Bayesian Factor Analysis

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*.

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

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. $\underline{\varepsilon}_i \sim \mathcal{N}\left(0, \underline{\Psi}\right)$ where $\underline{\Psi} = diag\left(\psi_1^2, \ldots, \psi_p^2\right)$

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

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

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

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

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

2 Bayesian Framework

2.1 Likelihood

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

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

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

The density of the data is then given by:

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

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

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

2.2 Posterior Set-Up

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

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

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

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 (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 - $\underline{\mathrm{f}}_i$

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

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

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

As this is the product of two $\mathcal{MVN}$ distributions we can expect the result to also be $\mathcal{MVN}$. Typically,

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

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

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

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

- Calculate $\underline{\Omega}_F = \left(\mathcal{I}_q + \underline{\Lambda}^T \underline{\Psi}^{-1} \underline{\Lambda}\right)^{-1}$
- Simulate at each iteration from an $\mathcal{MVN}_q\left(0, \underline{\Omega}_F\right)$ distribution instead.
- Then add on the mean of $\underline{\Omega}_F \underline{\Lambda}^T \underline{\Psi}^{-1} \left(\underline{X}_i - \underline{\mu}\right)$ (13)

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

A Gaussian distribution is a conjugate prior for $\underline{\Lambda}$, implying an $\mathcal{MVN}_q$ distribution prior for each row $\underline{\Lambda}_j$ of $\underline{\Lambda}$ s.t. $\underline{\Lambda}_j \sim \mathcal{MVN}_q\left(0, \Sigma_\lambda \mathcal{I}_q\right)$ 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 $\mathcal{MVN}_q$ distributions to itself be distributed in the same way.

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

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

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

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

- Calculate $\underline{\Omega}_{\lambda_j} = \left[\left(\Sigma_\lambda \mathcal{I}_q\right)^{-1} + \Psi_{jj}^{-1} \underline{F}^T \underline{F}\right]^{-1}$
- Simulate at each iteration from an $\mathcal{MVN}_q\left(0, \underline{\Omega}_{\lambda_j}\right)$ distribution instead.
- Then add on the mean of $\underline{\Omega}_{\lambda_j} \underline{F}^T \Psi_{jj}^{-1} \left(\underline{X}^j - \mu_j\right)$ (16)

3.3 Uniquenesses - $\underline{\Psi}$

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

$$\begin{aligned}
\mathrm{P}\left(\underline{\Psi}\right) &\sim \mathcal{W}^{-1}\left(\underline{\mathcal{S}}_\Psi, \nu\right) \\
&\propto \left|\underline{\Psi}^{-1}\right|^{\frac{\nu+p+1}{2}} \exp\left(-\frac{1}{2} \mathrm{tr}\left(\underline{\mathcal{S}}_\Psi \underline{\Psi}^{-1}\right)\right)
\end{aligned}$$

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

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

This suggests the prior for $\underline{\Psi}$ is a product of p $\mathcal{IG}\left({}^{\alpha_\psi}/_2, {}^{\beta_\psi}/_2\right)$ distributions.

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

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

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

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

$$\mathrm{P}\left(\Psi_{jj}|\underline{X}^\star, \underline{F}, \underline{\Lambda}\right) \sim \mathcal{IG}\left(\frac{n+\alpha_\psi}{2}, \frac{S_{jj}^2 + \beta_\psi}{2}\right) \qquad (18)$$

3.4 Reintroducing $\underline{\mu}$

We have already seen from (13), (16) and (18) that reintroducing the mean to the other parameters' full conditionals is trivial. All that remains is to specify the prior and derive the full conditional for $\underline{\mu}$ itself. A Gaussian distribution is a conjugate prior for $\underline{\mu}$, implying an $\mathcal{MVN}_p$ distribution prior s.t. $\underline{\mu} \sim \mathcal{MVN}_p\left(0, \Sigma_\mu\mathcal{I}_p\right)$ 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 $\mathcal{MVN}_p$ distributions to itself be distributed in the same way.

$$\begin{aligned}
\mathrm{P}\left(\underline{\mu}|\underline{X}, \underline{F}, \underline{\Psi}, \Lambda\right) &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\underline{X}_i - \underline{\mu} - \underline{\Lambda}\underline{\mathrm{f}}_i\right)^T\underline{\Psi}^{-1}\left(\underline{X}_i - \underline{\mu} - \underline{\Lambda}\underline{\mathrm{f}}_i\right)\right) \exp\left(-\frac{1}{2}\left(\underline{\mu}^T\left(\Sigma_\mu\mathcal{I}_p\right)^{-1}\underline{\mu}\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left[-2\underline{X}_i^T\Psi^{-1}\underline{\mu} + 2\left(\underline{\Lambda}\underline{\mathrm{f}}_i\right)^T\underline{\Psi}^{-1}\underline{\mu} + \underline{\mu}^T\underline{\Psi}^{-1}\underline{\mu} + \underline{\mu}^T\left(\Sigma_\mu\mathcal{I}_p\right)^{-1}\underline{\mu}\right]\right) \\
&\propto \exp\left(\sum_{i=1}^{n}\underline{X}_i^T\underline{\Psi}^{-1}\underline{\mu} - \sum_{i=1}^{n}\left(\underline{\Lambda}\underline{\mathrm{f}}_i\right)^T\underline{\Psi}^{-1}\underline{\mu} - \frac{1}{2}\left[\underline{\mu}^T\left(\left(\Sigma_\mu\mathcal{I}_p\right)^{-1} + n\underline{\Psi}^{-1}\right)\underline{\mu}\right]\right) \qquad (19) \\
\therefore \mathrm{P}\left(\underline{\mu}|\underline{X}, \underline{F}, \underline{\Psi}, \underline{\Lambda}\right) &\sim \mathcal{MVN}_p\Bigg(\left[\left(\Sigma_\mu\mathcal{I}_p\right)^{-1} + n\underline{\Psi}^{-1}\right]^{-1}\underline{\Psi}^{-1}\left(\sum_{i=1}^{n}\underline{X}_i^T - \sum_{i=1}^{n}\left(\underline{\Lambda}\underline{\mathrm{f}}_i\right)^T\right)^T, \\
&\qquad \left[\left(\Sigma_\mu\mathcal{I}_p\right)^{-1} + n\underline{\Psi}^{-1}\right]^{-1}\Bigg) \qquad (20)
\end{aligned}$$

However, we can save on computational time if we:

- Calculate $\underline{\Omega}_\mu = \left[\left(\Sigma_\mu\mathcal{I}_p\right)^{-1} + n\underline{\Psi}^{-1}\right]^{-1}$
- Simulate at each iteration from an $\mathcal{MVN}_p\left(0, \underline{\Omega}_\mu\right)$ distribution instead.
- Then add on the mean of $\underline{\Omega}_\mu\underline{\Psi}^{-1}\left(\sum_{i=1}^{n}\underline{X}_i^T - \sum_{i=1}^{n}\left(\underline{\Lambda}\underline{\mathrm{f}}_i\right)^T\right)^T$ (21)

3.5 Gibbs Sampler Pseudo-Code

i) Choose scalar hyperparameters $\Sigma_\mu, \Sigma_\lambda, \alpha_\psi,$ and β_ψ, and select q

ii) Initalise:
$$\begin{aligned}
\underline{\mu}^{(0)} &\sim \mathcal{MVN}_p\left(0, \Sigma_\mu \mathcal{I}_p\right) \\
\underline{F}^{(0)} &\sim \mathcal{MVN}_q\left(n, 0, \mathcal{I}_q\right) \\
\underline{\Lambda}^{(0)} &\sim \mathcal{MVN}_q\left(n, 0, \Sigma_\lambda \mathcal{I}_q\right) \\
\underline{\Psi}^{(0)} &\sim \mathcal{IG}\left(p, {}^{\alpha_\psi}/_2, {}^{\beta_\psi}/_2\right)
\end{aligned}$$

iii) For $t = 1, \ldots,$ n.iters

$$\begin{aligned}
\text{a)} \quad \underline{\Omega}_\mu^{(t)} &= \left[\left(\Sigma_\mu \mathcal{I}_p\right)^{-1} + n\underline{\Psi}^{(t-1)^{-1}}\right]^{-1} \\
\text{b)} \quad \underline{\Omega}_F^{(t)} &= \left(\mathcal{I}_q + \underline{\Lambda}^{(t-1)^T}\underline{\Psi}^{(t-1)^{-1}}\underline{\Lambda}^{(t-1)}\right)^{-1} \\
\text{c)} \quad \underline{\mu}^{(t)} &\sim \mathcal{MVN}_p\left(0, \underline{\Omega}_\mu^{(t)}\right) + \underline{\Omega}_\mu^{(t)}\underline{\Psi}^{(t-1)^{-1}}\left(\sum_{i=1}^n \underline{X}_i^T - \sum_{i=1}^n \left(\underline{\Lambda}^{(t-1)}\underline{\mathrm{f}}_i^{(t-1)}\right)^T\right)^T \\
\text{d)} \quad \text{For } i &= 1, \ldots, n \\
\bullet\ \underline{\mathrm{f}}_i^{(t)} &\sim \mathcal{MVN}_q\left(0, \underline{\Omega}_F^{(t)}\right) + \underline{\Omega}_F^{(t)}\underline{\Lambda}^{(t-1)^T}\underline{\Psi}^{(t-1)^{-1}}\left(\underline{X}_i - \underline{\mu}^{(t)}\right) \\
\text{e)} \quad \text{For } j &= 1, \ldots, p \\
\bullet\ \underline{\Omega}_{\lambda_j}^{(t)} &= \left[\left(\Sigma_\lambda \mathcal{I}_q\right)^{-1} + \Psi_{jj}^{(t-1)^{-1}}\underline{F}^{(t)^T}\underline{F}^{(t)}\right]^{-1} \\
\bullet\ \underline{\Lambda}_j^{(t)} &\sim \mathcal{MVN}_q\left(0, \underline{\Omega}_{\lambda_j}^{(t)}\right) + \underline{\Omega}_{\lambda_j}^{(t)}\underline{F}^{(t)^T}\Psi_{jj}^{(t-1)^{-1}}\left(\underline{X}^j - \underline{\mu}_j^{(t)}\right) \\
\bullet\ \Psi_{jj}^{(t)} &\sim \mathcal{IG}\left(\frac{n+\alpha_\psi}{2}, \frac{S_{jj}^{(t)^2} + \beta_\psi}{2}\right)
\end{aligned}$$

iv) Disregard the first $\mathcal{B}$ burn-in iterations and thin every $\mathcal{T}$-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 $\mathrm{E}\left(\underline{\mathrm{f}}^\star\right) = 0$ 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 letting each iteration's scores matrix equal the transpose of the product of the transpose of that iteration's score matrix *pre-multiplied* by the transpose of 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 accuracte 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:

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

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 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. 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.

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

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

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

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

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

- Calculate $\underline{\Omega}_{\lambda_j} = \left[\underline{\mathrm{D}}_j^{-1}+\Psi_{jj}^{-1}\underline{F}^T\underline{F}\right]^{-1}$
- Simulate at each iteration from an $\mathcal{MVN}_{q^\star}\left(0,\underline{\Omega}_{\lambda_j}\right)$ distribution instead.
- Then add on the mean of $\underline{\Omega}_{\lambda_j}\underline{F}^T\Psi_{jj}^{-1}\left(\underline{X}^j-\mu_j\right)$

4.2.2 Local Shrinkage – ϕ_{jk}

Using the conditional prior in (22) and the prior for ϕ in (23) we can derive the full conditional for the local shrinkage paramter 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:

$$\mathrm{P}\left(\phi_{jk}|—\right) \sim \mathcal{G}\left(\frac{\nu+1}{2},\frac{\nu+\tau_k\lambda_{jk}^2}{2}\right) \tag{26}$$

4.2.3 Global Shrinkage – τ_k

4.3 Adaptive Step

4.4 Adaptive Gibbs Sampler Pseudo-Code

5 References

A. Bhattacharya & D.B. Dunson. Sparse bayesian infinite factor models. *Biometrika*, **98**, 2: *pp.* 291–306, 2011.