

# Bayesian Factor Analysis

## Notes & Derivations

Keefe Murphy<sup>1, 2</sup>, Dr. Claire Gormley<sup>1, 2</sup>, and Prof. Brendan Murphy<sup>1, 2</sup>

<sup>1</sup>Department of Mathematics and Statistics, UCD

<sup>2</sup>Insight Centre for Data Analytics, UCD

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Model Set-Up . . . . .	3
1.2	Assumptions . . . . .	3
<b>2</b>	<b>Bayesian Framework</b>	<b>4</b>
2.1	Likelihood . . . . .	4
2.2	Posterior Set-Up . . . . .	5
<b>3</b>	<b>Sampling from the Full Conditionals</b>	<b>5</b>
3.1	Factor Scores . . . . .	5
3.2	Loadings Matrix . . . . .	6
3.3	Uniquenesses . . . . .	6
3.4	Reintroducing $\mu$ . . . . .	7
3.5	Gibbs Sampler Pseudo-Code . . . . .	8
3.6	Issues Around Identifiability . . . . .	8
<b>4</b>	<b>Introducing the Shrinkage Prior</b>	<b>9</b>

# 1 Introduction

## 1.1 Model Set-Up

Let  $\underline{x} = (x_1, x_2, \dots, x_p)^T$  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  $f_1, f_2, \dots, f_q$  called *common factors* and  $p$  additional sources of variation  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p$  called *specific factors*.

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

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
	$\underline{f}_i$	$\rightarrow$	$(q \times 1)$	vector of factor scores for obs $i$
	$\underline{\varepsilon}_i$	$\rightarrow$	$(p \times 1)$	vector of errors for obs $i$
	$i$	$=$	$1, \dots, n$	
	$j$	$=$	$1, \dots, p$	
	$k$	$=$	$1, \dots, q$	

$\Lambda_{jk}$  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

$$\left( \underline{x}_i - \underline{\mu} \right)_{(p \times 1)} = \underline{x}_{i(p \times 1)}^* = \underline{\Lambda}_{(p \times q)} \underline{f}_{i(q \times 1)} + \underline{\varepsilon}_{i(p \times 1)} \quad (1)$$

## 1.2 Assumptions

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

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

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

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

$$\therefore \text{E}(\underline{f}) = 0 \text{ and } \text{Cov}(\underline{f}) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} = \underline{\mathcal{I}}$$

$$\therefore \underline{f}_i \sim \mathcal{MVN}_q(0, \underline{\mathcal{I}}) \quad (3)$$

## 2 Bayesian Framework

### 2.1 Likelihood

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

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

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

The density of the data is then given by:

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

$$\begin{aligned}
\text{Where } \underline{\Lambda}_{(p \times q)} &= \begin{pmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1q} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \dots & \lambda_{pq} \end{pmatrix} \\
&\& \underline{F}_{(n \times q)} = \begin{pmatrix} f_{11} & f_{12} & \dots & f_{1q} \\ f_{21} & f_{22} & \dots & f_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \dots & f_{nq} \end{pmatrix} \& \underline{f}_i \text{ is a column vector containing the entries of row } i \text{ of } F
\end{aligned}$$

## 2.2 Posterior Set-Up

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

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

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

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 necessitate multiplying the quantity in (9) by  $\left[ \prod_{j=1}^p P(\mu_j) \right]$ . 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 - $\underline{f}_i$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

### 3.3 Uniquenesses - $\underline{\Psi}$

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

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

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

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

This suggests the prior for  $\underline{\Psi}$  is a product of  $p$   $\mathcal{IG}(\alpha/2, \beta/2)$  distributions.

$$\begin{aligned}
\therefore P(\underline{\Psi} \mid \alpha, \beta) &= \prod_{j=1}^p P(\Psi_{jj} \mid \alpha, \beta) \\
&\propto \prod_{j=1}^p (\Psi_{jj})^{-(\frac{\alpha}{2}-1)} \exp\left(-\frac{\beta}{2}\Psi_{jj}\right) \\
\therefore P(\underline{\Psi} \mid \underline{X}^*, \underline{F}, \underline{\Lambda}) &\propto P(\underline{X}^* \mid \underline{F}, \underline{\Lambda}) P(\underline{\Psi} \mid \alpha, \beta) \\
&\propto \prod_{j=1}^p (\Psi_{jj})^{-\frac{n}{2}} \exp\left(-\frac{1}{2} \sum_{i=1}^n (\underline{x}_i^* - \underline{\Lambda}_j \underline{f}_i)^T \Psi_{jj}^{-1} (\underline{x}_i^* - \underline{\Lambda}_j \underline{f}_i)\right) \prod_{j=1}^p (\Psi_{jj})^{-(\frac{\alpha}{2}-1)} \exp\left(-\frac{\beta}{2}\Psi_{jj}\right) \\
&\propto \prod_{j=1}^p \Psi_{jj}^{-(\frac{n+\alpha}{2}+1)} \exp\left(-\frac{\mathcal{S}_{jj}^{2*} + \beta}{2} \Psi_{jj}^{-1}\right) \tag{17}
\end{aligned}$$

$$\text{where } \mathcal{S}_{jj}^{2*} = \sum_{i=1}^n (x_{ij} - \underline{\Lambda}_j \underline{f}_i)^T (x_{ij} - \underline{\Lambda}_j \underline{f}_i)$$

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

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

Thus the posterior distribution of each  $\Psi_{jj}$  is given by:

$$P(\Psi_{jj} \mid \underline{X}^*, \underline{F}, \underline{\Lambda}) \sim \mathcal{IG}\left(\frac{n+\alpha}{2}, \frac{\mathcal{S}_{jj}^2 + \beta}{2}\right) \tag{18}$$

### 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.  $\underline{\mu} \sim \mathcal{MVN}_p(0, \Sigma_\mu \mathcal{I}_p)$  where  $\Sigma_\mu$  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.

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

$$\begin{aligned}
\therefore P(\underline{\mu} \mid \underline{X}, \underline{F}, \underline{\Psi}, \underline{\Lambda}) &\sim \mathcal{MVN}_p\left([\Sigma_\mu \mathcal{I}_p]^{-1} + n \underline{\Psi}^{-1}\right)^{-1} \underline{\Psi}^{-1} \left(\sum_{i=1}^n \underline{X}_i^T - \sum_{i=1}^n (\underline{\Lambda} \underline{f}_i)^T\right)^T, \\
&\quad \left([\Sigma_\mu \mathcal{I}_p]^{-1} + n \underline{\Psi}^{-1}\right)^{-1} \tag{20}
\end{aligned}$$

However, we can save on computational time if we:

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

### 3.5 Gibbs Sampler Pseudo-Code

i) Choose scalar hyperparameters  $\Sigma_\mu, \Sigma_\Lambda, \alpha$ , and  $\beta$ , and select  $q$

ii) Initialise:

$$\begin{aligned}\underline{\mu}^{(0)} &\sim \mathcal{MVN}_q(0, \Sigma_\mu \mathcal{I}_p) \\ \underline{F}^{(0)} &\sim \mathcal{MVN}_q(n, 0, \mathcal{I}_q) \\ \underline{\Lambda}^{(0)} &\sim \mathcal{MVN}_q(n, 0, \Sigma_\Lambda \mathcal{I}_q) \\ \underline{\Psi}^{(0)} &\sim \mathcal{IG}(p, \alpha/2, \beta/2)\end{aligned}$$

iii) For  $t = 1, \dots, \text{n.iters}$

- a)  $\underline{\Omega}_\mu^{(t)} = \left[ (\Sigma_\mu \mathcal{I}_p)^{-1} + n \underline{\Psi}^{(t-1)^{-1}} \right]^{-1}$
- b)  $\underline{\Omega}_F^{(t)} = \left( \mathcal{I} + \underline{\Lambda}^{(t-1)^T} \underline{\Psi}^{(t-1)^{-1}} \underline{\Lambda}^{(t-1)} \right)^{-1}$
- c)  $\underline{\mu}^{(t)} \sim \mathcal{MVN}_p \left( 0, \underline{\Omega}_\mu^{(t)} \right) + \underline{\Omega}_\mu^{(t)} \underline{\Psi}^{(t-1)^{-1}} \left( \sum_{i=1}^n \underline{X}_i^T - \sum_{i=1}^n \left( \underline{\Lambda}^{(t-1)} \underline{f}_i^{(t-1)} \right)^T \right)^T$
- d) For  $i = 1, \dots, n$ 
  - $\underline{f}_i^{(t)} \sim \mathcal{MVN}_q \left( 0, \underline{\Omega}_F^{(t)} \right) + \underline{\Omega}_F^{(t)} \underline{\Lambda}^{(t-1)^T} \underline{\Psi}^{(t-1)^{-1}} (\underline{X}_i - \underline{\mu}^{(t)})$
- e) For  $j = 1, \dots, p$ 
  - $\underline{\Omega}_{\lambda_j}^{(t)} = \left[ (\Sigma_\lambda \mathcal{I}_q)^{-1} + \underline{\Psi}_{jj}^{(t-1)^{-1}} \underline{F}^{(t)^T} \underline{F}^{(t)} \right]^{-1}$
  - $\underline{\Lambda}_j^{(t)} \sim \mathcal{MVN}_q \left( 0, \underline{\Omega}_{\lambda_j}^{(t)} \right) + \underline{\Omega}_{\lambda_j}^{(t)} \underline{F}^{(t)^T} \underline{\Psi}_{jj}^{(t-1)^{-1}} (\underline{X}^j - \underline{\mu}_j^{(t)})$
  - $\underline{\Psi}_{jj}^{(t)} \sim \mathcal{IG} \left( \frac{n + \alpha}{2}, \frac{S_{jj}^{(t)^2} + \beta}{2} \right)$

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  $\underline{T} \underline{T}^T = \mathcal{I}$ . Then :

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

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

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

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 Procrustean 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