Infinite Mixtures of Infinite Factor Analysers

Notes & Derivations

Keefe Murphy[1, 2], Dr. Claire Gormley[1, 2], and Prof. Brendan Murphy[1, 2]

[1]Department of Mathematics and Statistics, UCD
[2]Insight Centre for Data Analytics, UCD

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 $\underline{\Sigma}$. The factor model states that $\underline{x}$ is linearly independent upon a few ($q \ll p$) unobservable random variables $f_1, f_2, \ldots, f_q$ called *common factors* and p additional sources of variation $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ called *specific factors*, s.t.

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

where	$\underline{x}_i$	$\rightarrow$	$(p \times 1)$	observation vector
	$\underline{\mu}$	$\rightarrow$	$(p \times 1)$	overall mean vector
	$\underline{\Lambda}$	$\rightarrow$	$(p \times q)$	loadings matrix
	$\underline{\mathrm{f}}_i$	$\rightarrow$	$(q \times 1)$	vector of factor scores for obs i
	$\underline{\varepsilon}_i$	$\rightarrow$	$(p \times 1)$	vector of errors for obs i
	i	$=$	$1, \ldots, N$	
	j	$=$	$1, \ldots, p$	
	k	$=$	$1, \ldots, q$	

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

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

1.2 Assumptions

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

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

   $$\therefore \underline{\varepsilon}_i \sim \mathrm{MVN}_p\left(\underline{0}, \underline{\Psi}\right) \tag{2}$$

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

   $$\therefore \mathrm{E}\left(\underline{\mathrm{f}}\right) = \underline{0} \text{ and } \mathrm{Cov}\left(\underline{\mathrm{f}}\right) = \begin{pmatrix} 1_1 & 0 & \ldots & 0 \\ 0 & 1_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1_q \end{pmatrix} = \mathcal{I}_q$$

   $$\therefore \underline{\mathrm{f}}_i \sim \mathrm{MVN}_q\left(\underline{0}, \mathcal{I}_q\right) \tag{3}$$

2 Bayesian Framework

2.1 Likelihood

$$\begin{aligned}
\mathrm{E}\left(\underline{x}_i^\star\right) &= \mathrm{E}\left(\underline{\Lambda}\underline{\mathrm{f}}_i + \underline{\varepsilon}_i\right) \\
&= \underline{\Lambda}\mathrm{E}\left(\underline{\mathrm{f}}_i\right) + \mathrm{E}\left(\underline{\varepsilon}\right) \\
&= \underline{0} \\
\therefore \underline{X}_i^\star &\sim \mathrm{MVN}_p\left(\underline{0}, \underline{\Sigma}\right)
\end{aligned} \tag{4}$$

$$\begin{aligned}
\text{Since } \underline{\varepsilon}_i &= \underline{x}_i^\star - \underline{\Lambda}\underline{\mathrm{f}}_i \\
\underline{\Sigma} &= \mathrm{Cov}\left(X\right) \\
&= \mathrm{E}\left[\left(\underline{x} - \underline{\mu}\right)\left(\underline{x} - \underline{\mu}\right)^T\right] \\
&= \mathrm{E}\left[x^\star x^{\star T}\right] \\
&= \mathrm{E}\left[\left(\underline{\Lambda}\underline{\mathrm{f}} + \underline{\varepsilon}\right)\left(\underline{\Lambda}\underline{\mathrm{f}} + \underline{\varepsilon}\right)^T\right] \\
&= \mathrm{E}\left[\left(\underline{\Lambda}\underline{\mathrm{f}}\right) + \underline{\varepsilon}\left(\underline{\Lambda}\underline{\mathrm{f}}\right)^T + \left(\underline{\Lambda}\underline{\mathrm{f}}\right)\underline{\varepsilon}^T + \underline{\varepsilon}\underline{\varepsilon}^T\right] \\
&= \underline{\Lambda}\mathrm{E}\left(\underline{\mathrm{f}}\underline{\mathrm{f}}^T\right)\underline{\Lambda}^T + \mathrm{E}\left(\underline{\varepsilon}\underline{\mathrm{f}}^T\right)\Lambda^T + \Lambda\mathrm{E}\left(\underline{\mathrm{f}}\underline{\varepsilon}^T\right) + \mathrm{E}\left(\underline{\varepsilon}\underline{\varepsilon}^T\right) \\
&= \underline{\Lambda}\underline{\Lambda}^T + \underline{\Psi} \\
\therefore \underline{X}_i^\star &\sim \mathrm{MVN}_p\left(\underline{0}, \underline{\Lambda}\underline{\Lambda}^T + \underline{\Psi}\right)
\end{aligned} \tag{5}$$

$$\begin{aligned}
\mathrm{E}\left(\underline{X}_i^\star \,|\, \underline{\mathrm{f}}_i\right) &= \mathrm{E}\left(\underline{\Lambda}\underline{\mathrm{f}}_i + \underline{\varepsilon}_i \,|\, \underline{\mathrm{f}}_i\right) \\
&= \underline{\Lambda}\mathrm{E}\left(\underline{\mathrm{f}}_i \,|\, \underline{\mathrm{f}}_i\right) + \mathrm{E}\left(\underline{\varepsilon}_i \,|\, \underline{\mathrm{f}}_i\right) \\
&= \underline{\Lambda}\underline{\mathrm{f}}_i \\
\mathrm{Cov}\left(\underline{X}_i^\star \,|\, \underline{\mathrm{f}}_i\right) &= \mathrm{E}\left[\left(\underline{X}_i^\star - \underline{\Lambda}\underline{\mathrm{f}}_i\right)\left(\underline{X}_i^\star - \underline{\Lambda}\underline{\mathrm{f}}_i\right)^T \,|\, \underline{\mathrm{f}}_i\right] \\
&= \mathrm{E}\left(\underline{\varepsilon}_i\underline{\varepsilon}_i^T \,|\, \underline{\mathrm{f}}_i\right) \\
&= \underline{\Psi} \\
\therefore \underline{X}_i^\star \,|\, \underline{\mathrm{f}}_i, \underline{\Lambda}, \underline{\Psi} &\sim \mathrm{MVN}_p\left(\underline{\Lambda}\underline{\mathrm{f}}_i, \underline{\Psi}\right)
\end{aligned} \tag{6}$$

The density of the data is then given by:

$$\begin{aligned}
\mathrm{P}\left(\underline{X}_i^\star \,|\, \underline{\mathrm{f}}_i, \underline{\Lambda}, \underline{\Psi}\right) &= (2\pi)^{-\frac{p}{2}} \,|\, \underline{\Psi} \,|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N}\left(\underline{X}_i^\star - \underline{\Lambda}\underline{\mathrm{f}}_i\right)^T \underline{\Psi}^{-1}\left(\underline{X}_i^\star - \underline{\Lambda}\underline{\mathrm{f}}_i\right)\right) \\
&\propto \,|\, \underline{\Psi} \,|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left[\underline{\Psi}^{-1}\left(\underline{X} - \underline{F\Lambda}\right)^T\left(\underline{X} - \underline{F\Lambda}\right)\right]\right)
\end{aligned} \tag{7}$$

$$\text{Where } \underline{\Lambda}_{(p\times q)} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1q} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pq} \end{pmatrix}$$

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

2.2 Posterior Set-Up

$$\begin{aligned}
\text{Likelihood} &= \mathrm{P}\left(X^{\star} \mid \underline{\theta}\right) \\
&= \mathrm{P}\left(\underline{X}_i^{\star} \mid \underline{\mathrm{f}}_i, \underline{\Lambda}, \underline{\Psi}\right) \\
\therefore \mathrm{P}\left(\underline{X}_i^{\star} \mid \underline{\mathrm{f}}_i, \underline{\Lambda}, \underline{\Psi}\right) &\sim \mathrm{MVN}_p\left(\underline{\Lambda}\underline{\mathrm{f}}_i, \underline{\Psi}\right) \qquad (8)
\end{aligned}$$

$$\begin{aligned}
\text{Prior} &= \mathrm{P}\left(\underline{\theta}\right) \\
&= \mathrm{P}\left(\underline{F}\right) \mathrm{P}\left(\underline{\Lambda}\right) \mathrm{P}\left(\underline{\Psi}\right)
\end{aligned}$$

$$\begin{aligned}
\text{Posterior} &\propto \text{Likelihood} \times \text{Prior} \\
\therefore \mathrm{P}\left(\underline{\mathrm{F}}, \underline{\Lambda}, \underline{\Psi} \mid \underline{X}\right) &\propto \mathcal{L}\left(\underline{X}^{\star} \mid \underline{F}, \underline{\Lambda}, \underline{\Psi}\right) \mathrm{P}\left(\underline{F}\right) \mathrm{P}\left(\underline{\Lambda}\right) \mathrm{P}\left(\underline{\Psi}\right) \\
&\propto \left[\prod_{i=1}^{N} \mathrm{P}\left(\underline{X}_i^{\star} \mid \underline{\mathrm{f}}_i, \underline{\Lambda}, \underline{\Psi}\right)\right] \left[\prod_{i=1}^{N} \mathrm{P}\left(\underline{\mathrm{f}}_i\right)\right] \left[\prod_{j=1}^{p} \mathrm{P}\left(\underline{\Lambda}_j\right)\right] \left[\prod_{j=1}^{p} \mathrm{P}\left(\Psi_{jj}\right)\right] \qquad (9)
\end{aligned}$$

Later on, especially as we move into the mixture case, it will be necessary to undo the centering, thereby removing the * on $\underline{X}$, and reintroduce $\underline{\mu}$. This will necessetate multiplying the quantity in (Equation (9)) by $\left[\prod_{j=1}^{p} \mathrm{P}\left(\mu_j\right)\right]$. 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 - $\underline{\mathrm{f}}_i$

$$\begin{aligned}
\underline{\mathrm{f}}_i &\sim \mathrm{MVN}_q\left(\underline{0}, \mathcal{I}_q\right) \\
&= (2\pi)^{-\frac{q}{2}} \exp\left(-\frac{1}{2}\underline{\mathrm{f}}_i^T \underline{\mathrm{f}}_i\right) \qquad (10)
\end{aligned}$$

To obtain the full conditional for $\underline{\mathrm{f}}_i$ we can multiply the conditional likelihood by the marginal distribution in (Equation(10)) s.t.

$$\begin{aligned}
\mathrm{P}\left(\underline{\mathrm{f}}_i \mid, \underline{X}_i^{\star}, \underline{\Lambda}, \underline{\Psi}\right) &\sim \mathrm{P}\left(\underline{X}_i^{\star} \mid \underline{\mathrm{f}}_i, \underline{\Lambda}, \underline{\Psi}\right) \mathrm{P}\left(\underline{\mathrm{f}}_i\right) \\
&\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N}\left(\underline{X}_i^{\star} - \underline{\Lambda}\underline{\mathrm{f}}_i\right)^T \underline{\Psi}^{-1}\left(\underline{X}_i^{\star} - \underline{\Lambda}\underline{\mathrm{f}}_i\right) + \underline{\mathrm{f}}_i^T \underline{\mathrm{f}}_i\right) \\
&\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N}\left[-\underline{X}_i^{\star T}\underline{\Psi}^{-1}\underline{\Lambda}\underline{\mathrm{f}}_i - \left(\underline{\Lambda}\underline{\mathrm{f}}_i\right)^T \underline{\Psi}^{-1}\underline{X}_i^{\star} + \left(\underline{\Lambda}\underline{\mathrm{f}}_i\right)^T \underline{\Psi}^{-1}\left(\underline{\Lambda}\underline{\mathrm{f}}_i\right) - \underline{\mathrm{f}}_i^T \underline{\mathrm{f}}_i\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left\{\underline{\mathrm{f}}_i^T\left[\mathcal{I}_q + \underline{\Lambda}^T\underline{\Psi}^{-1}\underline{\Lambda}\right]\underline{\mathrm{f}}_i\right\} + \underline{X}_i^{\star T}\underline{\Psi}^{-1}\underline{\Lambda}\underline{\mathrm{f}}_i\right) \qquad (11)
\end{aligned}$$

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

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

$\therefore$ we can identify the μ and Σ^{-1} terms from (Equation (11)) above to yield

$$\mathrm{P}\left(\underline{\mathrm{f}}_i \mid, \underline{X}_i^{\star}, \underline{\Lambda}, \underline{\Psi}\right) \sim \mathrm{MVN}_q\left(\left[\mathcal{I}_q + \underline{\Lambda}^T\underline{\Psi}^{-1}\underline{\Lambda}\right]^{-1}\underline{\Lambda}^T\underline{\Psi}^{-1}\underline{X}_i^{\star}, \left[\mathcal{I}_q + \underline{\Lambda}^T\underline{\Psi}^{-1}\underline{\Lambda}\right]^{-1}\right) \qquad (12)$$

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 $\underline{\Omega}_F = \mathcal{I}_q + \underline{\Lambda}^T \underline{\Psi}^{-1} \underline{\Lambda}$
- Compute the Cholesky Factorization $\underline{\Omega}_F = U^T U$.
- Sample $z \sim \text{MVN}_q(0, \mathcal{I}_q)$ N times.
- Backsolve $Uv = z^T$.
- Compute $\underline{\Omega}_F^{-1}$ from U.
- Return $\left(\underline{\Omega}_F^{-1} \underline{\Lambda}^T \underline{\Psi}^{-1} \left(\underline{X} - \underline{\mu}\right)^T + v\right)^T$ (13).

3.2 Loadings Matrix - $\underline{\Lambda}$

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

$$
\begin{aligned}
\text{P}\left(\underline{\Lambda}_j \,|\, \underline{X}^\star, \underline{F}, \underline{\Psi}\right) &\sim \text{P}\left(\underline{X}^\star \,|\, \underline{F}, \underline{\Lambda}_j, \underline{\Psi}\right) \text{P}\left(\underline{\Lambda}_j \,|\, \Sigma_\lambda\right) \\
&\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N}\left(\underline{X}_i^\star - \underline{\Lambda}_j \underline{\text{f}}_i\right)^T \Psi_{jj}^{-1}\left(\underline{X}_i^\star - \underline{\Lambda}_j \underline{\text{f}}_i\right)\right) \exp\left(-\frac{1}{2}\left(\underline{\Lambda}_j^T \left(\Sigma_\lambda \mathcal{I}_q\right)^{-1} \underline{\Lambda}_j\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N}\left[-2\underline{X}_i^{\star T} \Psi_{jj}^{-1}\left(\underline{\Lambda}_j \underline{\text{f}}_i\right) + \left(\underline{\Lambda}_j \underline{\text{f}}_i\right)^T \Psi_{jj}^{-1}\left(\underline{\Lambda}_j \underline{\text{f}}_i\right) + \underline{\Lambda}_j^T \left(\Sigma_\lambda \mathcal{I}_q\right)^{-1} \underline{\Lambda}_j\right]\right) \\
&\propto \exp\left(\underline{\Lambda}_j \Psi_{jj}^{-1} \sum_{i=1}^{N} X_{ij}^{\star T} \underline{\text{f}}_i - \frac{1}{2}\underline{\Lambda}_j^T \left[\sum_{i=1}^{N} \Psi_{jj}^{-1} \underline{\text{f}}_i^T \underline{\text{f}}_i\right] \underline{\Lambda}_j - \frac{1}{2}\underline{\Lambda}_j^T \left(\Sigma_\lambda \mathcal{I}_q\right)^{-1} \underline{\Lambda}_j\right) \\
&\propto \exp\left(\underline{\Lambda}_j \left[\underline{F}^T \Psi_{jj}^{-1} \underline{X}^{j\star}\right] - \frac{1}{2}\underline{\Lambda}_j^T \left[\left(\Sigma_\lambda \mathcal{I}_q\right)^{-1} + \Psi_{jj}^{-1} \underline{F}^T \underline{F}\right] \underline{\Lambda}_j\right) \qquad (14)
\end{aligned}
$$

where $\underline{X}^{j\star}$ denotes the j-th column of $\underline{X}^\star$

$$
\therefore \text{P}\left(\underline{\Lambda}_j \,|\, \underline{X}^\star, \underline{F}, \underline{\Psi}\right) \sim \text{MVN}_q\left(\left[\left(\Sigma_\lambda \mathcal{I}_q\right)^{-1} + \Psi_{jj}^{-1} \underline{F}^T \underline{F}\right]^{-1} \underline{F}^T \Psi_{jj}^{-1} \underline{X}^{j\star}, \left[\left(\Sigma_\lambda \mathcal{I}_q\right)^{-1} + \Psi_{jj}^{-1} \underline{F}^T \underline{F}\right]^{-1}\right) \qquad (15)
$$

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

- Calculate $\underline{\Omega}_{\lambda_j} = \left(\Sigma_\lambda \mathcal{I}_q\right)^{-1} + \Psi_{jj}^{-1} \underline{F}^T \underline{F}$.
- Compute the Cholesky Factorization $\underline{\Omega}_{\lambda_j} = U^T U$.
- Sample $z \sim \text{N}(0, 1)$ q times.
- Back-solve $Uv = z$.
- Compute $\underline{\Omega}_{\lambda_j}^{-1}$ from U.
- Return $\underline{\Omega}_{\lambda_j}^{-1} \underline{F}^T \Psi_{jj}^{-1}\left(\underline{X}^j - \mu_j\right) + v$ (16).

[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. $U^T U = \Omega$, sample from $z \sim \text{N}(0, 1)$, then backsolve $L^T v = Uv = z$ s.t $x = \mu + v = \mu + L^{-T} z = \mu + U^{-1} z$. Then:

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

3.3 Uniquenesses - $\underline{\Psi}$

If we suggest an Inverse Wishart prior distribution for $\underline{\Psi}$, we have:

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

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

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

Since $\underline{\Psi}$ is a diagonal matrix:

$$\mathrm{P}\left(\underline{\Psi}^{-1}\right) \propto \prod_{j=1}^{p} |\,\Psi_{jj}^{-1}\,|^{\frac{N}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathcal{S}^{\star}_{jj}\Psi_{jj}^{-1}\right)\right)$$

This suggests the prior for $\underline{\Psi}^{-1}$ is a product of p $\mathrm{Ga}\left(\alpha/2, \beta/2\right)$ distributions.

$$\begin{aligned}
\therefore \mathrm{P}\left(\underline{\Psi}^{-1} \,|\, \alpha, \beta\right) &= \prod_{j=1}^{p} \mathrm{P}\left(\Psi_{jj}^{-1} \,|\, \alpha, \beta\right) \\
&\propto \prod_{j=1}^{p} \left(\Psi_{jj}^{-1}\right)^{\frac{\alpha}{2}-1} \exp\left(-\frac{\beta}{2}\Psi_{jj}^{-1}\right) \\
\therefore \mathrm{P}\left(\underline{\Psi}^{-1} \,|\, \underline{X}^{\star}, \underline{F}, \underline{\Lambda}\right) &\propto \mathrm{P}\left(\underline{X}^{\star} \,|\, \underline{F}, \underline{\Lambda}\right) \mathrm{P}\left(\underline{\Psi}^{-1} \,|\, \alpha, \beta\right) \\
&\propto \prod_{j=1}^{p} \left(\Psi_{jj}^{-1}\right)^{\frac{N}{2}} \exp\left(-\frac{\mathcal{S}^{\star}_{jj}}{2}\Psi_{jj}^{-1}\right) \prod_{j=1}^{p} \left(\Psi_{jj}^{-1}\right)^{\frac{\alpha}{2}-1} \exp\left(-\frac{\beta}{2}\Psi_{jj}^{-1}\right) \\
&\propto \prod_{j=1}^{p} \left(\Psi_{jj}^{-1}\right)^{\frac{N+\alpha}{2}-1} \exp\left(-\frac{\mathcal{S}^{\star}_{jj} + \beta}{2}\Psi_{jj}^{-1}\right) \qquad (17) \\
\text{where } \mathcal{S}^{\star}_{jj} &= \sum_{i=1}^{N} \left(x_{ij} - \underline{\Lambda}_j \underline{\mathrm{f}}_i\right)^2
\end{aligned}$$

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

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

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

$$\mathrm{P}\left(\Psi_{jj}^{-1} \,|\, \underline{X}, \underline{F}, \underline{\Lambda}\right) \sim \mathrm{Ga}\left(\frac{N+\alpha}{2}, \frac{S_{jj}+\beta}{2}\right) \qquad (18)$$

3.4 Reintroducing $\underline{\mu}$

We've already seen from (Equations (13), (16) and (18)) 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. $\underline{\mu} \sim \mathrm{MVN}_p\left(\underline{0}, \Sigma_\mu \mathcal{I}_p\right)$ where Σ_μ is a scalar hyperparameter which controls the diagonal covariance matrix of the prior. As above, we can expect the result of the product of two MVN_p distributions to itself be distributed in the same way.

$$\begin{aligned}
\mathrm{P}\left(\underline{\mu} \,|\, \underline{X}, \underline{F}, \underline{\Psi}, \Lambda\right) &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N} \left(\underline{X}_i - \underline{\mu} - \underline{\Lambda \mathrm{f}}_i\right)^T \underline{\Psi}^{-1} \left(\underline{X}_i - \underline{\mu} - \underline{\Lambda \mathrm{f}}_i\right)\right) \exp\left(-\frac{1}{2}\left(\underline{\mu}^T \left(\Sigma_\mu \mathcal{I}_p\right)^{-1} \underline{\mu}\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{N} \left[-2\underline{X}_i^T \Psi^{-1} \underline{\mu} + 2\left(\underline{\Lambda \mathrm{f}}_i\right)^T \underline{\Psi}^{-1} \underline{\mu} + \underline{\mu}^T \underline{\Psi}^{-1} \underline{\mu} + \underline{\mu}^T \left(\Sigma_\mu \mathcal{I}_p\right)^{-1} \underline{\mu}\right]\right) \\
&\propto \exp\left(\sum_{i=1}^{N} \underline{X}_i^T \underline{\Psi}^{-1} \underline{\mu} - \sum_{i=1}^{N} \left(\underline{\Lambda \mathrm{f}}_i\right)^T \underline{\Psi}^{-1} \underline{\mu} - \frac{1}{2}\left[\underline{\mu}^T \left(\left(\Sigma_\mu \mathcal{I}_p\right)^{-1} + N\underline{\Psi}^{-1}\right) \underline{\mu}\right]\right) \qquad (19)
\end{aligned}$$

$$\therefore \mathrm{P}\left(\underline{\mu} \,|\, \underline{X}, \underline{F}, \underline{\Psi}, \underline{\Lambda}\right) \sim \mathrm{MVN}_p\Big(\left[(\Sigma_\mu \mathcal{I}_p)^{-1} + n\underline{\Psi}^{-1}\right]^{-1} \underline{\Psi}^{-1}\Big(\sum_{i=1}^{N} \underline{X}_i - \sum_{i=1}^{N} \underline{\Lambda \mathrm{f}}_i\Big), \left[(\Sigma_\mu \mathcal{I}_p)^{-1} + N\underline{\Psi}^{-1}\right]^{-1} \Big) \tag{20}$$

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

- Calculate $\underline{\Omega}_\mu = (\Sigma_\mu \mathcal{I}_p)^{-1} + N\underline{\Psi}^{-1}$, which is a diagonal $p \times p$ matrix.
- Invert $\underline{\Omega}_\mu$ by inverting its diagonal elements.
- Compute the Cholesky Factorization $\underline{\Omega}_\mu^{-1} = U^T U$.
- Sample $z \sim \mathrm{N}(0,1)$ q times.
- Compute $v = U^T z$.
- Return $\underline{\Omega}_\mu^{-1}\underline{\Psi}^{-1}\Big(\sum_{i=1}^{N} \underline{X}_i - \sum_{i=1}^{N} \underline{\Lambda \mathrm{f}}_i\Big) + v$ (21).

3.5 Gibbs Sampler Pseudo-Code

i) Choose scalar hyperparameters $\Sigma_\mu, \Sigma_\lambda, \alpha,$ and β, and select q.

ii) Initalise:

$$\begin{aligned}
\underline{\mu}^{(0)} &\sim \mathrm{MVN}_p\left(\underline{0}, \Sigma_\mu \mathcal{I}_p\right) \\
\underline{F}^{(0)} &\sim \mathrm{MVN}_q\left(N, \underline{0}, \mathcal{I}_q\right) \\
\underline{\Lambda}^{(0)} &\sim \mathrm{MVN}_q\left(N, \underline{0}, \Sigma_\lambda \mathcal{I}_q\right) \\
\underline{\Psi}^{(0)^{-1}} &\sim \mathrm{Ga}_p\left(\alpha/2, \beta/2\right)
\end{aligned}$$

iii) For $t = 1, \ldots, T$, using the routines specified in (Equations (13), (16), (18) and (21)):

$$\begin{aligned}
\text{a)}\quad \underline{\Omega}_\mu^{(t)} &= (\Sigma_\mu \mathcal{I}_p)^{-1} + N\underline{\Psi}^{-1^{(t-1)}} \\
\text{b)}\quad \underline{\mu}^{(t)} &\sim \mathrm{MVN}_p\left(\underline{\Omega}_\mu^{-1^{(t)}} \underline{\Psi}^{-1^{(t-1)}} \Big(\sum_{i=1}^{N} \underline{X}_i - \sum_{i=1}^{N} \underline{\Lambda}^{(t-1)} \underline{\mathrm{f}}_i^{(t-1)}\Big), \underline{\Omega}_\mu^{(t)}\right) \\
\text{c)}\quad \underline{\Omega}_F^{(t)} &= \mathcal{I}_q + \underline{\Lambda}^{T^{(t-1)}} \underline{\Psi}^{-1^{(t-1)}} \underline{\Lambda}^{(t-1)} \\
\text{d)}\quad \underline{\mathrm{f}}_i^{(t)} &\sim \mathrm{MVN}_q\left(\underline{\Omega}_F^{-1^{(t)}} \underline{\Lambda}^{T^{(t-1)}} \underline{\Psi}^{-1^{(t-1)}} \left(\underline{X}_i - \underline{\mu}^{(t)}\right), \underline{\Omega}_F^{-1^{(t)}}\right) \\
\text{e)}\quad \text{For } j &= 1, \ldots, p \\
\bullet\ \underline{\Omega}_{\lambda_j}^{(t)} &= (\Sigma_\lambda \mathcal{I}_q)^{-1} + \Psi_{jj}^{-1^{(t-1)}} \underline{F}^{T^{(t-1)}} \underline{F}^{(t)} \\
\bullet\ \underline{\Lambda}_j^{(t)} &\sim \mathrm{MVN}_q\left(\underline{\Omega}_{\lambda_j}^{-1^{(t)}} \underline{F}^{T^{(t)}} \Psi_{jj}^{-1^{(t-1)}} \left(\underline{X}^j - \underline{\mu}_j^{(t)}\right), \underline{\Omega}_{\lambda_j}^{-1^{(t)}}\right) \\
\text{f)}\quad \Psi^{-1^{(t)}} &\sim \mathrm{Ga}_p\left(\frac{N+\alpha}{2}, \frac{S_{jj}^{(t)} + \beta}{2}\right)
\end{aligned}$$

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

3.6 Issues Around Identifiability

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

$$\begin{aligned}\underline{x} - \underline{\mu} &= \underline{\Lambda}\underline{\mathrm{f}} + \underline{\varepsilon} \\ &= \underline{\Lambda}\underline{T}\underline{T}^T\underline{\mathrm{f}} + \underline{\varepsilon} \\ &= \underline{\Lambda}^\star\underline{\mathrm{f}}^\star + \underline{\varepsilon}\end{aligned}$$

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

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

However, we can solve 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 burn-in iteration. This Procustean map is a rotation only, i.e. translation, scaling, dilation, etc. are not applied. We then also apply that same rotation matrix at each iteration to each iteration of the matrix of factor scores. This amounts to *post-multiplying* the loadings matrix at each iteration by the Procrustes rotation matrix that maps to the loadings template, and also, by using the identity $(AB)^T = B^T A^T$, letting each iteration's scores matrix equal the transpose of the *post-multiplication* of that iteration's score matrix by that same rotation matrix.

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 an accurate approximation to the infinite factor model.

The exact specification of this shrinkage-type prior has the degree of shrinkage increasing across the column index as follows:

$$\begin{aligned}\lambda_{jk} \,|\, \phi_{jk}, \tau_k &\sim \mathrm{N}\left(0, \phi_{jk}^{-1}\tau_k^{-1}\right) \\ \text{s.t.} \quad \underline{\lambda}_j \,|\, \underline{\phi}_j, \underline{\tau} &\sim \mathrm{MVN}_{q^\star}\left(\underline{0}, \underline{\mathrm{D}}_j\right) \qquad (22) \\ \text{where} \quad \underline{\mathrm{D}}_j^{-1} &= diag\left(\phi_{j1}\tau_1, \ldots, \phi_{jq^\star}\tau_{q^\star}\right) \\ \phi_{jk} &\sim \mathrm{Ga}\left(\nu/2, \nu/2\right) \qquad (23) \\ \tau_k &= \prod_{h=1}^{k} \delta_h \\ \delta_1 \sim \mathrm{Ga}\left(\alpha_1, 1\right), &\quad \delta_h \sim \mathrm{Ga}\left(\alpha_2, 1\right), \quad h \geq 2 \qquad (24)\end{aligned}$$

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 $\alpha_2 > 1$, which favours more shrinkage as the column index increases.

4.2 Deriving 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 (Section 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 s.t. $\mathrm{P}\left(\underline{\mathrm{f}}_i \,|\, —\right) \sim \mathrm{MVN}_{q^\star}\left(\left[\mathcal{I}_{q^\star} + \underline{\Lambda}_{q^\star}^T \underline{\Psi}^{-1} \underline{\Lambda}_{q^\star}\right]^{-1} \underline{\Lambda}^T \underline{\Psi}^{-1} \underline{X}_i^\star, \left[\mathcal{I}_{q^\star} + \underline{\Lambda}_{q^\star}^T \underline{\Psi}^{-1} \underline{\Lambda}_{q^\star}\right]^{-1}\right)$

4.2.1 Loadings Matrix - $\underline{\Lambda}$

Incorporating the new prior (Equation (22)), and following the same steps as (Section 3.2) above, it is trivial to show that the $\underline{\Lambda}_j$s now have independent conditionally conjugate posteriors given by:

$$\mathrm{P}\left(\underline{\Lambda}_j \,|\, —\right) \sim \mathrm{MVN}_{q^\star}\left(\left[\underline{\mathrm{D}}_j^{-1} + \Psi_{jj}^{-1} \underline{F}^T \underline{F}\right]^{-1} \underline{F}^T \Psi_{jj}^{-1} \underline{X}^{j^\star}, \left[\underline{\mathrm{D}}_j^{-1} + \Psi_{jj}^{-1} \underline{F}^T \underline{F}\right]^{-1}\right) \tag{25}$$

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

- Calculate $\underline{\Omega}_{\lambda_j} = \underline{\mathrm{D}}_j^{-1} + \Psi_{jj}^{-1} \underline{F}^T \underline{F}$
- Compute the Cholesky Factorization $\underline{\Omega}_{\lambda_j} = U^T U$.
- Sample $z \sim \mathrm{N}(0, 1)$ q times.
- Back-solve $Uv = z$.
- Compute $\underline{\Omega}_{\lambda_j}^{-1}$ from U.
- Return $\underline{\Omega}_{\lambda_j}^{-1} \underline{F}^T \Psi_{jj}^{-1} \left(\underline{X}^j - \mu_j\right) + v$ (26).

4.2.2 Local Shrinkage

Local Shrinkage – ϕ_{jk} Using the conditional prior in (Equation (22)) and the prior for ϕ_{jk} in (Equation (23)) we can derive the full conditional for the local shrinkage parameter as follows:

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

Thus the posterior distribution of each ϕ_{jk} is given by:

$$\mathrm{P}\left(\phi_{jk} \,|\, —\right) \sim \mathrm{Ga}\left(\frac{\nu+1}{2}, \frac{\nu + \tau_k \lambda_{jk}^2}{2}\right) \tag{27}$$

4.2.3 Global Shrinkage

Global Shrinkage – τ_k Using the conditional prior in (Equation (22)) and the prior for τ_k in (Equation (24)) we can derive the full conditional for the global shrinkage parameter, in three stages – first by deriving and sampling from $\mathrm{P}(\delta_1 \,|\, —)$ & $\mathrm{P}(\delta_k \,|\, —)$ for $k \geq 2$, as follows below — and then obtaining the product $\tau_k = \prod_{h=1}^{k} \delta_h$ thereafter:

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

$$
\text{where } \tau_h^{(k)} = \prod_{t=1}^{h} \frac{\delta_t}{\delta_k} \text{ for } k = 1, \ldots, q^\star \tag{28}
$$

$$
\therefore \mathrm{P}(\delta_1 \,|\, —) \sim \mathrm{Ga}\left(\alpha_1 + \frac{pq^\star}{2}, 1 + \frac{1}{2} \sum_{h=1}^{q^\star} \tau_h^{(1)} \sum_{j=1}^{p} \lambda_{jh}^2 \phi_{jh}\right) \tag{29}
$$

Similarly:

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

$$
\therefore \mathrm{P}(\delta_k \,|\, —) \sim \mathrm{Ga}\left(\alpha_2 + \frac{p}{2}(q^\star - k + 1), 1 + \frac{1}{2} \sum_{h=k}^{q^\star} \tau_h^{(k)} \sum_{j=1}^{p} \lambda_{jh}^2 \phi_{jh}\right) \tag{30}
$$

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 factor model for different choices of $q^\star$, and then using the BIC or another criteria for model selection. This approach can be difficult to implement for large p, small N problems, and the BIC itself is not 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 conservative 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 $\underline{\Lambda}_{\tilde{q}}$ from the Gibbs sampler contain information about the effective number of factors. At the t-th iteration of the Gibbs sampler, let $m^{(t)}$ denote the number of columns in $\underline{\Lambda}_{\tilde{q}}$ 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 is 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 (Section 3.5) above will be 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 the lesser of P and $5\ln(P)$.

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 $\mathrm{p}(t) = \exp(b_0 + b_1 t)$ 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 $\mathrm{p}(t)$ as -0.1 and -5×10^{-5} respectively. We generate a sequence u_t of uniform random numbers between 0 and 1. At the t-th iteration, if $u_t \leq \mathrm{p}(t)$, 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, we add a column to the loadings. Otherwise, we discard the redundant columns. The other parameters are also modified accordingly. When we add a factor, we sample parameters from their prior distributions to fill in additional columns, and otherwise retain parameters corresponding to the non-redundant columns. 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 $\left\{q^{\star(t)}\right\}$ after burn-in as an estimate of $q^\star$ with credible intervals quantifying uncertainty.

5 Extension to Clustering Heterogeneous Data

5.1 Introducing Mixture Models

Marginally, (Equation 5) provides a parsimonious covariance matrix, s.t. $\underline{X}_i \sim \mathrm{MVN}_p\left(\underline{\mu}, \underline{\Lambda}\underline{\Lambda}^T + \underline{\Psi}\right)$. 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 spectific latent Gaussian model with covariance specified according to the form above. Let's now introduce some basic notation at this stage:

$$N = \sum_{g=1}^{G} n_g \qquad \text{where } n_g \text{ is the size of the } g\text{-th group.}$$

$$\mathrm{P}\left(\underline{X} \,|\, \underline{\gamma}\right) = \sum_{g=1}^{G} \underline{\pi}_g \mathrm{P}_g\left(\underline{X} \,|\, \underline{\theta}_g\right) \qquad \text{where } \underline{\gamma} = (\underline{\theta}_1, \ldots, \underline{\theta}_G, \underline{\pi}_1, \ldots, \underline{\pi}_G) \text{ and the p.d.f } \mathrm{P}_g \text{ is parametrized by } \underline{\theta}_g$$

$\underline{\pi}_g$ are known as the *cluster mixing proportions*, which have the following properties

$$\underline{\pi}_g \geq 0 \quad \forall\, g = 1, \ldots, G$$

$$\sum_{g=1}^{G} \underline{\pi}_g = 1$$

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

$$\begin{aligned} (\underline{X}_i \,|\, z_{ig} = 1) &\sim \text{MVN}_p \left(\underline{\mu}_g, \underline{\Lambda}_g \underline{\Lambda}_g^T + \underline{\Psi}_g \right) \\ \therefore \text{P}\left(\underline{X}_i\right) &= \sum_{g=1}^{G} \pi_g \, \text{MVN}_p \left(\underline{\mu}_g, \underline{\Lambda}_g \underline{\Lambda}_g^T + \underline{\Psi}_g \right) \end{aligned} \tag{31}$$

5.1.1 Decomposable Prior for $\underline{\gamma}$

The posterior distribution of $\underline{\gamma}$ is

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

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

$$\begin{aligned} \text{P}\left(\underline{\gamma}\right) &= \text{P}\left(\underline{\pi}\right) \prod_{g=1}^{G} \text{P}\left(\underline{\theta}_g\right), \\ \text{P}\left(\underline{\gamma} \,|\, \underline{X}, \underline{z}\right) &\propto \text{P}\left(\underline{\pi}\right) \prod_{g=1}^{G} \prod_{i:z_i=g} \text{P}\left(\underline{\theta}_g\right) \text{P}_g \left(\underline{X}_i \,|\, \underline{\theta}_g\right) \end{aligned} \tag{32}$$

5.2 Deriving Posterior Distributions

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

- Component Parameters:
  $\text{P}\left(\underline{\theta}_g \,|\, \underline{\theta}_{-g}, \underline{X}, \underline{z}\right) \equiv \text{P}\left(\underline{\theta}_g \,|\, \underline{X}, \underline{z}\right) \propto \prod_{i:z_i=g} \text{P}\left(\underline{\theta}_g\right) \text{P}_g \left(\underline{X}_i \,|\, \underline{\theta}_g\right)$
  where $\underline{\theta}_{-1} = \left(\underline{\theta}_1, \ldots, \underline{\theta}_{g-1}, \underline{\theta}_{g+1}, \ldots, \underline{\theta}_G\right)$

- Component Weights:
  $\text{P}\left(\underline{\pi} \,|\, \underline{X}, \underline{z}\right) \equiv \text{P}\left(\underline{\pi} \,|\, \underline{z}\right) \propto \text{P}\left(\underline{\pi}\right) \prod_{g=1}^{G} \pi_g^{n_g}$
  where n_g is the number of observations in group g,
  since $\text{P}\left(\underline{z} \,|\, \underline{\pi}\right) \sim \text{Mult}\left(N, \underline{\pi}\right)$

- Latent Classifier:
  $\text{P}\left(\underline{z} \,|\, \underline{X}, \underline{\gamma}\right) \propto \prod_{i=1}^{N} \text{P}\left(\underline{z}_i\right) \text{P}\left(\underline{X}_i \,|\, \underline{\theta}_i, \underline{z}_i\right)$

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

Let the prior distribution of $\underline{\pi}$ be Dirichlet with parameter $\underline{\alpha}$ – a higher order generalisation of the Beta distribution.

$$\begin{aligned}
\mathrm{P}\left(\underline{\pi}\right) &\propto \prod_{g=1}^{G} \underline{\pi}_g^{\alpha_g - 1} \\
\therefore \mathrm{P}\left(\underline{\pi} \,|\, \underline{z}, \underline{X}\right) &\propto \prod_{g=1}^{G} \underline{\pi}_g^{\alpha_g - 1} \prod_{g=1}^{G} \underline{\pi}_g^{n_g} \\
&\propto \prod_{g=1}^{G} \underline{\pi}_g^{\alpha_g + n_g - 1} \\
\text{s.t. } \mathrm{P}\left(\underline{\pi} \,|\, \underline{z}, \underline{X}\right) &\sim \mathrm{Dir}\left(\underline{\alpha} + \underline{n}\right) \qquad (33) \\
\text{where } \underline{n} &= \left(n_1, \ldots, n_G\right)
\end{aligned}$$

5.2.2 Cluster Labels – $\underline{z}_i$

$$\begin{aligned}
\left(\underline{z}_i \,|\, \underline{X}_i, \underline{\gamma}\right) &\sim \{1, \ldots, G\}, \quad \forall\, i = 1, \ldots, N \\
\mathrm{P}\left(\underline{z}_i = g \,|\, \underline{X}_i, \underline{\gamma}\right) &= \frac{\underline{\pi}_g \mathrm{P}\left(\underline{X}_i \,|\, \underline{\theta}_g\right)}{\sum_{g=1}^{G} \underline{\pi}_g \mathrm{P}\left(\underline{X}_i \,|\, \underline{\theta}_g\right)}
\end{aligned}$$

5.2.3 MCMC Algorithm Pseudo-Code

1. Choose scalar hyperparameters as before.
2. Initialise:
   (a) a
   (b) b
   (c) c
3. Start with some initialisation of the cluster labels $\underline{z}^{(0)}$ from the Mult $(N, \underline{\pi})$ prior.
4. Compute $\underline{n}$.
5. For $g = 1, \ldots, G$, sample other parameters as before, but this time from their *group specific* full conditionals:
   (a) a
   (b) b
   (c) c
6. Follow the adaptation procedure outlined in (Section 4.3).
7. Sample $\underline{\pi}$ from Dir $(\underline{\alpha} + \underline{n})$.
8. For $i = 1, \ldots, N$, sample $\underline{z}_i$.
9. Repeat steps 4–8 for $t = 1, \ldots, T$.
10. Disregard the first B burn-in iterations and thin every K-th iteration.

5.3 Label Switching

When the main goal is identifying/interpreting the mixture components and/or clustering, then *label switching* should be treated. It is easy to see that $\mathrm{P}\left(\underline{X} \,|\, \underline{\gamma}\right) = \mathrm{P}\left(\underline{X} \,|\, \underline{\tilde{\gamma}}\right)$ where $\underline{\tilde{\gamma}} = \left(\underline{\theta}_{j_1}, \ldots, \underline{\theta}_{j_G}, \underline{\pi}_{j_1}, \ldots, \underline{\pi}_{j_G}\right)$ 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 a mixture distribution to relabelling the components: by interchanging the order of the components, the distributions induced by $\underline{\gamma}$ and $\underline{\tilde{\gamma}}$ are the same, although evidently the two parameters are distinct. For the finite mixture distribution as defined above with G components, there exist $G!$ equivalent ways of arranging them.

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.

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

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`.

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