Bayesian Inference for Sparse Factor Models

Yong See Foo

January 2020

1 Model

Suppose that we have a matrix $\mathbf{Y} \in \mathbb{R}^{G \times N}$ of observed gene expressions, where G is the number of genes, and N is the number of individuals. We wish to model gene expression as a weighted sum of K transcription factor activities: $y_{ij} = \sum_{k=1}^{K} l_{ik} f_{kj} + e_{ij}$, where l_{ik} is the regulatory weight of factor k on gene i, f_{kj} is the activation of factor k for individual i, and e_{ij} accounts for any corresponding residual noise. In matrix notation, the model is formulated as $\mathbf{Y} = \mathbf{LF} + \mathbf{E}$. By assuming that the noise is independently distributed and follows a Gaussian distribution with gene-specific variance, the distribution of the gene expression $\mathbf{Y}$ can be defined as

$$p(\boldsymbol{y}_{i\cdot} \mid \mathbf{L}, \mathbf{F}, \boldsymbol{\tau}) = \mathcal{N}\Big(\boldsymbol{y}_{i\cdot} \;\Big|\; \mathbf{F}^\top \boldsymbol{l}_{i\cdot}, \tau_i^{-1}\mathbf{I}\Big),$$

where $\boldsymbol{y}_{i\cdot}$ and $\boldsymbol{l}_{i\cdot}$ are column vectors indicating the ith row of $\mathbf{Y}$ and $\mathbf{L}$ repsectively, and τ_i is the precision of Gaussian noise. This may also be written as $p(\boldsymbol{y}_{i\cdot} \mid \mathbf{L}, \mathbf{F}, \boldsymbol{\tau}) = \mathcal{N}\big(\boldsymbol{y}_{\cdot j} \mid \mathbf{L}\boldsymbol{f}_{\cdot j}, D_{\boldsymbol{\tau}}^{-1}\big)$, where $\boldsymbol{y}_{\cdot j}$ indicates the jth column of $\mathbf{Y}$, and $D_{\boldsymbol{v}} = \operatorname{diag}(\boldsymbol{v})$ for any vector $\boldsymbol{v}$.

As only a small subset of genes are regulated by each transcription factor, the loading matrix $\mathbf{L}$ is known to be sparse. This is encoded with the following prior:

$$p(l_{ik} \mid z_{ik}, \alpha_k) = \begin{cases} \delta(l_{ik}) & \text{if } z_{ik} = 0 \\ \mathcal{N}\big(l_{ik} \mid 0, \alpha_k^{-1}\big) & \text{if } z_{ik} = 1 \end{cases}$$

where $z_{ik} = 0$ if gene i is not regulated by transcription factor k, otherwise l_{ik} follows a Gaussian distribution with factor-specific precision α_k. A connectivity matrix $\mathbf{Z}$ stores the latent binary variables z_{ik}, and we define a Bernoulli prior for each of its elements:

$$p(z_{ik}) = \operatorname{Bern}(z_{ik} \mid \pi_k),$$

where π_k are hyperparameters which control the sparsity of each factor.

To avoid identifiability issues caused by scaling, we define a unit Gaussian prior distribution for the factor matrix $\mathbf{F}$:

$$p\big(\boldsymbol{f}_{\cdot j}\big) = \mathcal{N}\big(\boldsymbol{f}_{\cdot j} \mid \mathbf{0}, \mathbf{I}\big).$$

Lastly, a gamma prior is defined for each of the precision parameters:

$$p(\tau_i) = \Gamma(\tau_i \mid a_\tau, b_\tau)$$
$$p(\alpha_k) = \Gamma(\alpha_k \mid a_\alpha, b_\alpha),$$

where $a_\tau, b_\tau, a_\alpha, b_\alpha$ are hyperparameters to be specified.

2 MCMC

2.1 Collapsed Gibbs sampling

We use collapsed Gibbs sampling to simulate the posterior, where the regulatory weights $\mathbf{L}$ are marginalised out when computing the conditional distribution of the connectivity matrix $\mathbf{Z}$. This results in a sampler that is more efficient than a vanilla Gibbs sampler, as the autocorrelation between samples of $\mathbf{Z}$ is reduced. We need to ensure that $l_{ik} = 0$ whenever $z_{ik} = 0$. This can achieved by introducing modifications to the conditional distribution of $\mathbf{Y}$. We have

$$\begin{aligned}
p(\boldsymbol{l}_{i\cdot}, \boldsymbol{z}_{i\cdot} \mid \mathbf{Y}, \mathbf{F}, \boldsymbol{\tau}, \boldsymbol{\alpha}) \propto & \prod_{k\,:\,z_{ik}=1} \pi_k \sqrt{\frac{\alpha_k}{2\pi}} \times \prod_{k\,:\,z_{ik}=0} (1-\pi_k)\,\delta(l_{ik}) \\
& \times \exp\Bigg\{ -\frac{\tau_i}{2} \left(\boldsymbol{y}_{i\cdot} - [\mathbf{F}]_{\boldsymbol{z}_{i\cdot}}^{\top} [\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}} \right)^{\top} \left(\boldsymbol{y}_{i\cdot} - [\mathbf{F}]_{\boldsymbol{z}_{i\cdot}}^{\top} [\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}} \right) \\
& \qquad - \frac{1}{2} [\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}^{\top} [D_{\boldsymbol{\alpha}}]_{\boldsymbol{z}_{i\cdot}} [\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}} \Bigg\} \\
\propto & \prod_{k\,:\,z_{ik}=1} \pi_k \sqrt{\frac{\alpha_k}{2\pi}} \times \prod_{k\,:\,z_{ik}=0} (1-\pi_k)\,\delta(l_{ik}) \\
& \times \exp\left\{ -\frac{1}{2} \left([\boldsymbol{l}]_{\boldsymbol{z}_{i\cdot}} - \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}} \right)^{\top} \Sigma_{\boldsymbol{l}_{i\cdot}}^{-1} \left([\boldsymbol{l}]_{\boldsymbol{z}_{i\cdot}} - \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}} \right) + \frac{1}{2} \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}}^{\top} \Sigma_{\boldsymbol{l}_{i\cdot}}^{-1} \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}} \right\}
\end{aligned} \tag{1}$$

where

$$\begin{aligned}
[\mathbf{F}]_{\boldsymbol{z}_{i\cdot}} &= \text{matrix consisting of rows of } \mathbf{F} \text{ whose corresponding entries of } \boldsymbol{z}_{i\cdot} \text{ are equal to 1} \\
[\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}} &= \text{vector consisting of entries of } \boldsymbol{l}_{i\cdot} \text{ whose corresponding entries of } \boldsymbol{z}_{i\cdot} \text{ are equal to 1} \\
[D_{\boldsymbol{\alpha}}]_{\boldsymbol{z}_{i\cdot}} &= \text{matrix consisting of rows of } D_{\boldsymbol{\alpha}} \text{ whose corresponding entries of } \boldsymbol{z}_{i\cdot} \text{ are equal to 1} \\
\Sigma_{\boldsymbol{l}_{i\cdot}} &= \left(\tau_i [\mathbf{F}]_{\boldsymbol{z}_{i\cdot}} [\mathbf{F}]_{\boldsymbol{z}_{i\cdot}}^{\top} + [D_{\boldsymbol{\alpha}}]_{\boldsymbol{z}_{i\cdot}} \right)^{-1} \\
\boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}} &= \tau_i \Sigma_{\boldsymbol{l}_{i\cdot}} [\mathbf{F}]_{\boldsymbol{z}_{i\cdot}} \boldsymbol{y}_{i\cdot},
\end{aligned}$$

and hence obtain the full conditional distribution of $\boldsymbol{l}_{i\cdot}$:

$$p\big([\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}} \mid \mathbf{Y}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha}\big) = \mathcal{N}\big([\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}} \mid \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}}, \Sigma_{\boldsymbol{l}_{i\cdot}}\big) \times \prod_{k\,:\,z_{ik}=0} \delta(l_{ik}). \tag{2}$$

Marginalising out $\boldsymbol{l}_{i\cdot}$ from Equation 1 gives a conditional distribution of z_{ik}:

$$p(z_{ik} \mid \mathbf{Y}, \mathbf{F}, \mathbf{Z}_{-ik}, \boldsymbol{\tau}, \boldsymbol{\alpha}) \propto \left(\frac{\alpha_k}{2\pi}\right)^{\frac{z_{ik}}{2}} \det|\Sigma_{\boldsymbol{l}_{i\cdot}}|^{\frac{1}{2}} \exp\left\{ \frac{1}{2} \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}}^{\top} \Sigma_{\boldsymbol{l}_{i\cdot}}^{-1} \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}} \right\} \pi_k^{z_{ik}} (1-\pi_k)^{1-z_{ik}}. \tag{3}$$

We also have

$$\begin{aligned}
p\big(\boldsymbol{f}_{\cdot j} \mid \mathbf{Y}, \mathbf{L}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha}\big) &\propto \exp\left\{ -\frac{1}{2} \left(\boldsymbol{y}_{\cdot j} - \mathbf{L}\boldsymbol{f}_{\cdot j} \right)^{\top} D_{\boldsymbol{\tau}} \left(\boldsymbol{y}_{\cdot j} - \mathbf{L}\boldsymbol{f}_{\cdot j} \right) - \frac{1}{2} \boldsymbol{f}_{\cdot j}^{\top} \boldsymbol{f}_{\cdot j} \right\} \\
&\propto \exp\left\{ -\frac{1}{2} \left(\boldsymbol{f}_{\cdot j} - \boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}} \right)^{\top} \Sigma_{\boldsymbol{f}_{\cdot j}}^{-1} \left(\boldsymbol{f}_{\cdot j} - \boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}} \right) \right\}
\end{aligned}$$

where

$$\begin{aligned}
\Sigma_{\boldsymbol{f}_{\cdot j}} &= \left(\mathbf{L}^{\top} D_{\boldsymbol{\tau}} \mathbf{L} + \mathbf{I} \right)^{-1} \\
\boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}} &= \Sigma_{\boldsymbol{f}_{\cdot j}} \mathbf{L}^{\top} D_{\boldsymbol{\tau}} \boldsymbol{y}_{\cdot j},
\end{aligned}$$

thus arriving at the full conditional distribution of $\boldsymbol{f}_{\cdot j}$:

$$p\big(\boldsymbol{f}_{\cdot j} \bigm| \mathbf{Y}, \mathbf{L}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha}\big) = \mathcal{N}\Big(\boldsymbol{f}_{\cdot j} \Bigm| \boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}}, \Sigma_{\boldsymbol{f}_{\cdot j}}\Big). \tag{4}$$

Lastly, we have the full conditional distribution of τ_i:

$$p(\tau_i \mid \mathbf{Y}, \mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\alpha}) = \Gamma\left(\tau_i \,\middle|\, a_\tau + \frac{N}{2}, b_\tau + \frac{1}{2}\Big(\boldsymbol{y}_{i\cdot} - [\mathbf{F}]^{\mathsf{T}}_{\boldsymbol{z}_{i\cdot}} \, [\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}\Big)^{\mathsf{T}} \Big(\boldsymbol{y}_{i\cdot} - [\mathbf{F}]^{\mathsf{T}}_{\boldsymbol{z}_{i\cdot}} \, [\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}\Big)\right), \tag{5}$$

and the full conditional distribution of α_k:

$$p(\alpha_k \mid \mathbf{Y}, \mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\tau}) = \Gamma\left(\alpha_k \,\middle|\, a_\alpha + \frac{1}{2}\sum_{i=1}^{G} z_{ik}, b_\alpha + \frac{1}{2}\sum_{i\,:\,z_{ik}=1} l_{ik}^2\right). \tag{6}$$

2.2 Empirical Bayes for estimating sparsity hyperparameters $\boldsymbol{\pi}$

It turns out that inference is sensitive to the sparsity hyperparameters $\boldsymbol{\pi}$, resulting in poor inference in the case of incorrectly specified values. In some applications, information about the sparsity of factors may not be available. Thus, we incorporate an Empirical Bayes approach into the Gibbs sampler, as an attempt to learn the sparsity hyperparameters $\boldsymbol{\pi}$ from the data.

2.3 Identifiability issues

As the priors are exchangeable, this results the model being non-identifiabile. Given a mode of the posterior distribution, if the latent factors are permuted, or if the sign of the entries corresponding to some factor are all switched, one will obtain another equivalent mode. These symmetries result in $2^K K!$ equivalent modes in the posterior, a subset of which is explored by the sampler. This causes most posterior summaries (e.g. posterior mean) to be of little utility.

A relabelling algorithm, similar to that of Erosheva and Curtis (2017), is used to deal with these issues of label switching and sign switching. Following the method of Stephens (2000), a decision-theoretic approach is to define a loss function for a set of actions and relabellings, and select the action and relabelling which minimises the posterior expected loss. This is done with the aim of relabelling samples such that they correspond to being sampled around the same mode.

Define an action

$$\mathbf{a} = \left(\{m_{l_{ik}}\}_{\substack{i=1:G\\k=1:K}}, \{s^2_{l_{ik}}\}_{\substack{i=1:G\\k=1:K}}, \{m_{f_{kj}}\}_{\substack{j=1:N\\k=1:K}}, \left\{s^2_{f_{kj}}\right\}_{\substack{j=1:N\\k=1:K}}, \{p_{z_{ik}}\}_{\substack{i=1:G\\k=1:K}}\right)$$

to be a choice of means and variances of the entries of $\mathbf{L}$ and $\mathbf{F}$, and also the means of the entries of $\mathbf{Z}$. Let $\sigma \in S_K$ and $\boldsymbol{\nu} \in \{-1, 1\}^K$, where S_K is the set of permutations on the set $\{1, 2, \ldots, K\}$. We define a loss function as follows:

$$\begin{aligned}\mathcal{L}(\mathbf{a}, \sigma, \boldsymbol{\nu}; \mathbf{L}, \mathbf{F}, \mathbf{Z}) = -\sum_{k=1}^{K}\Bigg\{&\sum_{i=1}^{G} \log \mathrm{Bern}(z_{i\sigma(k)} \mid p_{z_{ik}}) + \mathbb{I}\big(l_{i\sigma(k)} \neq 0\big) \log \mathcal{N}(\nu_k l_{i\sigma(k)} \mid m_{l_{ik}}, s^2_{l_{ik}}) \\ &+ \sum_{j=1}^{N} \mathcal{N}(\nu_k f_{\sigma(k)j} \mid m_{f_{kj}}, s^2_{f_{kj}})\Bigg\}.\end{aligned} \tag{7}$$

Suppose we want to relabel T samples of $\left\{\left(\mathbf{L}^{(t)}, \mathbf{F}^{(t)}, \mathbf{Z}^{(t)}\right)\right\}_{t=1:T}$ obtained from MCMC. We seek to choose $\mathbf{a}$ and $\left\{\left(\sigma^{(t)}, \boldsymbol{\nu}^{(t)}\right)\right\}_{t=1:T}$ such that the Monte Carlo risk

$$\sum_{t=1}^{T} \mathcal{L}(\mathbf{a}, \left\{\left(\sigma^{(t)}, \boldsymbol{\nu}^{(t)}\right)\right\}_{t=1:T}; \mathbf{L}^{(t)}, \mathbf{F}^{(t)}, \mathbf{Z}^{(t)})$$

is minimised. After initialising $\mathbf{a}$ and $\left\{\left(\sigma^{(t)}, \boldsymbol{\nu}^{(t)}\right)\right\}_{t=1:T}$, a local optimum may be obtained by alternating between the following steps:

1. Choose $\boldsymbol{a}$ such that the Monte Carlo risk is minimised given the current values of $\left\{\left(\sigma^{(t)}, \boldsymbol{\nu}^{(t)}\right)\right\}_{t=1:T}$.
2. Choose $\left\{\left(\sigma^{(t)}, \boldsymbol{\nu}^{(t)}\right)\right\}_{t=1:T}$ such that the Monte Carlo risk is minimised given the current action $\boldsymbol{a}$.

The procedure may be terminated when a fixed point is reached. Step 1 may be solved analytically, whereas Step 2 is equivalent to the linear assignment problem, which may be solved by a $\mathcal{O}(K^3)$ algorithm of Jonker and Volgenant (1987).

3 Variational Inference

3.1 Mean-field approximation

Use the variational factorisation

$$q(\mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha}) = \prod_{i=1}^{G}\left[\prod_{k=1}^{K} q(l_{ik} \mid z_{ik}) q(z_{ik})\right] q(\tau_i) \times \prod_{j=1}^{N} q\big(\boldsymbol{f}_{\cdot j}\big) \times \prod_{k=1}^{K} q(\alpha_k) \tag{8}$$

as an approximation to the posterior distribution, where

$$\begin{aligned}
q(l_{ik} \mid z_{ik}) &= \mathcal{N}\big(l_{ik} \mid \mu_{l_{ik}}, \sigma^2_{l_{ik}}\big)^{z_{ik}} \times \delta(l_{ik})^{1-z_{ik}} \\
q(z_{ik}) &= \mathrm{Bern}(z_{ik} \mid \eta_{ik}) \\
q(\boldsymbol{f}_{\cdot j}) &= \mathcal{N}\Big(\boldsymbol{f}_{\cdot j} \;\Big|\; \boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}}, \Sigma_{\boldsymbol{f}_{\cdot j}}\Big) \\
q(\tau_i) &= \Gamma\Big(\tau_i \;\Big|\; \hat{a}_{\tau_i}, \hat{b}_{\tau_i}\Big) \\
q(\alpha_k) &= \Gamma\Big(\alpha_k \;\Big|\; \hat{a}_{\alpha_k}, \hat{b}_{\alpha_k}\Big).
\end{aligned}$$

Coordinate ascent for $\boldsymbol{l}_{i\cdot}$ and $\boldsymbol{z}_{i\cdot}$ gives

$$
\begin{aligned}
q^*(l_{ik}, z_{ik}) &\propto \exp\left\{\mathbb{E}_{\boldsymbol{l}_{-ik},\mathbf{F},\boldsymbol{z}_{-ik},,\tau_i,\boldsymbol{\alpha}}[\log p(l_{ik}, z_{ik} \mid \mathbf{Y}, \boldsymbol{l}_{-ik}, \mathbf{F}, \boldsymbol{z}_{-ik}, \boldsymbol{\tau}, \boldsymbol{\alpha})]\right\} \\
&\propto \exp\left\{ \frac{z_{ik}}{2}\mathbb{E}_{\alpha_k}\left[\log\frac{\alpha_k}{2\pi}\right] - \mathbb{E}_{\boldsymbol{l}_{-ik},\mathbf{F},\boldsymbol{z}_{-ik},\tau_i}\left[\frac{\tau_i}{2}\left(\boldsymbol{y}_{i\cdot} - [\mathbf{F}]^\top_{\boldsymbol{z}_{i\cdot}}[\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}\right)^\top \left(\boldsymbol{y}_{i\cdot} - [\mathbf{F}]^\top_{\boldsymbol{z}_{i\cdot}}[\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}\right)\right]\right. \\
&\qquad \left. - \frac{z_{ik}}{2}\mathbb{E}_{\boldsymbol{\alpha}}\left[\alpha_k l_{ik}^2\right]\right\} \times \prod_{k=1}^K \pi_k^{z_{ik}}(1-\pi_k)^{1-z_{ik}}, \\
&\propto \exp\left\{ -\frac{\hat{a}_{\tau_i}}{2\hat{b}_{\tau_i}}\mathbb{E}_{\boldsymbol{l}_{-ik},\mathbf{F},\boldsymbol{z}_{-ik}}\left[-2z_{ik}\boldsymbol{y}_{i\cdot}^\top\boldsymbol{f}_k l_{ik} + 2z_{ik}\boldsymbol{f}_k^\top\sum_{k'\neq k} z_{ik'}\boldsymbol{f}_{k'}l_{ik'}l_{ik} + z_{ik}\boldsymbol{f}_k^\top\boldsymbol{f}_k l_{ik}^2\right]\right. \\
&\qquad \left. + \frac{z_{ik}}{2}\left(\psi(\hat{a}_{\alpha_k}) - \log 2\pi\hat{b}_{\alpha_k} - \frac{\hat{a}_{\alpha_k}}{\hat{b}_{\alpha_k}}l_{ik}^2\right)\right\} \times \prod_{k=1}^K \pi_k^{z_{ik}}(1-\pi_k)^{1-z_{ik}} \\
&\propto \exp\left\{ -\frac{\hat{a}_{\tau_i}}{2\hat{b}_{\tau_i}}\left[-2z_{ik}\boldsymbol{y}_{i\cdot}^\top\boldsymbol{\mu}_{\boldsymbol{f}_k}l_{ik} + 2z_{ik}\boldsymbol{\mu}_{\boldsymbol{f}_k}^\top\sum_{k'\neq k}\eta_{ik'}\boldsymbol{\mu}_{\boldsymbol{f}_{k'}}\mu_{\boldsymbol{l}_{ik'}}l_{ik}\right.\right. \\
&\qquad \left.\left. + z_{ik}\sum_{j=1}^N\left(\left[\Sigma_{\boldsymbol{f}_{\cdot j}}\right]_{kk} + \left[\boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}}\right]_k^2\right)l_{ik}^2\right] + \frac{z_{ik}}{2}\left(\psi(\hat{a}_{\alpha_k}) - \log 2\pi\hat{b}_{\alpha_k} - \frac{\hat{a}_{\alpha_k}}{\hat{b}_{\alpha_k}}l_{ik}^2\right)\right\} \\
&\qquad \times \pi_k^{z_{ik}}(1-\pi_k)^{1-z_{ik}},
\end{aligned}
$$

which corresponds to the updates

$$
\begin{aligned}
\sigma^{2*}_{l_{ik}} &= \left(\frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}}\sum_{j=1}^N\left(\left[\Sigma_{\boldsymbol{f}_{\cdot j}}\right]_{kk} + \left[\boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}}\right]_k^2\right) + \frac{\hat{a}_{\alpha_k}}{\hat{b}_{\alpha_k}}\right)^{-1} \\
\mu^*_{l_{ik}} &= \frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}}\sigma^{2*}_{l_{ik}}\boldsymbol{\mu}_{\boldsymbol{f}_k}^\top\left(\boldsymbol{y}_{i\cdot} - \sum_{k'\neq k}\eta_{ik'}\boldsymbol{\mu}_{\boldsymbol{f}_{k'}}\mu_{\boldsymbol{l}_{ik'}}\right)
\end{aligned} \tag{9}
$$

$$
q(z_{ik}) \propto \exp\left\{\frac{z_{ik}}{2}\left(\psi(\hat{a}_{\alpha_k}) - \log 2\pi\hat{b}_{\alpha_k} + \frac{\mu^{2*}_{l_{ik}}}{\sigma^{2*}_{l_{ik}}}\right)\right\}\left(\sqrt{\sigma^{2*}_{l_{ik}}}\pi_k\right)^{z_{ik}}(1-\pi_k)^{1-z_{ik}}. \tag{10}
$$

Coordinate ascent for $\boldsymbol{f}_{\cdot j}$ gives

$$
\begin{aligned}
q^*(\boldsymbol{f}_{\cdot j}) &\propto \exp\left\{\mathbb{E}_{,\mathbf{L},\mathbf{Z},\boldsymbol{\tau}}\left[\log p(\boldsymbol{f}_{\cdot j} \mid \mathbf{Y}, \mathbf{L}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha})\right]\right\} \\
&\propto \exp\left\{\mathbb{E}_{,\mathbf{L},\mathbf{Z},\boldsymbol{\tau}}\left[-\frac{1}{2}\left(\boldsymbol{y}_{\cdot j} - \mathbf{L}\boldsymbol{f}_{\cdot j}\right)^\top D_{\boldsymbol{\tau}}\left(\boldsymbol{y}_{\cdot j} - \mathbf{L}\boldsymbol{f}_{\cdot j}\right)\right] - \frac{1}{2}\boldsymbol{f}_{\cdot j}^\top\boldsymbol{f}_{\cdot j}\right\} \\
&\propto \exp\left\{\boldsymbol{y}_{\cdot j}^\top D_{\overline{\boldsymbol{\tau}}}\overline{\mathbf{L}}\boldsymbol{f}_{\cdot j} - \frac{1}{2}\boldsymbol{f}_{\cdot j}^\top\overline{\mathbf{L}^\top D_{\boldsymbol{\tau}}\mathbf{L}}\boldsymbol{f}_{\cdot j} - \frac{1}{2}\boldsymbol{f}_{\cdot j}^\top\boldsymbol{f}_{\cdot j}\right\}
\end{aligned}
$$

where

$$
\begin{aligned}
D_{\overline{\boldsymbol{\tau}}} &= \mathrm{diag}\left(\left\{\frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}}\right\}_{i=1}^G\right) \\
\left[\overline{\mathbf{L}}\right]_{ik} &= \eta_{ik}\mu_{l_{ik}} \\
\left[\overline{\mathbf{L}^\top D_{\boldsymbol{\tau}}\mathbf{L}}\right]_{kk'} &= \sum_{i=1}^G\frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}}\eta_{ik}\eta_{ik'}^{1-\delta_{kk'}}\left(\delta_{kk'}\sigma^2_{l_{ik}} + \mu_{l_{ik}}\mu_{l_{ik'}}\right),
\end{aligned}
$$

which corresponds to the updates

$$
\begin{aligned}
\Sigma^*_{\boldsymbol{f}_{\cdot j}} &= \left(\overline{\mathbf{L}^\top D_{\boldsymbol{\tau}} \mathbf{L}} + \mathbf{I}\right)^{-1} \\
\boldsymbol{\mu}^*_{\boldsymbol{f}_{\cdot j}} &= \Sigma^*_{\boldsymbol{f}_{\cdot j}} \overline{\mathbf{L}}^T D_{\overline{\boldsymbol{\tau}}} \boldsymbol{y}_{\cdot j}.
\end{aligned}
\tag{11}
$$

Coordinate ascent for τ_i gives

$$
\begin{aligned}
q^*(\tau_i) &\propto \exp\left\{\mathbb{E}_{\mathbf{L},\mathbf{F},\mathbf{Z}}[\log p(\tau_i \mid \mathbf{Y}, \mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\alpha})]\right\} \\
&\propto \exp\left\{\left(a_\tau + \frac{N}{2}\right)\log\tau_i - b_\tau\tau_i - \frac{\tau_i}{2}\mathbb{E}_{\mathbf{L},\mathbf{F},\mathbf{Z}}\left[\left(\boldsymbol{y}_{i\cdot} - \mathbf{F}^\top\boldsymbol{l}_{i\cdot}\right)^\top\left(\boldsymbol{y}_{i\cdot} - \mathbf{F}^\top\boldsymbol{l}_{i\cdot}\right)\right]\right\} \\
&\propto \exp\left\{\left(a_\tau + \frac{N}{2}\right)\log\tau_i - \left(b_\tau + \frac{1}{2}\left(\boldsymbol{y}_{i\cdot}^\top\boldsymbol{y}_{i\cdot} - 2\overline{\boldsymbol{l}_{i\cdot}}^\top\overline{\mathbf{F}}\boldsymbol{y}_{i\cdot} + \overline{\boldsymbol{l}_{i\cdot}^\top\mathbf{F}\mathbf{F}^\top\boldsymbol{l}_{i\cdot}}\right)\right)\tau_i\right\}
\end{aligned}
$$

where

$$
\begin{aligned}
\overline{\boldsymbol{l}_{i\cdot}} &= \{\eta_{ik}\mu_{l_{ik}}\}_{k=1}^K \\
\left[\overline{\mathbf{F}}\right]_{kj} &= \mu_{f_{kj}} \\
\overline{\boldsymbol{l}_{i\cdot}^\top\mathbf{F}\mathbf{F}^\top\boldsymbol{l}_{i\cdot}} &= \sum_{k=1}^K\sum_{k'=1}^K\left(\eta_{ik}\eta_{ik'}^{1-\delta_{kk'}}\left(\delta_{kk'}\sigma^2_{l_{ik}} + \mu_{l_{ik}}\mu_{l_{ik'}}\right)\sum_{j=1}^N\left(\left[\Sigma_{\boldsymbol{f}_{\cdot j}}\right]_{kk'} + \left[\boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}}\right]_k\left[\boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}}\right]_{k'}\right)\right),
\end{aligned}
$$

which corresponds to the updates

$$
\begin{aligned}
\hat{a}^*_{\tau_i} &= a_\tau + \frac{N}{2} \\
\hat{b}^*_{\tau_i} &= b_\tau + \frac{1}{2}\left(\boldsymbol{y}_{i\cdot}^\top\boldsymbol{y}_{i\cdot} - 2\overline{\boldsymbol{l}_{i\cdot}}^\top\overline{\mathbf{F}}\boldsymbol{y}_{i\cdot} + \overline{\boldsymbol{l}_{i\cdot}^\top\mathbf{F}\mathbf{F}^\top\boldsymbol{l}_{i\cdot}}\right).
\end{aligned}
\tag{12}
$$

Coordinate ascent for α_k gives

$$
\begin{aligned}
q^*(\alpha_k) &\propto \exp\left\{\mathbb{E}_{\mathbf{L},\mathbf{Z}}[\log p(\alpha_k \mid \mathbf{Y}, \mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\tau})]\right\} \\
&\propto \exp\left\{\left(a_\alpha + \frac{1}{2}\mathbb{E}_{\mathbf{Z}}\left[\sum_{i=1}^G z_{ik}\right]\right)\log\alpha_k - b_\alpha\alpha_k - \frac{\alpha_k}{2}\mathbb{E}_{\mathbf{L},\mathbf{Z}}\left[\sum_{i\,:\,z_{ik}=1} l_{ik}^2\right]\right\}
\end{aligned}
$$

which corresponds to the updates

$$
\begin{aligned}
\hat{a}^*_{\alpha_k} &= a_\alpha + \frac{1}{2}\sum_{i=1}^G \eta_{ik} \\
\hat{b}^*_{\alpha_k} &= b_\alpha + \frac{1}{2}\sum_{i=1}^G \eta_{ik}\left(\sigma^2_{l_{ik}} + \mu^2_{l_{ik}}\right).
\end{aligned}
\tag{13}
$$

4 Unused attempts

4.1 Capturing dependency within $\boldsymbol{l}_{i\cdot}$

To capture dependency within $\boldsymbol{l}_{i\cdot}$, the variational factorisation is modified:

$$
q(\mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha}) = \prod_{i=1}^G q(\boldsymbol{l}_{i\cdot}, \boldsymbol{z}_{i\cdot})q(\tau_i) \times \prod_{j=1}^N q(\boldsymbol{f}_{\cdot j}) \times \prod_{k=1}^K q(\alpha_k) \tag{14}
$$

where

$$
\begin{aligned}
q(\boldsymbol{l}_{i\cdot}, \boldsymbol{z}_{i\cdot}) &= \mathcal{N}\big([\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}} \;\big|\; \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}}, \Sigma_{\boldsymbol{l}_{i\cdot}}\big) \times \prod_{k \,:\, z_{ik}=0} \delta(l_{ik}) \times q(\boldsymbol{z}_{i\cdot}) \\
q(\boldsymbol{f}_{\cdot j}) &= \mathcal{N}\Big(\boldsymbol{f}_{\cdot j} \;\Big|\; \boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}}, \Sigma_{\boldsymbol{f}_{\cdot j}}\Big) \\
q(\tau_i) &= \Gamma\Big(\tau_i \;\Big|\; \hat{a}_{\tau_i}, \hat{b}_{\tau_i}\Big) \\
q(\alpha_k) &= \Gamma\Big(\alpha_k \;\Big|\; \hat{a}_{\alpha_k}, \hat{b}_{\alpha_k}\Big).
\end{aligned}
$$

Note that the dimensions of $\mu_{\boldsymbol{l}_{i\cdot}}$ and $\Sigma_{\boldsymbol{l}_{i\cdot}}$ depend on $\boldsymbol{z}_{i\cdot}$, so there is no sensible way to numerically record its value. This will prove to be problematic.

Coordinate ascent for $\boldsymbol{l}_{i\cdot}$ and $\boldsymbol{z}_{i\cdot}$ gives

$$
\begin{aligned}
q^*(\boldsymbol{l}_{i\cdot}, \boldsymbol{z}_{i\cdot}) &\propto \exp\left\{\mathbb{E}_{\mathbf{F},\tau_i,\boldsymbol{\alpha}}[\log p(\boldsymbol{l}_{i\cdot}, \boldsymbol{z}_{i\cdot} \mid \mathbf{Y}, \mathbf{F}, \boldsymbol{\tau}, \boldsymbol{\alpha})]\right\} \\
&\propto \exp\left\{\frac{1}{2}\sum_{k \,:\, z_{ik}=1} \mathbb{E}_{\alpha_k}\left[\log\frac{\alpha_k}{2\pi}\right] - \mathbb{E}_{\mathbf{F},\tau_i}\left[\frac{\tau_i}{2}\Big(\boldsymbol{y}_{i\cdot} - [\mathbf{F}]_{\boldsymbol{z}_{i\cdot}}^{\mathsf{T}}[\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}\Big)^{\mathsf{T}}\Big(\boldsymbol{y}_{i\cdot} - [\mathbf{F}]_{\boldsymbol{z}_{i\cdot}}^{\mathsf{T}}[\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}\Big)\right]\right. \\
&\qquad \left. - \frac{1}{2}\mathbb{E}_{\boldsymbol{\alpha}}\Big[[\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}^{\mathsf{T}}[D_{\boldsymbol{\alpha}}]_{\boldsymbol{z}_{i\cdot}}[\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}\Big]\right\} \times \prod_{k=1}^{K}\pi_k^{z_{ik}}(1-\pi_k)^{1-z_{ik}} \\
&\propto \exp\left\{\frac{1}{2}\sum_{k \,:\, z_{ik}=1}\left(\overline{\log\alpha_k} - \log 2\pi\right) - \frac{\hat{a}_{\tau_i}}{2\hat{b}_{\tau_i}}\mathbb{E}_{\mathbf{F}}\left[-2\boldsymbol{y}_{i\cdot}^{\mathsf{T}}[\mathbf{F}]_{\boldsymbol{z}_{i\cdot}}^{\mathsf{T}}[\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}} + \Big[\boldsymbol{l}_{i\cdot}^{\mathsf{T}}\Big]_{\boldsymbol{z}_{i\cdot}}[\mathbf{F}]_{\boldsymbol{z}_{i\cdot}}[\mathbf{F}]_{\boldsymbol{z}_{i\cdot}}^{\mathsf{T}}[\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}\right]\right. \\
&\qquad \left. - \frac{1}{2}[\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}^{\mathsf{T}}[D_{\overline{\boldsymbol{\alpha}}}]_{\boldsymbol{z}_{i\cdot}}[\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}\right\} \times \prod_{k=1}^{K}\pi_k^{z_{ik}}(1-\pi_k)^{1-z_{ik}} \\
&\propto \exp\left\{\frac{1}{2}\sum_{k \,:\, z_{ik}=1}\left(\overline{\log\alpha_k} - \log 2\pi\right) + \frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}}\sum_{k \,:\, z_{ik}=1} l_{ik}\sum_{j=1}^{N} y_{ij}\overline{f_{kj}}\right. \\
&\qquad \left. - \frac{\hat{a}_{\tau_i}}{2\hat{b}_{\tau_i}}\sum_{k \,:\, z_{ik}=1}\sum_{k' \,:\, z_{ik'}=1} l_{ik}l_{ik'}\sum_{j=1}^{N}\overline{f_{kj}f_{k'j}} - \frac{1}{2}[\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}^{\mathsf{T}}[D_{\overline{\boldsymbol{\alpha}}}]_{\boldsymbol{z}_{i\cdot}}[\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}\right\} \\
&\qquad \times \prod_{k=1}^{K}\pi_k^{z_{ik}}(1-\pi_k)^{1-z_{ik}},
\end{aligned}
$$

where

$$
\begin{aligned}
\overline{\log\alpha_k} &= \psi(\hat{a}_{\alpha_k}) - \log\hat{b}_{\alpha_k} \\
D_{\overline{\boldsymbol{\alpha}}} &= \mathrm{diag}\left(\left\{\frac{\hat{a}_{\alpha_k}}{\hat{b}_{\alpha_k}}\right\}_{k=1}^{K}\right) \\
\overline{f_{kj}} &= \Big[\boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}}\Big]_k \\
\overline{f_{kj}f_{k'j}} &= \Big[\Sigma_{\boldsymbol{f}_{\cdot j}}\Big]_{kk'} + \Big[\boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}}\Big]_k\Big[\boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}}\Big]_{k'}.
\end{aligned}
$$

The variational parameters corresponding to $[\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}$ are thus updated to

$$\begin{aligned}
\Sigma^*_{\boldsymbol{l}_{i\cdot}} &= \left(\frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}} \sum_{j=1}^{N} \left(\left[\Sigma_{\boldsymbol{f}_{\cdot j}}\right]_{\boldsymbol{z}_{i\cdot},\boldsymbol{z}_{i\cdot}} + \left[\boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}}\right]_{\boldsymbol{z}_{i\cdot}} \left[\boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}}\right]^{\mathsf{T}}_{\boldsymbol{z}_{i\cdot}} \right) + [D_{\overline{\boldsymbol{\alpha}}}]_{\boldsymbol{z}_{i\cdot}} \right)^{-1} \\
\boldsymbol{\mu}^*_{\boldsymbol{l}_{i\cdot}} &= \frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}} \Sigma^*_{\boldsymbol{l}_{i\cdot}} \sum_{j=1}^{N} y_{ij} \left[\boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}}\right]_{\boldsymbol{z}_{i\cdot}}
\end{aligned} \tag{15}$$

where $\left[\Sigma_{\boldsymbol{f}_{\cdot j}}\right]_{\boldsymbol{z}_{i\cdot},\boldsymbol{z}_{i\cdot}}$ is a principal minor of $\Sigma_{\boldsymbol{f}_{\cdot j}}$ whose rows and columns indices correspond to the entries of 1s in $\boldsymbol{z}_{i\cdot}$, and $\left[\boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}}\right]_{\boldsymbol{z}_{i\cdot}}$ is a vector consisting of entries of $\boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}}$ whose corresponding entries of $\boldsymbol{z}_{i\cdot}$ are equal to 1.

Marginalising out $\boldsymbol{l}_{i\cdot}$ then gives

$$q^*(\boldsymbol{z}_{i\cdot}) \propto \exp\left\{ \frac{1}{2} \sum_{k\,:\,z_{ik}=1} \left(\overline{\log \alpha_k} - \log 2\pi\right) + \frac{1}{2} \boldsymbol{\mu}^{\mathsf{T}}_{\boldsymbol{l}_{i\cdot}} \Sigma^{-1}_{\boldsymbol{l}_{i\cdot}} \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}} \right\} \det |\Sigma_{\boldsymbol{l}_{i\cdot}}|^{\frac{1}{2}} \prod_{k=1}^{K} \pi_k^{z_{ik}} (1-\pi_k)^{1-z_{ik}} \,. \tag{16}$$

Unfortunately, this distribution is intractable to obtain. We instead approximate $q^*(\boldsymbol{z}_{i\cdot})$ with

$$\hat{q}(\boldsymbol{z}_{i\cdot}) = \prod_{k=1}^{K} \hat{q}(z_{ik}) = \prod_{k=1}^{K} \mathrm{Bern}(z_{ik} \mid \gamma_{ik}). \tag{17}$$

Methods for estimating γ_{ik} will be addressed in Section 4.2.

For the remaining variational parameters, computations similar to the previous section give the following updates:

$$\begin{aligned}
\Sigma^*_{\boldsymbol{f}_{\cdot j}} &= \left(\overline{\mathbf{L}^{\mathsf{T}} D_{\boldsymbol{\tau}} \mathbf{L}} + \mathbf{I}\right)^{-1} \\
\boldsymbol{\mu}^*_{\boldsymbol{f}_{\cdot j}} &= \Sigma^*_{\boldsymbol{f}_{\cdot j}} \overline{\mathbf{L}}^{T} D_{\overline{\boldsymbol{\tau}}} \boldsymbol{y}_{\cdot j}
\end{aligned} \tag{18}$$

$$\begin{aligned}
\hat{a}^*_{\tau_i} &= a_\tau + \frac{N}{2} \\
\hat{b}^*_{\tau_i} &= b_\tau + \frac{1}{2}\left(\boldsymbol{y}^{\mathsf{T}}_{i\cdot}\boldsymbol{y}_{i\cdot} - 2\overline{\boldsymbol{l}_{i\cdot}}^{\mathsf{T}} \overline{\mathbf{F}} \boldsymbol{y}_{i\cdot} + \overline{\boldsymbol{l}^{\mathsf{T}}_{i\cdot} \mathbf{F}\mathbf{F}^{\mathsf{T}} \boldsymbol{l}_{i\cdot}}\right)
\end{aligned} \tag{19}$$

$$\begin{aligned}
\hat{a}^*_{\alpha_k} &= a_\alpha + \frac{1}{2} \sum_{i=1}^{G} \eta_{ik} \\
\hat{b}^*_{\alpha_k} &= b_\alpha + \frac{1}{2} \sum_{i=1}^{G} \eta_{ik} \left(\sigma^2_{l_{ik}} + \mu^2_{l_{ik}}\right) .
\end{aligned} \tag{20}$$

However, these expectations depend on $\mu_{\boldsymbol{l}_{i\cdot}}$ and $\Sigma_{\boldsymbol{l}_{i\cdot}}$, which we do not have values of due to their varying dimensions.

4.2 Estimating γ_{ik}

Let $B = \{0,1\}^K$, the set of binary vectors of size K. Through this section, q^* refers to the unnormalised density stated in Equation (16) for some fixed i. The direct approach of estimating γ_{ik} is to compute

$$\gamma^*_{ik} = \sum_{\substack{\boldsymbol{\zeta}\in B \\ \zeta_k = 1}} q^*(\boldsymbol{\zeta}) \Big/ \sum_{\boldsymbol{\zeta}\in B} q^*(\boldsymbol{\zeta}) \,. \tag{21}$$

However, this would take 2^K matrix inversions, which is infeasible for large K. Instead, we seek to estimate γ_{ik} (for all k with some fixed i) using values of $q^*(\boldsymbol{\zeta})$ for $\boldsymbol{\zeta} \in \mathbf{B}_T$, where $\mathbf{B}_T$ is a random subset of B of size T.

Define

$$g_{ik}(z) = \sum_{\substack{\boldsymbol{\zeta}\in\mathbf{B}_T \\ \zeta_k = z}} q^*(\boldsymbol{\zeta}) \bigg/ |\{\boldsymbol{\zeta} \in \mathbf{B}_T : \zeta_k = z\}| \quad \text{for } z = 0, 1.$$

We then have an aggregation-based method for estimating γ_{ik}:

$$\gamma_{ik}^* = \frac{g_{ik}(1)}{g_{ik}(0) + g_{ik}(1)}. \tag{22}$$

Another *ad hoc* method is to use the independence assumption in Equation (17). Let

$$\gamma_{ik} = \frac{1}{1 + \exp(-u_{ik})},$$

the approximation

$$q^*(\boldsymbol{\zeta}) \approx \prod_{k=1}^{K} \gamma_{ik}^{\zeta_k} (1 - \gamma_{ik})^{1-\zeta_k}$$

is then equivalent to

$$q^*(\boldsymbol{\zeta}) \approx \prod_{k=1}^{K} \frac{\exp(u_{ik}\zeta_k)}{1 + \exp(u_{ik})}.$$

Taking logs of both sides then gives a regression problem:

$$\log q^*(\boldsymbol{\zeta}) \approx u_0 + \sum_{k=1}^{K} u_{ik}\zeta_k \tag{23}$$

where u_0 is some constant. Each element $\boldsymbol{\zeta}$ of $\mathbf{B}_T$ and its corresponding value of $q(\boldsymbol{\zeta})$ serves as a data point to be used for regression.

One last approach is to apply coordinate ascent to only a random subset of $\boldsymbol{l}_{i\cdot}$ and $\boldsymbol{z}_{i\cdot}$ during each iteration. If the subset is small enough, the direct approach found in Equation 21 can then be feasible. This idea is motivated by stochastic variational inference, but does not share the same theoretical guarantees.

References

Erosheva, E. A. and Curtis, S. M. (2017). Dealing with Reflection Invariance in Bayesian Factor Analysis. *Psychometrika*, 82(2):295–307.

Jonker, R. and Volgenant, A. (1987). A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809.