Infinite Mixtures of Infinite Factor Analysers

Notes & Derivations

Keefe Murphy[1, 2], Dr. Claire Gormley[1, 2], Prof. Brendan Murphy[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 Model Set-Up

Let $\underline{x} = (x_1, x_2, \ldots, x_p)^T$ have mean $\underline{\mu}$ and covariance matrix Σ. The factor model states that $\underline{x}$ is linearly dependent upon a few ($q \ll p$) unobservable random variables $\underline{\eta}_i, \underline{\eta}_i, \ldots, \underline{\eta}_q$ 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 $(p \times q)\,\Lambda$ matrix of factor loadings. If we assume the data have been centred to have column means of 0 then we have

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

1.2 Assumptions

1. $\underline{\mu} = 0$
2. $\underline{\varepsilon}_i$ and $\underline{\eta}_i$ are independent: $\text{Cov}\left(\underline{\eta}_i, \underline{\varepsilon}_i\right) = \text{E}\left(\underline{\eta}_i\underline{\varepsilon}_i^T\right) = 0$
3. $\underline{\varepsilon}_i \sim \text{MVN}_p\left(\underline{0}, \Psi\right)$ where $\Psi = diag\left(\psi_1, \ldots, \psi_p\right)$

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

   $$\therefore \underline{\varepsilon}_i \sim \text{MVN}_p\left(\underline{0}, \Psi\right) \tag{1.2}$$

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

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

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

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 - $\underline{\eta}_i$

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

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

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

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

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

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

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

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 $\underline{z} \sim \mathrm{MVN}_q\left(\underline{0}, \mathcal{I}_q\right)$ 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.

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. $\underline{\Lambda}_j \sim \mathrm{MVN}_q\left(\underline{0}, \Sigma_\lambda\right)$ where Σ_λ is a diagonal covariance matrix. As above, we can

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

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

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 $\Omega_{\lambda_j} = U^T U$.
- 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:

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

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:

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

Since Ψ is a diagonal matrix:

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

This suggests the prior for Ψ^{-1} is a product of p $\text{Ga}\left(\alpha, \beta\right)$ 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.

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

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:

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

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 covariance matrix, and $\underline{\tilde{\mu}}$ is a vector of prior mean means, typically the sample mean for each group. As above, we can expect the result of the product of two MVN_p distributions to itself be distributed this way.

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

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

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 $\underline{z} \sim \text{MVN}_p\left(\underline{0}, \mathcal{I}_p\right)$.
- Compute $\underline{v} = U^T\underline{z}$.
- Return $\Omega_\mu^{-1}\left(\Psi^{-1}(\sum_{i=1}^{N}\underline{x}_i - \sum_{i=1}^{N}\Lambda\underline{\eta}_i) + \Sigma_\mu^{-1}\underline{\tilde{\mu}}\right) + \underline{v}$ (3.11).

3.5 Gibbs Sampler Pseudo-Code

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

ii) Initalise:

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

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.

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 $\text{E}\left(\underline{\eta}_i^\star\right) = \underline{0}$ and $\text{Cov}\left(\underline{\eta}_i^\star\right) = \mathcal{I}_q$. 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.

4 Introducing the Shrinkage Prior

4.1 Multiplicative Gamma Process Shrinkage Priors

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. 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.

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 with the dimension p of the outcomes. 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 large p, small N 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 would be small when the truth is sparse. However, we need a computational strategy for choosing an appropriate level of truncation $q^\star$. We would like to strike a balance between missing important factors by choosing $q^\star$ too small and wasting computation 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. We thereby save on computation by discarding unimportant factors. For this reason, the sampler described in 3.5 above is modified to an adaptive Gibbs sampler, which tunes the number of factors as the sampler progresses. We begin with a default value for $\tilde{q}$ of $\min\left(\lfloor 3\ln(p)\rfloor, p, N-1\right)$. We adapt only after the burn-in period has elapsed, 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 after burn-in, with b_0, b_1 chosen so that adaptation occurs around every 10 iterations at the beginning of the chain but decreases in frequency exponentially fast. We chose b_0 and b_1 in the adaptation probability as -0.1 and -5×10^{-5} respectively. We generate a sequence u_t of uniform random numbers between 0 and 1. If $u_t \leq \mathrm{p}\left(t\right)$ at the t-th iteration, 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 the non-redundant columns are retained. The other parameters are also modified accordingly. 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 as an estimate of $q^\star$ with credible intervals quantifying uncertainty. Thus a histogram approximation to the posterior for $q^\star$ is introduced and may be used to address the question about the number of latent factors.

5 Extension to Clustering Heterogeneous Data

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 a 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

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

5.2.3 Mixtures of Infinite Factor Analyzers Pseudo-Code

1. Choose scalar hyperparameters as before.

2. Start by initialising the cluster labels $Z^{(0)}$: simulate from the Mult $(1, \underline{\pi})$ prior N times, or employ another clustering algorithm, such as K-Means. Compute $\underline{n}$, and $\underline{\tilde{\mu}}_g$ for each group.

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

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

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

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

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 $j = 1, \ldots, p$

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

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

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

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

5. Re-compute $\underline{n}$ and sample $\underline{\pi}$ from Dir $(\underline{\alpha} + \underline{n})$.

6. For $i = 1, \ldots, N$, sample $\underline{z}_i$ as outlined in (5.5).

7. Follow the adaptation procedure outlined in 4.3[2].

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

9. Disregard the first B burn-in iterations and thin every K-th iteration [3].

5.3 Label Switching

It's easy to see that $\mathrm{P}(X \mid \gamma) = \mathrm{P}(X \mid \tilde{\gamma})$ 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 Overfitting 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. Overfitting mixtures, along with Dirichlet Processes (5.5), are a means of extending the MIFA methodology in order to estimate G in a similarly choice-free manner. The prior in 5.2.1 plays an important role. This method approaches mixture model estimation by initially overfitting the number of clusters expected to be present, and specified conditions on the Dirichlet hyperparameter for the cluster mixing proportions encourage the emptying out of excess components in the posterior distribution.

To initialise the method, a conservatively high number of groups $G^\star$ 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.

[2] Our R-package also implements MFA, without the MGP shrinkage prior in 4.1 and adaptation.

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

5.5 Dirichlet Process Mixtures

6 References

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

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

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.

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`.

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`.

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.