Bayesian Inference for Sparse Factor Models

Yong See Foo

February 24, 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 samples. We wish to model gene expression as a weighted combination of K latent factors:

$$y_{ij} = \sum_{k=1}^{K} l_{ik} f_{kj} + e_{ij},$$

where l_{ik} is the regulatory effect of factor k on gene i, f_{kj} is the activation of factor k for sample j, 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 Gaussian distributed noise with gene-specific variance, the distribution of the gene expression $\mathbf{Y}$ can be defined as

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

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

$$p(\boldsymbol{y}_{i\cdot} \mid \mathbf{L}, \mathbf{F}, \boldsymbol{\tau}) = \mathcal{N}\big(\boldsymbol{y}_{\cdot j} \;\big|\; \mathbf{L} \boldsymbol{f}_{\cdot j}, D_{\boldsymbol{\tau}}^{-1}\big),$$

where $\boldsymbol{y}_{\cdot j}$ indicates the jth column of $\mathbf{Y}$, and $D_{\boldsymbol{v}} = \operatorname{diag}(\boldsymbol{v})$ for any vector $\boldsymbol{v}$.

It is desirable for the loading matrix $\mathbf{L}$ to be sparse for better interpretability of the fitted factors as biological pathways. Typically, biological pathways regulate only a small subset of genes. The *spike and slab prior* is employed to allow entries of $\mathbf{L}$ to possibly be exactly zero:

$$p(l_{ik} \mid z_{ik}, \alpha_k) = \begin{cases} \delta_0(l_{ik}) & \text{if } z_{ik} = 0 \\ \mathcal{N}\big(l_{ik} \;\big|\; 0, \alpha_k^{-1}\big) & \text{if } z_{ik} = 1 \end{cases},$$

where $z_{ik} = 0$ if gene i is not regulated by 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.

The precision of the loading matrix $\mathbf{L}$ is set to be factor-specific to model different regulatory importance across factors. This is achieved via the *automatic relevance determination prior*:

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

where a_α and b_α are hyperparameters to be specified.

To avoid non-identifiability issues caused by scaling, we employ a normal prior with unit variance on 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 the precision parameters of the noise model:

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

where a_τ and b_τ are hyperparameters to be specified.

2 Markov chain Monte Carlo (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

$$\begin{aligned} p(\boldsymbol{l}_{i\cdot}, \boldsymbol{z}_{i\cdot} \mid \mathbf{Y}, \mathbf{F}, \boldsymbol{\tau}, \boldsymbol{\alpha}) \propto & \prod_{k\,:\,z_{ik}=1} \pi_k \sqrt{\frac{\alpha_k}{2\pi}} \times \prod_{k\,:\,z_{ik}=0} (1-\pi_k)\, \delta_0(l_{ik}) \\ & \times \exp\Bigg\{ -\frac{\tau_i}{2} \left(\boldsymbol{y}_{i\cdot} - [\mathbf{F}]^\top_{\boldsymbol{z}_{i\cdot}} [\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}\right)^\top \left(\boldsymbol{y}_{i\cdot} - [\mathbf{F}]^\top_{\boldsymbol{z}_{i\cdot}} [\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}\right) \\ & \qquad - \frac{1}{2} [\boldsymbol{l}_{i\cdot}]^\top_{\boldsymbol{z}_{i\cdot}} [D_{\boldsymbol{\alpha}}]_{\boldsymbol{z}_{i\cdot}} [\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}} \Bigg\} \\ \propto & \prod_{k\,:\,z_{ik}=1} \pi_k \sqrt{\frac{\alpha_k}{2\pi}} \times \prod_{k\,:\,z_{ik}=0} (1-\pi_k)\, \delta_0(l_{ik}) \\ & \times \exp\left\{ -\frac{1}{2} \left([\boldsymbol{l}]_{\boldsymbol{z}_{i\cdot}} - \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}}\right)^\top \Sigma^{-1}_{\boldsymbol{l}_{i\cdot}} \left([\boldsymbol{l}]_{\boldsymbol{z}_{i\cdot}} - \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}}\right) + \frac{1}{2} \boldsymbol{\mu}^\top_{\boldsymbol{l}_{i\cdot}} \Sigma^{-1}_{\boldsymbol{l}_{i\cdot}} \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}} \right\} \end{aligned} \tag{1}$$

where

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

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

$$p\big([\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}} \mid \mathbf{Y}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha}\big) = \mathcal{N}\big([\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}} \mid \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}}, \Sigma_{\boldsymbol{l}_{i\cdot}}\big) \times \prod_{k\,:\,z_{ik}=0} \delta_0(l_{ik}). \tag{2}$$

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

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

We also have

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

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 $\boldsymbol{f}_{\cdot j}$:

$$
p(\boldsymbol{f}_{\cdot j} \mid \mathbf{Y}, \mathbf{L}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha}) = \mathcal{N}\left(\boldsymbol{f}_{\cdot j} \,\middle|\, \boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}}, \Sigma_{\boldsymbol{f}_{\cdot j}}\right). \tag{4}
$$

Lastly, we have the full conditional distribution of τ_i:

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

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 Model non-identifiability

Given a mode of the posterior distribution, if the factors (of equal sparsity hyperparameters) 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 at least 2^K equivalent modes in the posterior distribution. For this reason, averages over the entire posterior space, i.e. the true posterior means, do not provide useful summaries of the information available.

However, due to the poor mixing of the Gibbs sampler, it is observed that only one of these equivalent modes is explored. This is because when there is sufficient data, these modes are well-separated in the posterior space. In the case of the Gibbs sampler, since $\mathbf{L}$ and $\mathbf{F}$ are simulated in separate steps, it is practically impossible for label switching or sign switching to happen, given a large enough number of variables to simulate. The issue is thus avoided, as sampling from one mode is sufficient for generic inference purposes. Similar behaviour of the Gibbs sampler is noted by Pritchard et al. (2000), in a work that infers population structure using genotype data.

If the sampler used is capable of moving between modes, or if multiple chains with different starting points are to be combined, then the non-identifiability issue needs to be dealt with. To this end, a relabelling algorithm following the framework of Stephens (2000) is presented in Appendix A.

3 Variational inference

3.1 Mean-field approximation

The variational factorisation

$$
q(\mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha}) = \prod_{i=1}^{G}\left[q(\tau_i)\prod_{k=1}^{K} q(l_{ik} \mid z_{ik}) q(z_{ik})\right] \times \prod_{k=1}^{K}\left[q(\alpha_k)\prod_{j=1}^{N} q(f_{kj})\right] \tag{7}
$$

is used as an approximation to the posterior distribution, where

$$
\begin{aligned}
q(l_{ik} \mid z_{ik}) &= \mathcal{N}\big(l_{ik} \mid \mu_{l_{ik}}, \sigma^2_{l_{ik}}\big)^{z_{ik}} \times \delta_0(l_{ik})^{1-z_{ik}} \\
q(z_{ik}) &= \mathrm{Bern}(z_{ik} \mid \eta_{ik}) \\
q(f_{kj}) &= \mathcal{N}\Big(f_{kj} \;\Big|\; \mu_{f_{kj}}, \sigma^2_{f_{kj}}\Big) \\
q(\tau_i) &= \Gamma\Big(\tau_i \;\Big|\; \hat{a}_{\tau_i}, \hat{b}_{\tau_i}\Big) \\
q(\alpha_k) &= \Gamma\Big(\alpha_k \;\Big|\; \hat{a}_{\alpha_k}, \hat{b}_{\alpha_k}\Big).
\end{aligned}
$$

Throughout this section, all expectations are taken over the distribution $q(\mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha})$.

3.2 Coordinate ascent variational inference (CAVI)

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\Bigg\{ -\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] \\
&\qquad + \frac{z_{ik}}{2}\mathbb{E}_{\alpha_k}\left[\log\frac{\alpha_k}{2\pi} - \alpha_k l_{ik}^2\right]\Bigg\} \times \pi_k^{z_{ik}}\left((1-\pi_k)\,\delta_0(l_{ik})\right)^{1-z_{ik}}, \\
&\propto \exp\Bigg\{ -\frac{\hat{a}_{\tau_i}}{2\hat{b}_{\tau_i}}\mathbb{E}_{\boldsymbol{L}_{-ik},\mathbf{F},\boldsymbol{Z}_{-ik}}\left[-2\boldsymbol{y}_{i\cdot}^{\top}\boldsymbol{f}_{k\cdot}l_{ik} + 2\sum_{k'\neq k} z_{ik'}\boldsymbol{f}_k^{\top}\boldsymbol{f}_{k'\cdot}l_{ik'}l_{ik} + \boldsymbol{f}_k^{\top}\boldsymbol{f}_k l_{ik}^2\right] \\
&\qquad + \frac{1}{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)\Bigg\}^{z_{ik}} \times \pi_k^{z_{ik}}\left((1-\pi_k)\,\delta_0(l_{ik})\right)^{1-z_{ik}} \\
&\propto \exp\Bigg\{\frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}}\left(-\sum_{j=1}^{N}\left(y_{ij}\mu_{f_{kj}} - \sum_{k'\neq k}\eta_{ik'}\mu_{f_{kj}}\mu_{f_{k'j}}\mu_{l_{ik'}}\right)l_{ik} - \sum_{j=1}^{N}\left(\mu^2_{f_{kj}} + \sigma^2_{f_{kj}}\right)\frac{l_{ik}^2}{2}\right) \\
&\qquad + \frac{1}{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)\Bigg\}^{z_{ik}} \times \pi_k^{z_{ik}}\left((1-\pi_k)\,\delta_0(l_{ik})\right)^{1-z_{ik}},
\end{aligned}
$$

which corresponds to the updates

$$
\begin{aligned}
\sigma^{2*}_{l_{ik}} &= \left(\frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}}\sum_{j=1}^{N}\left(\mu^2_{f_{kj}} + \sigma^2_{f_{kj}}\right) + \frac{\hat{a}_{\alpha_k}}{\hat{b}_{\alpha_k}}\right)^{-1} \\
\mu^*_{l_{ik}} &= \frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}}\sigma^{2*}_{l_{ik}}\sum_{j=1}^{N}\left(y_{ij}\mu_{f_{kj}} - \sum_{k'\neq k}\eta_{ik'}\mu_{f_{kj}}\mu_{f_{k'j}}\mu_{l_{ik'}}\right)
\end{aligned} \tag{8}
$$

$$
q(z_{ik}) \propto \exp\left\{\frac{z_{ik}}{2}\left(\psi(\hat{a}_{\alpha_k}) - \log 2\pi\hat{b}_{\alpha_k} + \frac{\mu^{2*}_{l_{ik}}}{\sigma^{2*}_{l_{ik}}}\right)\right\}\left(\sqrt{2\pi\sigma^{2*}_{l_{ik}}}\pi_k\right)^{z_{ik}}(1-\pi_k)^{1-z_{ik}}. \tag{9}
$$

Coordinate ascent for $\boldsymbol{f}_{kj}$ gives

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

$$\propto \exp\left\{ \boldsymbol{y}_{\cdot j}^{\top} D_{\overline{\boldsymbol{\tau}}} \overline{\boldsymbol{l}_{\cdot k}} f_{kj} - \sum_{k' \neq k} \overline{f_{k'j} \boldsymbol{l}_{\cdot k'}^{\top} D_{\boldsymbol{\tau}} \boldsymbol{l}_{\cdot k}} f_{kj} - \frac{1}{2} \left(\overline{\boldsymbol{l}_{\cdot k}^{\top} D_{\boldsymbol{\tau}} \boldsymbol{l}_{\cdot k}} + 1 \right) f_{kj}^2 \right\}$$

where

$$\begin{aligned}
D_{\overline{\boldsymbol{\tau}}} &= \operatorname{diag}\left(\left\{ \frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}} \right\}_{i=1}^{G} \right) \\
\overline{\boldsymbol{l}_{\cdot k}} &= \{\eta_{ik} \mu_{l_{ik}}\}_{i=1}^{G} \\
\overline{f_{k'j} \boldsymbol{l}_{\cdot k'}^{\top} D_{\boldsymbol{\tau}} \boldsymbol{l}_{\cdot k}} &= \mu_{f_{k'j}} \sum_{i=1}^{G} \frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}} \eta_{ik} \eta_{ik'} \mu_{l_{ik}} \mu_{l_{ik'}} \\
\overline{\boldsymbol{l}_{\cdot k}^{\top} D_{\boldsymbol{\tau}} \boldsymbol{l}_{\cdot k}} &= \sum_{i=1}^{G} \frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}} \eta_{ik} \left(\mu_{l_{ik}}^2 + \sigma_{l_{ik}}^2 \right),
\end{aligned}$$

which corresponds to the updates

$$\begin{aligned}
\sigma_{f_{kj}}^{2*} &= \left(\overline{\boldsymbol{l}_{\cdot k}^{\top} D_{\boldsymbol{\tau}} \boldsymbol{l}_{\cdot k}} + 1 \right)^{-1} \\
\mu_{f_{kj}}^{*} &= \sigma_{f_{kj}}^{2*} \left(\boldsymbol{y}_{\cdot j}^{\top} D_{\overline{\boldsymbol{\tau}}} \overline{\boldsymbol{l}_{\cdot k}} - \sum_{k' \neq k} \overline{f_{k'j} \boldsymbol{l}_{\cdot k'}^{\top} D_{\boldsymbol{\tau}} \boldsymbol{l}_{\cdot k}} \right).
\end{aligned} \tag{10}$$

Coordinate ascent for τ_i gives

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

where

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

which corresponds to the updates

$$\begin{aligned}
\hat{a}_{\tau_i}^{*} &= a_\tau + \frac{N}{2} \\
\hat{b}_{\tau_i}^{*} &= b_\tau + \frac{1}{2} \left(\boldsymbol{y}_{i\cdot}^{\top} \boldsymbol{y}_{i\cdot} - 2 \overline{\boldsymbol{l}_{i\cdot}}^{\top} \overline{\mathbf{F}} \boldsymbol{y}_{i\cdot} + \overline{\boldsymbol{l}_{i\cdot}^{\top} \mathbf{F} \mathbf{F}^{\top} \boldsymbol{l}_{i\cdot}} \right).
\end{aligned} \tag{11}$$

Coordinate ascent for α_k gives

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

which corresponds to the updates

$$\begin{aligned}
\hat{a}^*_{\alpha_k} &= a_\alpha + \frac{1}{2}\sum_{i=1}^{G} \eta_{ik} \\
\hat{b}^*_{\alpha_k} &= b_\alpha + \frac{1}{2}\sum_{i=1}^{G} \eta_{ik}\left(\sigma^2_{l_{ik}} + \mu^2_{l_{ik}}\right).
\end{aligned} \tag{12}$$

3.3 Computing the evidence lower bound

Variational inference optimises a quantity known as the *evidence lower bound* (ELBO). In coordinate ascent variational inference, the ELBO increases during each parameter update, and can be monitored for convergence (Blei et al., 2017). For this model, the ELBO is given by

$$\text{ELBO}(q) = \mathbb{E}_{\mathbf{L},\mathbf{F},\mathbf{Z},\boldsymbol{\tau},\boldsymbol{\alpha}}[\log p(\mathbf{Y},\mathbf{L},\mathbf{F},\mathbf{Z},\boldsymbol{\tau},\boldsymbol{\alpha}) - \log q(\mathbf{L},\mathbf{F},\mathbf{Z},\boldsymbol{\tau},\boldsymbol{\alpha})].$$

Breaking this down into components, we first have

$$\begin{aligned}
\mathbb{E}_{\mathbf{L},\mathbf{F},\boldsymbol{\tau}}[\log p(y_{ij} \mid \mathbf{L},\mathbf{F},\boldsymbol{\tau})] &= \frac{1}{2}\mathbb{E}_{\mathbf{L},\mathbf{F},\boldsymbol{\tau}}\left[\log\frac{\tau_i}{2\pi} - \tau_i\left(y_{ij} - \boldsymbol{l}_{i\cdot}^{\mathsf{T}}\boldsymbol{f}_{\cdot j}\right)^2\right] \\
&= \frac{1}{2}\Bigg(\psi(\hat{a}_{\tau_i}) - \log 2\pi\hat{b}_{\tau_i} - \frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}}\Bigg(-2y_{ij}\sum_{k=1}^{K}\eta_{ik}l_{ik}\left[\boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}}\right]_k \\
&\qquad + y_{ij}^2 + \overline{\left(\boldsymbol{l}_{i\cdot}^{\mathsf{T}}\boldsymbol{f}_{\cdot j}\right)^2}\Bigg)\Bigg) \\
\mathbb{E}_{\mathbf{L},\mathbf{Z},\boldsymbol{\alpha}}[\log p(l_{ik} \mid \mathbf{Z},\boldsymbol{\alpha})] &= \mathbb{E}_{\mathbf{L},\mathbf{Z},\boldsymbol{\alpha}}\left[\frac{z_{ik}}{2}\left(\log\frac{\alpha_k}{2\pi} - \alpha_k l_{ik}^2\right) + (1-z_{ik})\log\delta_0(l_{ik})\right] \\
&= \frac{\eta_{ik}}{2}\left(\psi(\hat{a}_{\alpha_k}) - \log 2\pi\hat{b}_{\alpha_k} - \frac{\hat{a}_{\alpha_k}}{\hat{b}_{\alpha_k}}\left(\mu^2_{l_{ik}} + \sigma^2_{l_{ik}}\right)\right) \\
&\qquad + \mathbb{E}_{\mathbf{L},\mathbf{Z}}[(1-z_{ik})\log\delta_0(l_{ik})] \\
\mathbb{E}_{\mathbf{Z}}[\log p(z_{ik})] &= \mathbb{E}_{\mathbf{Z}}[z_{ik}\log\pi_k + (1-z_{ik})\log(1-\pi_k)] \\
&= \eta_{ik}\log\pi_k + (1-\eta_{ik})\log(1-\pi_k) \\
\mathbb{E}_{\mathbf{F}}[\log p(f_{kj})] &= -\frac{1}{2}\mathbb{E}_{\mathbf{F}}\left[f_{kj}^2\right] - \log 2\pi \\
&= -\frac{1}{2}\left(\mu^2_{f_{kj}} + \sigma^2_{f_{kj}}\right) - \log 2\pi \\
\mathbb{E}_{\boldsymbol{\tau}}[\log p(\tau_i)] &= \mathbb{E}_{\boldsymbol{\tau}}[(a_\tau - 1)\log\tau_i - b_\tau\tau_i] + a_\tau\log b_\tau - \log\Gamma(a_\tau) \\
&= (a_\tau - 1)\left(\psi(\hat{a}_{\tau_i}) - \log\hat{b}_{\tau_i}\right) - \frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}}b_\tau + a_\tau\log b_\tau - \log\Gamma(a_\tau) \\
\mathbb{E}_{\boldsymbol{\alpha}}[\log p(\alpha_k)] &= \mathbb{E}_{\boldsymbol{\alpha}}[(a_\alpha - 1)\log\alpha_k - b_\alpha\alpha_k] + a_\alpha\log b_\alpha - \log\Gamma(a_\alpha) \\
&= (a_\alpha - 1)\left(\psi(\hat{a}_{\alpha_k}) - \log\hat{b}_{\alpha_k}\right) - \frac{\hat{a}_{\alpha_k}}{\hat{b}_{\alpha_k}}b_\alpha + a_\alpha\log b_\alpha - \log\Gamma(a_\alpha)
\end{aligned}$$

where

$$\overline{\left(\boldsymbol{l}_{i\cdot}^{\mathsf{T}}\boldsymbol{f}_{\cdot j}\right)^2} = \sum_{k=1}^{K}\sum_{k'=1}^{K}\eta_{ik}\eta_{ik'}^{1-\delta_{kk'}}\left(\mu_{l_{ik}}\mu_{l_{ik'}} + \delta_{kk'}\sigma^2_{l_{ik}}\right)\left(\mu_{f_{kj}}\mu_{f_{k'j}} + \delta_{kk'}\sigma^2_{f_{kj}}\right).$$

Next, using standard differential entropy results, we have

$$\mathbb{E}_{\mathbf{L},\mathbf{Z}}[-\log q(l_{ik}, z_{ik})] = \frac{\eta_{ik}}{2}\left(\log 2\pi\sigma^2_{l_{ik}} + 1\right) - \eta_{ik}\log\eta_{ik} - (1-\eta_{ik})\log(1-\eta_{ik})$$

$$
\begin{aligned}
&\quad - \mathbb{E}_{\mathbf{L},\mathbf{Z}}[(1 - z_{ik}) \log \delta_0(l_{ik})] \\
\mathbb{E}_{\mathbf{F}}[-\log q(f_{kj})] &= \frac{1}{2} \log 2\pi\sigma^2_{f_{kj}} \\
\mathbb{E}_{\boldsymbol{\tau}}[-\log q(\tau_i)] &= \hat{a}_{\tau_i} - \log \hat{b}_{\tau_i} + \log \Gamma(\hat{a}_{\tau_i}) + (1 - \hat{a}_{\tau_i})\, \psi(\hat{a}_{\tau_i}) \\
\mathbb{E}_{\boldsymbol{\alpha}}[-\log q(\alpha_k)] &= \hat{a}_{\alpha_k} - \log \hat{b}_{\alpha_k} + \log \Gamma(\hat{a}_{\alpha_k}) + (1 - \hat{a}_{\alpha_k})\, \psi(\hat{a}_{\alpha_k}).
\end{aligned}
$$

The ELBO may be calculated by summing up these results appropriately.

References

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877. arXiv: 1601.00670.

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.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

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

Appendix A Relabelling samples

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

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

to be a choice of means and variances of the entries of $\mathbf{F}$. 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:

$$
\mathcal{L}(\mathbf{a}, \sigma, \boldsymbol{\nu}; \mathbf{F}) = -\sum_{k=1}^{K} \sum_{j=1}^{N} \mathcal{N}(\nu_{\sigma(k)} f_{\sigma(k)j} \mid m_{f_{kj}}, s^2_{f_{kj}}). \tag{13}
$$

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

$$
\mathcal{R}_{\mathrm{MC}} = \sum_{t=1}^{T} \mathcal{L}\Big(\mathbf{a}, \Big\{\Big(\sigma^{(t)}, \boldsymbol{\nu}^{(t)}\Big)\Big\}_{t=1:T}; \mathbf{F}^{(t)}\Big)
$$

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

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

The procedure is terminated when a fixed point is reached. The final signflips and permutations $\left\{\left(\sigma^{(t)}, \boldsymbol{\nu}^{(t)}\right)\right\}_{t=1:T}$ are then applied to all relevant variables simulated.

Step 1 may be solved analytically, by setting partial derivatives of the Monte Carlo risk with respect to the action parameters to zero. This is equivalent to finding the maximum likelihood estimators, summarised by the following updates:

$$\begin{aligned}
\widehat{m_{f_{kj}}} &= \frac{1}{T}\sum_{t=1}^{T} \nu^{(t)}_{\sigma^{(t)}(k)} f^{(t)}_{\sigma^{(t)}(k)j} \\
\widehat{s^2_{f_{kj}}} &= \frac{1}{T}\sum_{t=1}^{T} \left(\nu^{(t)}_{\sigma^{(t)}(k)} f^{(t)}_{\sigma^{(t)}(k)j} - \widehat{m_{f_{kj}}}\right)^2.
\end{aligned}$$

Step 2 is equivalent to the linear assignment problem. For each simulated sample, this may be solved by an $\mathcal{O}(K^3)$ algorithm of Jonker and Volgenant (1987) after a cost matrix is constructed. The construction of the cost matrix itself takes $\mathcal{O}\big(K^2(G+N)\big)$ time (for each simulated sample).