_g^{-1^{(t-1)}} \left(\underline{x}_{i:z_{ig}=1} - \underline{\mu}_g^{(t)}\right), \Omega_{\eta_g}^{-1^{(t)}}\right)
\end{aligned}
$$

c) For $j = 1, \ldots, p$

- $$
\begin{aligned}
\Omega_{\lambda_{jg}}^{(t)} &= \text{D}_j^{-1} + \psi_{jg}^{-1^{(t-1)}} \eta_{i:z_{ig}=1}^{T^{(t)}} \eta_{i:z_{ig}=1}^{(t)} \\
\underline{\Lambda}_{jg}^{(t)} &\sim \text{MVN}_{q_g^\star}\left(\Omega_{\lambda_{jg}}^{-1^{(t)}} \eta_{i:z_{ig}=1}^{T^{(t)}} \psi_{jg}^{-1^{(t-1)}} \left(\underline{x}_{i:z_{ig}=1}^{j} - \underline{1}\mu_{jg}^{(t)}\right), \Omega_{\lambda_{jg}}^{-1^{(t)}}\right)
\end{aligned}
$$

- $$\psi_{jg}^{-1^{(t)}} \sim \text{Ga}\left(\alpha + \frac{n_g}{2}, \beta_j + \frac{S_{jg}^{(t)}}{2}\right)$$

- $$\phi_{jkg}^{(t)} \sim \text{Ga}\left(\nu + \frac{3}{2}, \nu + \frac{\tau_{kg}^{(t-1)} \lambda_{jkg}^{2^{(t)}}}{2}\right) \quad \forall\ k = 1, \ldots, q_g^\star$$

d)
$$
\begin{aligned}
\delta_{1g}^{(t)} &\sim \text{Ga}\left(\alpha_1 + \frac{p q_g^\star}{2}, \beta_1 + \frac{1}{2} \sum_{h=1}^{q_g^\star} \tau_{hg}^{(1)^{(t-1)}} \sum_{j=1}^{p} \lambda_{jhg}^{2^{(t)}} \phi_{jhg}^{(t)}\right) \\
\delta_{hg}^{(t)} &\sim \text{Ga}\left(\alpha_2 + \frac{p}{2}\left(q_g^\star - k + 1\right), \beta_2 + \frac{1}{2} \sum_{h=k}^{q_g^\star} \tau_{hg}^{(k)^{(t)}} \sum_{j=1}^{p} \lambda_{jhg}^{2^{(t-1)}} \phi_{jhg}^{(t)}\right), \quad h \geq 2 \\
\tau_{kg}^{(t)} &= \prod_{h=1}^{k} \delta_{hg}^{(t)} \qquad \forall\ k = 1, \ldots, q_g^\star
\end{aligned}
$$

4. Re-compute $\underline{n}$, sample $\underline{\pi}$ from $\mathrm{Dir}\,(\underline{\alpha} + \underline{n})$ and sample $\underline{z}_i$ as outlined in (5.5).

5. Follow the adaptation procedure outlined in 4.3.

6. Repeat steps 4–7 for $t = 2, \ldots, T$ using the current value for $q_g^\star$.

7. Disregard the first B burn-in iterations and thin every K-th iteration [2].

5.3 Label Switching

It's easy to see that $\mathrm{P}\,(X \,|\, \gamma) = \mathrm{P}\,(X \,|\, \tilde{\gamma})$ where $\tilde{\gamma} = (\theta_{j_1}, \ldots, \theta_{j_G}, \pi_{j_1}, \ldots, \pi_{j_G})$ and $j_1, \ldots, j_G$ is any permutation of $1, \ldots, G$. This type of finite mixture distribution nonidentifiability is caused by the invariance of mixture distributions to component relabelling: by interchanging the order of components, the distributions induced by γ and $\tilde{\gamma}$ are the same, although evidently the two parameters are distinct. For finite mixture distribution as defined above with G components, there exist $G!$ equivalent ways of arranging them. Generally as the Markov chain progresses, we will observe switches between these equivalent modes. When the main goal is identifying/interpreting mixture components &/or clustering, this *label switching* phenomenon needs to be addressed. The approach we adopt to do so is applied post-hoc, after the chain has finished running, and has the advantage of not involving loss functions based on sampled model parameters. We only require samples of Z, which are matched to a template vector of cluster labels at burnin using the cost-minimizing permutation suggested by the square assignment algorithm of Carpaneto & Toth (1980). This same permutation is applied to all other parameters which vary by group, namely the means, loadings, uniquenesses, and mixing proportions, prior to computing their posterior mean estimates.

5.4 Overfitted Mixtures

The need to choose the optimal number of latent factors in a mixture of factor analysers has been obviated using MIFA, but the issue of model choice is still not entirely resolved. Overfitted mixtures, along with Dirichlet processes (6) and transdimensional MCMC, are a means of extending the MIFA method in order to estimate G in a similarly choice-free manner. The prior in 5.2.1 plays an important role. Mixture model estimation is approached by initially overfitting the number of clusters expected to be present, and specified conditions on the Dirichlet hyperparameter for the mixing proportions encourage the emptying out of excess components in the posterior distribution.

To initialise the method, a conservatively high number of groups $G^\star = \max\left(\lfloor 3\ln(N) \rfloor, 25, N-1\right)$ is chosen, and fixed for the entire length of the MCMC chain. It's assumed that $G^\star > G$. Each $\alpha_g = 0.5/G^\star$ is set small enough to favour empty groups a priori Ishwaran et al. (2001). The symmetric uniform prior $\mathrm{Dir}\,(1, \ldots, 1)$ used previously is rather indifferent in this respect. The number of non-empty groups at each iteration G_0 is recorded thusly:

$$G_0 = G^\star - \sum_{g=1}^{G} \mathbb{1}\Big(\sum_i z_{ig} = 0\Big) \tag{5.6}$$

The true G is estimated by the G_0 value visited most often by the sampler. Component specific inference is conducted only on the M_0 samples corresponding to those visits. However, this method is not ideal in the sense that there is a conflict of opinion on 'how small' α needs to be: too large and no/few clusters will be emptied, too large and the estimate will shrink close to the true G, but mixing proportions will become so small that new clusters will rarely be formed, and. Furthermore, the sampler needs to carry round and simulate empty groups from priors, with the adaptation scheme modified to exclude empty groups about which we have no information.

[2] With MFA, one chooses between competing models according to the pair of G and q values which optimise any of the model selection criteria outlined in 3.5. With the MIFA approach, one chooses G using BICM or AICM only.

6 Dirichlet Process Mixture Models

Traditional mixtures of factor analyzers using a fixed and finite number of components and factors can suffer from over and under-fitting, where there is a misfit between the complexity of the model and the amount of data available. Model selection is a difficult issue and it involves the double and interdependent choice of number of factors and number of mixture components. The proposed IMIFA model represents an alternative to parametric modeling by allowing theoretically infinite factors and mixture components. IMIFA is a similarly Bayesian nonparametric extension designed to address the issue of determining the number of clusters, G, by allowing infinitely many of them through a Dirichlet process prior, thereby yielding a choice-free approach which obviates the need for model selection criteria. In so doing, it represents also a way to sidestep the fraught task of determining the number of model parameters, bringing significant flexibility. It must be noted that, even if theoretically the number of mixture components is infinite under a Dirichlet process prior, practically they are at most equal to the sample size N, if we ignore empty components, and the growth rate of $\mathbb{E}(G)$ is known to be logarthmic in N. In the same way the infinite factor model detailed above can retain a number of effective factors at most equal to the number of variables P in practice, by ignoring redundant factors.

6.1 Dirichlet Processes

Dirichlet processes are stochastic processes whose draws are random probability measures. A probability distribution H is a Dirichlet process with parameters H_0, the *base distribution*, and α, the *concentration parameter*, denoted $H \sim \mathrm{DP}(\alpha, H_0)$, if every marginal of H on finite partitions of the domain Ω are Dirichlet distributed, as per the following definition due to Ferguson (1973):

$$\begin{aligned}(H(A_1), \ldots, H(A_r)) &\sim \mathrm{Dir}(\alpha H_0(A_1), \ldots, \alpha H_0(A_r)) \\ A_1 \cup \ldots \cup A_r &= \Omega\end{aligned} \tag{6.1}$$

The base distribution H_0 can be interpreted as the prior guess for the parameters of the model or the mean of the DP:

$$\mathbb{E}[H(A)] = H_0(A) \tag{6.2}$$

The concentration parameter α expresses the strength of belief in H_0:

$$\mathbb{V}[H(A)] = \frac{H_0(A)(1 - H_0(A))}{\alpha + 1} \tag{6.3}$$

The choice of the base distribution for the model parameters is important for the model performance. In our implementation, the base distribution comes from the factor analytic structure given variously above. This allows us define the general model $f(\underline{x}_i) = \sum_{g=1}^{\infty} \pi_g \mathrm{MVN}_p\left(\underline{\mu}_g, \Lambda_g\Lambda_g^\top + \Psi_g\right)$. We consider conjugate prior distributions for the model parameters, with additional layers for the hyperparameters, as specified above.

There exist several equivalent metaphors which motivate methods of yielding samples from a DP without representing the infinite dimensional variable G explicitly, which make its key properties more clear, e.g. Chinese Restaurant Process, Pólya-Urn Scheme, and Stick-Breaking Representation. We focus on the latter. Furthermore, MCMC sampling strategies can be divided into two families; marginal methods, which integrate out the infinite dimensional probability measure H and directly represent the partition structure of the data e.g. Escobar (1994), Escobar & West (1995), and Neal (2000), as well as conditional methods, which sample a sufficient but finite number of groups at each iteration, e.g. truncation Ishwaran et al. (2001), retrospective sampling Papaspiliopoulos & Roberts (2008), and slice sampling Walker (2007); Kalli et al. (2011). We adopt a slice sampling approach.

6.2 Stick-Breaking Construction

The stick-breaking construction of Sethuraman (1994) metaphorically views the mixing proportions $\{\pi_1, \pi_2, \ldots\}$ as pieces of a unit-length stick that is sequentially broken in an infinite process, with stick-breaking proportions $\{V_1, V_2, \ldots\}$ according to realisations of a Beta distribution, and can be summarised as follows:

$$\begin{aligned} V_g &\sim \text{Beta}(1, \alpha) & \theta_g &\sim H_0 \\ \pi_g &= V_g \prod_{l=1}^{g-1}(1 - V_l) & H &= \sum_{g=1}^{\infty} \pi_g \delta_{\theta_g} \sim \text{DP}\left(\alpha, H_0\right) \end{aligned} \tag{6.4}$$

where δ_θ is the Dirac delta centered at θ and $\theta = \left\{\underline{\mu}, \Lambda, \Psi\right\}$ denotes the whole set of parameters of the cluster-specific FA models, s.t. draws are composed of a sum of infinitely many point masses. We use this as a prior process for generating the weights of the infinite mixture distribution.

6.3 Slice Sampling

The slice sampler of Walker (2007) introduces an auxiliary variable $u > 0$ which preserves the marginal distribution of the data x and facilitates writing the conditional density of $x \,|\, u$ as a finite mixture, by letting $\underline{\xi} = \{\xi_1, \xi_2, \ldots\}$ be a decreasing sequence of infinite quantities which sum to 1, s.t. the joint distribution of (x, u) is

$$f(x, u \,|\, \theta, \underline{\xi}) = \sum_{g=1}^{\infty} \pi_g \text{Unif}(u; 0, \xi_g) f(x; \theta_g) \tag{6.5}$$

with

$$f(x; \theta) = \int f(x, u) du \quad = \sum_{g=1}^{\infty} \pi_g f(x; \theta_g) \int \text{Unif}(u; 0, \xi_g) du = \sum_{g=1}^{\infty} \pi_g f(x; \theta_g) \tag{6.6}$$

and

$$f(u; \underline{\xi}) = \int f(x, u; \theta, \underline{\xi}) = \sum_{g=1}^{\infty} \pi_g \text{Unif}(u; 0, \xi_g) \int f(x; \theta_g) dx = \sum_{g=1}^{\infty} \pi_g \text{Unif}(u; 0, \xi_g) = \sum_{g=1}^{\infty} \frac{\pi_g}{\xi_g} \mathbb{1}(u < \xi_g) \tag{6.7}$$

Since only a finite number of ξ_g are greater than u, by denoting $\mathcal{A}_\xi(u) = \{g : u < \xi_g\}$, the conditional density of $x \,|\, u$ can be written as a *finite* mixture model:

$$f\left(x \,|\, u; \theta\right) = \frac{f\left(x, u; \theta, \underline{\xi}\right)}{f\left(u; \underline{\xi}\right)} \quad = \sum_{g \in \mathcal{A}_\xi(u)} \frac{\pi_g}{\xi_g f\left(u; \underline{\xi}\right)} f\left(x; \theta_g\right) \tag{6.8}$$

because

$$\text{Unif}\left(u; 0, \xi_g\right) = \frac{1}{\xi_g} \mathbb{1}(u < \xi_g)$$

$$\text{Now } f\left(\underline{x}_i\right) = \sum_{g=1}^{\infty} \pi_g \text{MVN}_p\left(\underline{\mu}_g, \Lambda_g \Lambda_g^\top + \Psi_g\right) \text{ can be sampled from.}$$

Typically $\xi_g = \pi_g$, but 'independent' slice-efficient sampling Kalli et al. (2011) allows for a deterministic decreasing sequence, e.g. geometric decay $\xi_g = (1 - \rho)\, \rho^{g-1}$,where ρ is a fixed value $\in (0, 1)$. The ρ parameter must be chosen delicately; higher values will be lead to better mixing, but longer running times, since the size of the set $\mathcal{A}_\xi(u)$ increases. We find $\rho = 0.75$ strikes an appropriate balance. For slice sampling, components and corresponding parameters are reordered at each iteration such that the mixing proportions form a decreasing sequence.

6.4 Infinite Mixtures of Infinite Factor Analysers

We impose a hierarchical structure on the latent variables and parameters which utilizes the auxiliary variable of the independent slice efficient sampler, with geometric decay values. The joint distribution of the factor-analytic mixture-model with infinite components is proportional to:

$$\begin{aligned} f\left(x,\eta,z,\underline{u},\underline{V},\theta\right) &\propto f\left(x\,|\,\eta,z,\underline{u},\underline{V},\theta\right) f\left(\eta\,|\,\underline{u}\right) f\left(z,\underline{u}\,|\,\underline{V},\underline{\pi}\right) f\left(\underline{V}\,|\,\alpha\right) f\left(\theta\right) \\ &\propto \left\{\prod_{i=1}^{N}\prod_{g\in\mathcal{A}_\xi(u_i)} \mathrm{MVN}_p\left(x_i;\underline{\mu}_g+\Lambda_g\underline{\eta}_i,\Psi_g\right)^{z_{ig}}\right\} \\ &\quad \left\{\prod_{i=1}^{N}\prod_{g\in\mathcal{A}_\xi(u_i)} \mathrm{MVN}_q\left(\underline{\eta}_i;0,\mathcal{I}_q\right)\right\} \\ &\quad \left\{\prod_{i=1}^{N}\prod_{g=1}^{\infty}\left(\frac{\pi_g}{\xi_g}\mathbb{1}(u_i<\xi_g)\right)^{z_{ig}}\right\}\left\{\prod_{g=1}^{\infty}\frac{(1-V_g)^{\alpha-1}}{\mathrm{Beta}\,(1,\alpha)}\right\} f(\theta) \end{aligned} \tag{6.9}$$

where $f(\theta)$ is the full collection of conjugate priors defined previously, under the trick that only the 'active components' such that $g\in\mathcal{A}_\xi\left(u_i\right)$ have to be estimated. Thus the number of active components which have to be estimated at each iteration is given by $G=\max_{1\leq i\leq N}|\mathcal{A}_\xi\left(u_i\right)|$, where $|\mathcal{A}_\xi\left(u_i\right)|$ is the cardinality of $\mathcal{A}_\xi\left(u_i\right)$. This discrete number varies but is fixed at each iteration, even if theoretically infinite. However, it's the non-empty subset rather than active clusters that are of interest. As with the OMIFA model 5.4, the algorithm is initialised with a conservatively high starting value for the number of groups, above the truth in the spirit of Hastie et al. (2014), and the true G is estimated by the number of non-empty clusters visited most often, with cluster-specific inference is conducted only on samples corresponding to those visits. All conditional posteriors have standard form and the IMIFA algorithm can proceed via efficient Gibbs updates.

6.5 IMIFA Full Conditionals

From (6.9)

$$\underline{V}\,|\,-\propto\prod_{g=1}^{\infty}\pi_g^{n_g}(1-V_g)^{\alpha-1}=\prod_{g=1}^{\infty}V_g^{n_g}\left(\prod_{l<g}(1-V_l)\right)^{n_g}(1-V_g)^{\alpha-1}$$

The previous expression can be expanded as:

$$\begin{aligned} \underline{V}\,|\,-\propto\ & V_1^{n_1}(1-V_1)^{\alpha-1}V_2^{n_2}(1-V_1)^{n_2}(1-V_2)^{\alpha-1}V_3^{n_3}(1-V_1)^{n_3}(1-V_2)^{n_2}(1-V_3)^{\alpha-1} \\ & V_4^{n_4}(1-V_1)^{n_4}(1-V_2)^{n_4}(1-V_3)^{n_4}(1-V_4)^{\alpha-1}\ldots \\ =\ & V_1^{n_1}(1-V_1)^{\alpha-1}(1-V_1)^{N-n_1}V_2^{n_2}(1-V_2)^{\alpha-1}(1-V_2)^{N-n_1-n_2}\ldots \end{aligned}$$

from which for V_g

$$V_g\,|\,-\sim\mathrm{Beta}\left(1+n_g,\alpha+N-\sum_{l=1}^{g}n_l\right) \tag{6.10}$$

To derive a full conditional for the auxiliary variable u_i, observe that $f\left(z,\underline{u}\,|\,\underline{V},\underline{\pi}\right)$ in (6.9) can be rewritten as $f\left(\underline{z}_i,u_i\,|\,\underline{V},\underline{\pi}\right)=f\left(\underline{z}_i,u_i\,|\,\underline{\pi}\right)=\prod_{g=1}^{\infty}\left(\pi_g\mathrm{Unif}\left(u;0,\xi_g\right)\right)^{z_{ig}}$ because the weights $\underline{\pi}$ are deterministic functions of $\underline{V}$. We thus derive the full conditional for u_i as:

$$u_i\,|\,z_{ig}=1\sim\mathrm{Unif}\left(0,\xi_g\right) \tag{6.11}$$

From (5.5) and (6.9), it is straightforward to show that

$$z_{ig} = 1 \,|\, — \propto \pi_g \mathrm{P}\left(\underline{x}_i \,|\, \theta_g\right) \frac{\pi_g}{\xi_g} \mathbb{1}(u_i < \xi_g) = \pi_g f\left(\underline{x}_i \,|\, \underline{\mu}_g, \Lambda_g \Lambda_g^T + \Psi_g\right) \frac{\pi_g}{\xi_g} \mathbb{1}(u_i < \xi_g) \tag{6.12}$$

Though it remains fixed in many applications, we also add a $\mathrm{Ga}\,(a, b)$ prior for the concentration parameter α, according to the auxiliary variable routine of West (1992) reproduced below, so that it can be learned from the data.

$$\mathrm{P}\left(G \,|\, \alpha, N\right) \propto \alpha^{G-1} \frac{(\alpha + N)\, \beta\,(\alpha + 1, N)}{\Gamma\,(N)} \qquad (G = 1, \ldots, N)$$

$$\therefore \alpha \,|\, G, \chi \sim \omega_\chi \mathrm{Ga}\,(a + G, b - \log(\chi)) + (1 - \omega_\chi)\,\mathrm{Ga}\,(a + G - 1, b - \log(\chi)) \tag{6.13}$$

with weights defined by: $\dfrac{\omega_\chi}{1 - \omega_\chi} = \dfrac{(a + G - 1)}{N\,(b - \log(\chi))}$,
where β is the Beta function and $(\chi \,|\, \alpha, G) \sim \mathrm{Beta}\,(\alpha + 1, N)$.

Finally, state spaces for real data IMIFA applications can be highly multimodal, with well separated regions of high posterior probability coexisting, corresponding to clusterings of different sizes, we incorporate the label switching moves suggested by Papaspiliopoulos & Roberts (2008). These are complimentary moves which are effective at swapping similar and unequal clusters, respectively. Parameters are reordered accordingly after each accepted move.

- Change labels of two randomly chosen non-empty clusters g and h with probability: $\min\left\{1, (\pi_h/\pi_g)^{n_g - n_h}\right\}$
- Change labels of neighbouring non-empty clusters g and $g+1$ with probability: $\min\left\{1, {}^{(1-V_g)^{n_g}}/_{(1-V_{g+1})^{n_{g+1}}}\right\}$

7 Results

7.1 Simulation Study

A simulation study with the following design was conducted in order to assess the performance of the IMIFA model: data with $G = 3$ groups and $P = 50$ variables was simulated according to the factor analytic structure, with $q_g = 4$ latent factors in each group. This was done with $N = 25 (N \ll P)$, $N = 50 (N = P)$, and $N = 300 (N \gg P)$ in order to evaluate model performance in different dimensionality scenarios. The groups are roughly balanced. Sensitivity to the concentration parameter α is examined by running the model at values of 0.5, 1, and 5, and by allowing α to be learned according to (6.12). In order to reduce error due to simulation, the results below are based on repeating the study over ten replicate datasets meeting the same criteria. With 12,500 iterations, of which the first 2,500 are discarded due to burn-in and every second sample was thinned, this gives 50,000 samples contributing to each row of the aggregated table below.

Dimension	α	G	q_1	q_2	q_3	Time (s)	Error (%)
	0.5	3 [3,3]	4 [2,8]	4 [2,8]	4 [2,8]	327.03	0
N = 25	1	3 [3,3]	4 [2,8]	4 [2,8]	4 [2,8]	328.11	0
(N $\ll$ p)	5	3 [3,4]	4 [2,8]	4 [2,8]	4 [2,8]	330.50	0
	Gibbs	3 [3,3]	4 [2,8]	3 [2,7]	4 [2,8]	327.56	0
	0.5	3 [3,3]	5 [3,6]	5 [3,6]	5 [3,7]	361.66	0
N = 50	1	3 [3,3]	5 [3,7]	5 [3,7]	5 [3,7]	363.06	0
(N = p)	5	3 [3,3]	5 [3,6]	5 [3,6]	4 [3,7]	367.44	0
	Gibbs	3 [3,3]	5 [3,6]	5 [3,6]	5 [3,7]	366.57	0
	0.5	3 [3,3]	5 [4,6]	5 [4,6]	5 [4,6]	555.82	0
N = 300	1	3 [3,3]	5 [4,6]	5 [4,6]	5 [4,6]	558.52	0
(N $\gg$ p)	5	3 [3,3]	5 [4,6]	5 [4,6]	5 [4,6]	560.75	0
	Gibbs	3 [3,3]	5 [4,6]	5 [4,6]	5 [4,6]	560.46	0

This demonstrates good performance in the sense that the modal estimate of G uncovered the truth in all cases, with only the scenario $N \ll P$ and α fixed too large showing some deviation in the 95% credible interval, given in brackets. Furthermore, estimates of q_g are within the limits of the credible intervals in every case also. Lastly, clustering performance is perfect. Overall, the IMIFA model demonstrates excellent ability to uncover the truth of simulated data.

7.2 Olive Oil Benchmark

Further assessment of IMIFA's clustering performance was conducted using data due to Forina et al. (1983) on the percentage composition of 8 fatty acids found by lipid fraction of 572 Italian olive oils, from three areas: Southern Italy, Sardinia, and Northern Italy. Within each area there are a number of different regions. Southern Italy comprises North Apulia, Calabria, South Apulia, and Sicily. Sardinia is divided into Inland Sardinia and Coastal Sardinia. Northern Italy comprises Umbria, East Liguria, and West Liguria. As such the true number of groups is hypothesised to correspond to either 3 *areas* or 9 *regions*. The full suite of models, from (Infinite) Factor Analysis through Mixtures of (Infinite) Factor Analysers, to (Overfitted/Infinite) Mixtures of (Infinite) Factor Analysers, were run on the data for 50,000 iterations, with the first 10,000 discarded and every 2^{nd} sample thereafter thinned. Results for methods that rely on ranges of G and/or q values being supplied are based on $G = 1, \ldots, 9$ and $q = 0, \ldots, 6$, respectively, with the optimal model chosen by BICM where necessary. The results are summarised in the table below, with clusterings evaluated using the Adjusted Rand Index, and the percentage error rate, compared to the known *area* labels. Though the α parameter plays different roles in each of these models, it is given as its fixed value or posterior mean, as appropriate. The IMIFA algorithm was used as the baseline for the relative time.

Method	**# Models**	**Time (s)**	**Rel. Time**	$\boldsymbol{\alpha}$	**G**	**q**	**Adj. Rand**	**Error (%)**
IMIFA	1	1814.59	1.000	4.54	4	6, 2, 3, 2	0.9345	8.57
IMFA	7	13152.41	7.248	4.79	6	5, 5, 5, 5, 5, 5	0.5164	37.24
OMIFA	1	1860.42	1.025	0.025	6	6, 2, 2, 2, 2, 2	0.9037	15.91
OMFA	7	9547.15	5.261	0.025	6	5, 5, 5, 5, 5, 5	0.5153	37.41
MIFA	9	6848.13	3.774	1	1	6	0	43.53
MFA	63	33636.22	18.537	1	2	5, 5	0.8192	17.13
IFA	1	240.82	0.133	–	1	6	–	–
FA	7	1092.95	0.602	–	1	6	–	–

Of the 572 observations, 323 originate from Southern Italy. This large cluster requires the largest amount of factors. It's clear that the flexibilty to model remaining clusters by different, in this case lower, numbers of factors greatly improves the clustering performance. Indeed, the best solution is given by the IMIFA model, which requires only one run and doesn't rely on any model selection criteria. The IMIFA performance compares favourably to the PGMM solution also, which gives a 5 group, 5 factor model (Adj. Rand = 0.5586, Error Rate = 33.56). It's also worth noting that the optimal models chosen by algorithms which do rely on model selection criteria were not all optimal in a clustering sense – for instance, in this case where the truth is known, the candidate MIFA model with $G = 4$ yields an Adj. Rand of 0.9304 and an error rate of 10.49%, despite having a higher BICM. The cross tabulation of the MAP clustering from the IMIFA model against the known area labels suggests the possibility of a fourth group, whereby Umbria is separated from East and West Liguria in the North. This makes sense, geographically. The confusion matrix using this relabelling yields an Adj. Rand of 0.9961 and an error rate of 0.7%.

	1	2	3	4
South	323	0	0	0
Sardinia	0	97	1	0
North	0	0	103	48

	1	2	3	4
South	323	0	0	0
Sardinia	0	97	1	0
Liguria	0	0	100	0
Umbria	0	0	3	48

7.3 Real Data

8 Extensions

The Pitman-Yor process is a popular generalisation of the Dirichlet Process Perman et al. (1992), and is sometimes referred to as the two-parameter Poisson-Dirichlet process. This prior introduces a discount parameter, d, which lies in the interval $[0, 1)$. In order to satisfy the rules of probability, the α parameter must be strictly greater than $-d$. The PY prior reduces to the Dirichlet process prior when $d = 0$. Nonetheless, some important distributional features of the PY process are fundamentally different when this parameter is non-zero. Conveniently, this prior has its own stick-breaking construction, given by

$$V_g \sim \text{Beta}\,(1 - d, \alpha + gd) \tag{8.1}$$

which means this generalisation can be easily incorporated into our Gibbs sampling framework. In particular, non-zero discount has the effect of flattening the Dirichlet process prior and mitigating against its 'rich-get-richer' property. The PY prior allows a small number of large groups plus some small groups with the growth rate of $\mathbb{E}\,(G)$ now Zipfian rather than logarithmic in N. Though the discount parameter remains fixed in our implementation, the gamma prior on α in (6.13) is shifted to account for non-zero d values.

Below are some possible further extensions to this work:

- Rewrite code bottlenecks using **RCPP**.
- Investigate block updating the loadings matrix using a matrix normal prior Viroli (2011). This works, albeit surprisingly slower than the current implementation, for the models where q_g is fixed. It remains to reformulate the MGP prior as a matrix normal MGP prior. Bhattacharya et al. (2015) could be of some help in this regard.
- Implement the third label switching move of Hastie et al. (2014).
- Learn the shrinkage hyperparameters α_1 & α_2, which was proposed but not implemented by Bhattacharya & Dunson (2011). This requires the introduction of Metropolis Hastings steps.
- Yu (2009) gives a strategy for Collapsed Gibbs Sampling of Dirichlet Process Mixture Models.
- The present work could be extended to the supervised/semi-supervised settings where some or all of the data is labelled, in order to uses Mixtures of Infinite Factor Analysers for (semi-) supervised Model Based Classification.
- Covariates could be incorporated in the spirit of Bayesian Factor Regression Models West (2003); Wang et al. (2007); Carvalho et al. (2008); for instance, the spectral metabolomic urine data also contains information on gender, weight and age.
- The **IMIFA** package includes functions to assist in soliciting good priors for either Dirichlet or Pitman-Yor processes, by finding the expected number of groups for a given combination of α, N, and d values, as well as the variance. A third function exists to actually plot the distribution of the number of groups, but at present this is only possible in the case where $d = 0$. Extending this function to plot Pitman-Yor priors with non-zero discount requires computation of generalized Stirling numbers of the second kind.
- Furthermore, the discount parameter currently remains fixed. Some papers introduce a Beta $(1, 1)$ prior on it, but strategies for subsequently updating the parameter are unclear, at least in the papers I have read to date.

- Practically useful extensions to heavier-tailed errors using multivariate t distributions in place of the multivariate normal here are easily encompassed within the simulation based Bayesian analysis we develop and use, as per Peel & Mclachlan (2000). For many applied problems, the tails of the normal distribution are often shorter than required. Also, the estimates of the component means and covariance matrices can be affected by observations that are atypical of the components in the normal mixture model being fitted. The problem of providing protection against outliers in multivariate data is a very difficult problem and increases with the difficulty of the dimension of the data. With this t mixture model-based approach, the normal distribution for each component in the mixture is embedded in a wider class of elliptically symmetric distributions with an additional parameter called the degrees of freedom κ. The normal-scale characterisation of the mixture of MVt distributions is given as follows:

$$p_g = \mathrm{P}\left(z_{ig} = 1 \,|\, \underline{x}_i, \gamma, \kappa_g\right) = \frac{\pi_g \mathrm{P}\left(\underline{x}_i \,|\, \theta_g, \kappa_g\right)}{\sum_{g=1}^{G} \pi_g \mathrm{P}\left(\underline{x}_i \,|\, \theta_g, \kappa_g\right)} = \frac{\pi_g f\left(\underline{x}_i \,|\, \underline{\mu}_g, {}^{\left(\Lambda_g \Lambda_g^T + \Psi_g\right)}/_{U_i}\right)}{\sum_{g=1}^{G} \pi_g f\left(\underline{x}_i \,|\, \underline{\mu}_g, {}^{\left(\Lambda_g \Lambda_g^T + \Psi_g\right)}/_{U_i}\right)} \tag{8.2}$$

where

$$U_i \,|\, z_{ig} = 1 \sim \mathrm{Ga}\left(\frac{1}{2}\kappa_g, \frac{1}{2}\kappa_g\right) \tag{8.3}$$

As κ tends to infinity, the t distribution approaches normality, as each U_i will tend to one with probability one. Hence this parameter κ may be viewed as a robustness tuning parameter. It can be fixed in advance or it can be inferred from the data for each component.

9 References

A. Bhattacharya & D. B. Dunson. *Sparse Bayesian infinite factor models. Biometrika*, 98(2): 291–306, 2011. ISSN 00063444. doi: 10.1093/biomet/asr013.

A. Bhattacharya, A. Chakraborty, & B. K. Mallick. Fast sampling with Gaussian scale-mixture priors in high-dimensional regression. pages 1–11, 2015. URL `http://arxiv.org/abs/1506.04778`.

G. Carpaneto & P. Toth. *Algorithm 548 Solution of the assignment problem. ACM Transactions on Mathematical Software*, 6: 104–111, 1980.

C. M. Carvalho, J/ Chang, J. E. Lucas, J. R. Nevins, Q. Wang, & M. West. High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. *Journal of the American Statistical Association*, 103(484): 1438–1456, 2008. ISSN 0162-1459. doi: 10.1198/016214508000000869. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3017385{&}tool=pmcentrez{&}rendertype=abstract`.

D. Durante. A note on the multiplicative gamma process. 2016. URL `http://arxiv.org/abs/1610.03408`.

M. D. Escobar. Estimating normal means with a dirichlet process prior. *Journal of the American Statistical Association*, 89(425): 268–277, 1994. ISSN 01621459. URL `http://www.jstor.org/stable/2291223`.

M. D. Escobar & M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430): 577–588, 1995. ISSN 01621459. URL `http://www.jstor.org/stable/2291069`.

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1(2): 209–230, 03 1973. doi: 10.1214/aos/1176342360. URL `http://dx.doi.org/10.1214/aos/1176342360`.

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

S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer series in statistics, 2010. ISBN 9780387775005. doi: 10.1007/978-0-387-98135-2.

S. Frühwirth-Schnatter. *Dealing with label switching under model uncertainty*, pages 193–218. Mixture estimation and applications. Wiley, Chichester, 2011. ISBN ISBN-10: 11199938.

S. Frühwirth-Schnatter & H. F. Lopes. Parsimonious bayesian factor analysis when the number of factors is unknown. (July 2015): 1–37, 2010.

Z. Ghahramani & G. E. Hinton. The em algorithm for mixtures of factor analyzers. Technical report, 1996.

D. I. Hastie, S. Liverani, & S. Richardson. Sampling from Dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. *Statistics and Computing*, 25(5): 1023–1037, 2014. ISSN 09603174. doi: 10.1007/s11222-014-9471-3.

H. Ishwaran, L.F. James, & J. Sun. Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 96: 1316–1332, 2001. URL `http://EconPapers.repec.org/RePEc:bes:jnlasa:v:96:y:2001:m:december:p:1316-1332`.

M. Kalli, J. E. Griffin, & S. G. Walker. Slice sampling mixture models. *Statistics and Computing*, 21(1): 93–105, 2011. ISSN 09603174. doi: 10.1007/s11222-009-9150-y.

G. J. McLachlan & D. Peel. *Finite mixture models*. Wiley series in probability and statistics. J. Wiley & Sons, New York, 2000. ISBN 0471006262. URL `http://opac.inria.fr/record=b1097397`.

P. D. McNicholas & T. B. Murphy. Parsimonious Gaussian Mixture Models. *Statistics and Computing*, 18(3): 285–296, 2008.

R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2): 249–265, 2000. ISSN 10618600. URL `http://www.jstor.org/stable/1390653`.

O. Papaspiliopoulos & G. O. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1): 169–186, 2008. ISSN 00063444. doi: 10.1093/biomet/asm086.

D. Peel & G. J. Mclachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10: 339–348, 2000. ISSN 0960-3174. doi: http://dx.doi.org/10.1023/A:1008981510081.

M. Perman, J. Pitman, & M. Yor. Size-biased sampling of poisson point processes and excursions. *Probability Theory and Related Fields*, 92(1): 21–39, 1992. ISSN 1432-2064. doi: 10.1007/BF01205234. URL `http://dx.doi.org/10.1007/BF01205234`.

A. E. Raftery, M. Newton, P. N. Krivitsky, & J. M. Satagopan. Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity. *Bayesian Statistics*, (8): 1–45, 2007.

H. Rue & L. Held. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 2005.

J. Sethuraman. A constructive definition of Dirichlet priors, 1994. URL `http://www3.stat.sinica.edu.tw/statistica/j4n2/j4n27/..$\backslash$j4n216$\backslash$j4n216.htm`.

C. Viroli. Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*, 21(4): 511–522, 2011. ISSN 09603174. doi: 10.1007/s11222-010-9188-x.

S. G. Walker. Sampling the dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36(1): 45–54, 2007. doi: 10.1080/03610910601096262. URL http://dx.doi.org/10.1080/03610910601096262.

Q. Wang, C. M. Carvalho, J. E. Lucas, & M. West. Bfrm: Bayesian factor regression modelling. *Bulletin of the International Society for Bayesian Analysis*, 14: 4–5, 2007. URL http://www.isds.duke.edu/mw/.downloads/bfrm-isbabull07.pdf.

M. West. Hyperparameter estimation in Dirichlet process mixture models. *ISDS discussion paper series*, pages #92–03, 1992. URL http://www.stat.duke.edu/{~}mw/.downloads/DP.learnalpha.pdf.

M. West. Bayesian factor regression models in the lqlq large p, small n" paradigm. In *Bayesian Statistics 7*, pages 723–732. Oxford University Press, 2003. URL http://ftp.isds.duke.edu/WorkingPapers/02-12.html.

X. Yu. Gibbs Sampling Methods for Dirichlet Process Mixture Model: Technical Details. pages 1–18, 2009.