Derivation of VEM for M&MASH

David Gerard

April 11, 2017

Abstract

Here, I derive the VEM algorithm for M&MASH. This is not for final publication, but only for my own notes.

1 Setup

I now describe the notation used in this text. I denote matrices by boldface uppercase letters ($\boldsymbol{A}$), vectors are denoted by boldface lowercase letters ($\boldsymbol{a}$), and scalars are denoted by non-boldface letters (a or A). All vectors are column-vectors. Lowercase letters may represent elements of a vector or matrix if they have subscripts. For example, a_{ij} is the (i,j)th element of $\boldsymbol{A}$, a_i is the ith element of $\boldsymbol{a}$, and $\boldsymbol{a}_i$ is either the ith row or ith column of $\boldsymbol{A}$. For indexing, we will generally use capital non-boldface letters to denote the total number of elements and their lowercase non-boldface versions to denote the index. For example, $i = 1, \ldots, I$. We let $\boldsymbol{A}_{n \times p}$ denote that $\boldsymbol{A} \in \mathbb{R}^{n \times p}$. We denote the matrix transpose by $\boldsymbol{A}^\intercal$, the matrix inverse by $\boldsymbol{A}^{-1}$, and the matrix determinant by $\det(\boldsymbol{A})$. Finally, sets will be denoted by calligraphic letters ($\mathcal{A}$).

We assume the following model

$$\boldsymbol{Y}_{N \times M} = \boldsymbol{X}_{N \times K}\boldsymbol{B}_{K \times M} + \boldsymbol{Z}_{N \times R}\boldsymbol{A}_{R \times M} + \boldsymbol{E}_{N \times M}, \tag{1}$$

where

$$\boldsymbol{E} \sim \underset{N \times M}{\mathrm{N}}(\boldsymbol{0}, \boldsymbol{\Lambda}^{-1} \otimes \boldsymbol{I}_N), \tag{2}$$

$$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_M), \tag{3}$$

$$\boldsymbol{Z} \sim \underset{N \times R}{\mathrm{N}}(\boldsymbol{0}, \boldsymbol{I}_R \otimes \boldsymbol{I}_N). \tag{4}$$

That is, the overall covariance of the columns of $\boldsymbol{Y}$ is $\boldsymbol{A}^\intercal\boldsymbol{A} + \boldsymbol{\Lambda}^{-1}$, a diagonal plus low rank matrix. We place iid mixtures of normals prior on the rows of $\boldsymbol{B}$ and gamma priors on $\boldsymbol{\Lambda}$.

$$\lambda_m \overset{iid}{\sim} \mathrm{Gamma}(\alpha, \beta), \tag{5}$$

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{b}_1^\intercal \\ \vdots \\ \boldsymbol{b}_K^\intercal \end{pmatrix}, \tag{6}$$

$$\boldsymbol{b}_k \text{ i.i.d. s.t. } p(\boldsymbol{b}_k) = \sum_{t=0}^{T} \pi_t \underset{M}{\mathrm{N}}(\boldsymbol{b}_k|\mathbf{0}, \boldsymbol{V}_t), \tag{7}$$

where the $\boldsymbol{V}_t$'s are $m \times m$ known positive semi-definite covariance matrices and the π_t's are proportions to be estimated. We augment the prior of $\boldsymbol{b}_k$ by including a 1-of-T indicator vector $\boldsymbol{w}_k \in \mathbb{R}^T$ denoting membership of $\boldsymbol{b}_k$ into one of the T mixture groups and write

$$p(\boldsymbol{b}_k|\boldsymbol{w}_k) = \prod_{t=0}^{T} \left[\mathrm{N}(\boldsymbol{b}_k|\mathbf{0}, \boldsymbol{V}_t)\right]^{w_{kt}}, \tag{8}$$

$$p(\boldsymbol{w}_k|\boldsymbol{\pi}) = \prod_{t=0}^{T} \pi_t^{w_{kt}}. \tag{9}$$

We could also place a normal prior on $\boldsymbol{A}$, but we will instead also estimate it by ML.

The densities of all elements involved are

$$p(\boldsymbol{Y}|\boldsymbol{B}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{\Lambda}) = (2\pi)^{-NM/2} \prod_{m=1}^{M} \lambda_m^{N/2} \exp\left\{-\frac{1}{2} \operatorname{tr}\left[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B} - \boldsymbol{Z}\boldsymbol{A})\boldsymbol{\Lambda}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B} - \boldsymbol{Z}\boldsymbol{A})^\intercal\right]\right\} \tag{10}$$

$$p(\boldsymbol{Z}) = (2\pi)^{-NR/2} \exp\left\{-\frac{1}{2} \operatorname{tr}[\boldsymbol{Z}\boldsymbol{Z}^\intercal]\right\} \tag{11}$$

$$p(\lambda_m|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_m^{\alpha-1} \exp\{-\beta\lambda_m\} \tag{12}$$

$$p(\boldsymbol{b}_k|\boldsymbol{w}_k) = \sum_{t=0}^{T} \left[(2\pi)^{-M/2} \det(\boldsymbol{V}_t)^{-1/2} \exp\left\{-\frac{1}{2}\boldsymbol{b}_k^\intercal \boldsymbol{V}_t^{-1} \boldsymbol{b}_k\right\}\right]^{w_{kt}} \tag{13}$$

$$p(\boldsymbol{w}_k|\boldsymbol{\pi}) = \prod_{t=0}^{T} \pi_t^{w_{kt}}. \tag{14}$$

2 VEM

Let q be a function. The goal is to maximize over $(q, \boldsymbol{\pi}, \boldsymbol{A})$ the following lower-bound of the log-marginal likelihood:

$$\log p(\boldsymbol{Y}|\boldsymbol{\pi}, \boldsymbol{A}) \geq \mathcal{L}(q, \boldsymbol{\pi}, \boldsymbol{A}) = \int q(\boldsymbol{B}, \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Lambda}) \log\left\{\frac{p(\boldsymbol{Y}, \boldsymbol{B}, \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Lambda}|\boldsymbol{\pi}, \boldsymbol{A})}{q(\boldsymbol{B}, \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Lambda})}\right\} \mathrm{d}\boldsymbol{B}\, \mathrm{d}\boldsymbol{W}\, \mathrm{d}\boldsymbol{Z}\, \mathrm{d}\boldsymbol{\Lambda}, \tag{15}$$

where

$$p(\boldsymbol{Y}, \boldsymbol{B}, \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Lambda}|\boldsymbol{\pi}, \boldsymbol{A}) = p(\boldsymbol{Y}|\boldsymbol{B}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{\Lambda}) p(\boldsymbol{B}|\boldsymbol{W}) p(\boldsymbol{W}|\boldsymbol{\pi}) p(\boldsymbol{Z}) p(\boldsymbol{\Lambda}), \tag{16}$$

where these densities are defined in equations (10) to (14). We constrain q to take the following form

$$q(\boldsymbol{B}, \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Lambda}) = q(\boldsymbol{Z})q(\boldsymbol{\Lambda}) \prod_{k=1}^{K} q(\boldsymbol{b}_k, \boldsymbol{w}_k), \tag{17}$$

and thus perform mean-field variational inference.

2.1 Update for $q(\boldsymbol{b}_k, \boldsymbol{w}_k)$

By a general result in mean-field variational inference, we have

$$\log q(\boldsymbol{b}_k, \boldsymbol{w}_k) \propto E_{-(\boldsymbol{b}_k, \boldsymbol{w}_k)} \left\{ \log p(\boldsymbol{Y}, \boldsymbol{B}, \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Lambda} | \boldsymbol{\pi}, \boldsymbol{A}) \right\}, \tag{18}$$

where $E_{-(\boldsymbol{b}_k, \boldsymbol{w}_k)}[\cdot]$ is notation for the expectation with respect to the q distributions over all variables except $\boldsymbol{b}_k$ and $\boldsymbol{w}_k$. See (10.9) of **?**, for example. Here, I let "$\propto$" denote equality to up to an additive constant that is independent of $\boldsymbol{b}_k$ and $\boldsymbol{w}_k$. So

$$\begin{aligned}
(18) &\propto E_{-(\boldsymbol{b}_k, \boldsymbol{w}_k)} \left\{ \log p(\boldsymbol{Y} | \boldsymbol{B}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{\Lambda}) + \log p(\boldsymbol{B} | \boldsymbol{W}) + \log p(\boldsymbol{W} | \boldsymbol{\pi}) \right\} && (19) \\
&\propto E_{-(\boldsymbol{b}_k, \boldsymbol{w}_k)} \left\{ -\frac{1}{2} \operatorname{tr}\left[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B} - \boldsymbol{Z}\boldsymbol{A}) \boldsymbol{\Lambda} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B} - \boldsymbol{Z}\boldsymbol{A})^{\intercal} \right] \right. \\
&\left. + \sum_{t=0}^{T} w_{kt} \left[-\frac{M}{2} \log(2\pi) - \frac{1}{2} \log\det(\boldsymbol{V}_t) - \frac{1}{2} \boldsymbol{b}_k^{\intercal} \boldsymbol{V}_t^{-1} \boldsymbol{b}_k + \log(\pi_t) \right] \right\} && (20)
\end{aligned}$$

We note that

$$\boldsymbol{X}\boldsymbol{B} = \sum_{\ell=1}^{K} \boldsymbol{x}_\ell \boldsymbol{b}_\ell^{\intercal}, \tag{21}$$

where $\boldsymbol{x}_k$ is the kth column of $\boldsymbol{X}$ and $\boldsymbol{b}_k$ is the kth row of $\boldsymbol{B}$. So

$$\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{A} - \boldsymbol{X}\boldsymbol{B} = \boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{A} - \sum_{\ell \neq k} \boldsymbol{x}_\ell \boldsymbol{b}_\ell^{\intercal} - \boldsymbol{x}_k \boldsymbol{b}_k^{\intercal}. \tag{22}$$

Let

$$\boldsymbol{R}_{-k} := \boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{A} - \sum_{\ell \neq k} \boldsymbol{x}_\ell \boldsymbol{b}_\ell^{\intercal}, \tag{23}$$

then

$$\begin{aligned}
&\operatorname{tr}\left[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B} - \boldsymbol{Z}\boldsymbol{A}) \boldsymbol{\Lambda} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B} - \boldsymbol{Z}\boldsymbol{A})^{\intercal} \right] = \operatorname{tr}\left[(\boldsymbol{R}_{-k} - \boldsymbol{x}_k \boldsymbol{b}_k^{\intercal}) \boldsymbol{\Lambda} (\boldsymbol{R}_{-k} - \boldsymbol{x}_k \boldsymbol{b}_k^{\intercal})^{\intercal} \right] && (24) \\
&\propto \operatorname{tr}\left[\boldsymbol{x}_k \boldsymbol{b}_k^{\intercal} \boldsymbol{\Lambda} \boldsymbol{b}_k \boldsymbol{x}_k^{\intercal} \right] - 2 \operatorname{tr}\left[\boldsymbol{x}_k \boldsymbol{b}_k^{\intercal} \boldsymbol{\Lambda} \boldsymbol{R}_{-k}^{\intercal} \right] && (25) \\
&= \|\boldsymbol{x}_k\|^2 \boldsymbol{b}_k^{\intercal} \boldsymbol{\Lambda} \boldsymbol{b}_k - 2 \boldsymbol{b}_k^{\intercal} \boldsymbol{\Lambda} \boldsymbol{R}_{-k}^{\intercal} \boldsymbol{x}_k && (26)
\end{aligned}$$

Hence,

$$\begin{aligned}(20) \propto &-\frac{1}{2}\left[\|\boldsymbol{x}_k\|^2\boldsymbol{b}_k^\intercal E\left[\boldsymbol{\Lambda}\right]\boldsymbol{b}_k - 2\boldsymbol{b}_k^\intercal E\left[\boldsymbol{\Lambda}\right]E\left[\boldsymbol{R}_{-k}\right]^\intercal \boldsymbol{x}_k\right] \\ &+\sum_{t=0}^{T} w_{kt}\left[-\frac{M}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{V}_t) - \frac{1}{2}\boldsymbol{b}_k^\intercal \boldsymbol{V}_t^{-1}\boldsymbol{b}_k + \log(\pi_t)\right]\end{aligned} \tag{27}$$

$$\begin{aligned}&\propto \sum_{t=0}^{T} w_{kt}\left[-\frac{M}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{V}_t) + \log(\pi_t)\right. \\ &\left.-\frac{1}{2}\left\{\boldsymbol{b}_k^\intercal\left(\boldsymbol{V}_t^{-1} + \|\boldsymbol{x}_k\|^2 E\left[\boldsymbol{\Lambda}\right]\right)\boldsymbol{b}_k - 2\boldsymbol{b}_k^\intercal E\left[\boldsymbol{\Lambda}\right]E\left[\boldsymbol{R}_{-k}\right]^\intercal \boldsymbol{x}_k\right\}\right]\end{aligned} \tag{28}$$

We are going to complete the square in (28). Let

$$\boldsymbol{\Sigma}_{kt} := \left(\boldsymbol{V}_t^{-1} + \|\boldsymbol{x}_k\|^2 E\left[\boldsymbol{\Lambda}\right]\right)^{-1}, \text{ and} \tag{29}$$

$$\boldsymbol{\mu}_{kt} := \boldsymbol{\Sigma}_{kt} E\left[\boldsymbol{\Lambda}\right]E\left[\boldsymbol{R}_{-k}\right]^\intercal \boldsymbol{x}_k = \left(\boldsymbol{V}_t^{-1} + \|\boldsymbol{x}_k\|^2 E\left[\boldsymbol{\Lambda}\right]\right)^{-1} E\left[\boldsymbol{\Lambda}\right]E\left[\boldsymbol{R}_{-k}\right]^\intercal \boldsymbol{x}_k. \tag{30}$$

Then,

$$(28) \propto \sum_{t=0}^{T} w_{kt}\left[-\frac{M}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{V}_t) + \log(\pi_t) - \frac{1}{2}\left\{\boldsymbol{b}_k^\intercal\boldsymbol{\Sigma}_{kt}^{-1}\boldsymbol{b}_k - 2\boldsymbol{b}_k^\intercal\boldsymbol{\Sigma}_{kt}^{-1}\boldsymbol{\mu}_{kt}\right\}\right] \tag{31}$$

$$\begin{aligned}&\propto \sum_{t=0}^{T} w_{kt}\left[-\frac{M}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{V}_t) + \log(\pi_t) + \frac{1}{2}\boldsymbol{\mu}_{kt}^\intercal\boldsymbol{\Sigma}_{kt}^{-1}\boldsymbol{\mu}_{kt}\right. \\ &\left.-\frac{1}{2}\left\{\boldsymbol{b}_k^\intercal\boldsymbol{\Sigma}_{kt}^{-1}\boldsymbol{b}_k - 2\boldsymbol{b}_k^\intercal\boldsymbol{\Sigma}_{kt}^{-1}\boldsymbol{\mu}_{kt} + \boldsymbol{\mu}_{kt}^\intercal\boldsymbol{\Sigma}_{kt}^{-1}\boldsymbol{\mu}_{kt}\right\}\right]\end{aligned} \tag{32}$$

$$\propto \sum_{t=0}^{T} w_{kt}\left[-\frac{M}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{V}_t) + \log(\pi_t) + \frac{1}{2}\boldsymbol{\mu}_{kt}^\intercal\boldsymbol{\Sigma}_{kt}^{-1}\boldsymbol{\mu}_{kt} - \frac{1}{2}(\boldsymbol{b}_k - \boldsymbol{\mu}_{kt})^\intercal\boldsymbol{\Sigma}_{kt}^{-1}(\boldsymbol{b}_k - \boldsymbol{\mu}_{kt})\right] \tag{33}$$

$$\begin{aligned}&\propto \sum_{t=0}^{T} w_{kt}\left[-\frac{1}{2}\log\det(\boldsymbol{V}_t) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}_{kt}) + \log(\pi_t) + \frac{1}{2}\boldsymbol{\mu}_{kt}^\intercal\boldsymbol{\Sigma}_{kt}^{-1}\boldsymbol{\mu}_{kt}\right. \\ &\left.-\frac{M}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{\Sigma}_{kt}) - \frac{1}{2}(\boldsymbol{b}_k - \boldsymbol{\mu}_{kt})^\intercal\boldsymbol{\Sigma}_{kt}^{-1}(\boldsymbol{b}_k - \boldsymbol{\mu}_{kt})\right].\end{aligned} \tag{34}$$

Equation (34) is the log of a kernal of mixture of normals. The mixing means are the $\boldsymbol{\mu}_{kt}$'s and the mixing covariances are the $\boldsymbol{\Sigma}_{kt}$'s. The mixing proportions can be found by looking at

$$\log\rho_{kt} := -\frac{1}{2}\log\det(\boldsymbol{V}_t) + \frac{1}{2}\log\det(\boldsymbol{\Sigma}_{kt}) + \log(\pi_t) + \frac{1}{2}\boldsymbol{\mu}_{kt}^\intercal\boldsymbol{\Sigma}_{kt}^{-1}\boldsymbol{\mu}_{kt}, \tag{35}$$

$$\rho_{kt} = \pi_t \det(\boldsymbol{V}_t)^{-1/2}\det(\boldsymbol{\Sigma}_{kt})^{1/2}\exp\left\{\frac{1}{2}\boldsymbol{\mu}_{kt}^\intercal\boldsymbol{\Sigma}_{kt}^{-1}\boldsymbol{\mu}_{kt}\right\} \tag{36}$$

We first work with

$$\begin{aligned}
\det(\boldsymbol{V}_t)^{-1/2}\det(\boldsymbol{\Sigma}_{kt})^{1/2} &= \det(\boldsymbol{V}_t)^{-1/2}\det\left(\boldsymbol{V}_t^{-1} + \|\boldsymbol{x}_k\|^2 E\left[\boldsymbol{\Lambda}\right]\right)^{-1/2} && (37)\\
&= \det\left(\boldsymbol{I}_M + \boldsymbol{V}_t\|\boldsymbol{x}_k\|^2 E\left[\boldsymbol{\Lambda}\right]\right)^{-1/2} && (38)\\
&\propto \det\left(\|\boldsymbol{x}_k\|^{-2} E\left[\boldsymbol{\Lambda}\right]^{-1} + \boldsymbol{V}_t\right)^{-1/2}. && (39)
\end{aligned}$$

The "$\propto$" are multiplicative proportionalities for things that do not depend on t. We can ignore other terms because we are just going to normalize the ρ_{kt}'s to sum to 1 anyway. We now work with

$$\exp\left\{\frac{1}{2}\boldsymbol{\mu}_{kt}^{\intercal}\boldsymbol{\Sigma}_{kt}^{-1}\boldsymbol{\mu}_{kt}\right\} = \exp\left\{\frac{1}{2}\boldsymbol{x}_k^{\intercal} E\left[\boldsymbol{R}_{-k}\right] E\left[\boldsymbol{\Lambda}\right]\left(\boldsymbol{V}_t^{-1} + \|\boldsymbol{x}_k\|^2 E\left[\boldsymbol{\Lambda}\right]\right)^{-1} E\left[\boldsymbol{\Lambda}\right] E\left[\boldsymbol{R}_{-k}\right]^{\intercal}\boldsymbol{x}_k\right\} \tag{40}$$

$$= \exp\left\{\frac{1}{2}\|\boldsymbol{x}_k\|^{-2}\boldsymbol{x}_k^{\intercal} E\left[\boldsymbol{R}_{-k}\right]\left(\|\boldsymbol{x}_k\|^{-2} E\left[\boldsymbol{\Lambda}\right]^{-1}\boldsymbol{V}_t^{-1} E\left[\boldsymbol{\Lambda}\right]^{-1}\|\boldsymbol{x}_k\|^{-2} + \|\boldsymbol{x}_k\|^{-2} E\left[\boldsymbol{\Lambda}\right]^{-1}\right)^{-1} E\left[\boldsymbol{R}_{-k}\right]^{\intercal}\boldsymbol{x}_k\|\boldsymbol{x}_k\|^{-2}\right\} \tag{41}$$

Let

$$\boldsymbol{\xi}_k := E\left[\boldsymbol{R}_{-k}\right]^{\intercal}\boldsymbol{x}_k\|\boldsymbol{x}_k\|^{-2}, \tag{42}$$

then

$$\begin{aligned}
(41) &= \exp\left\{\frac{1}{2}\boldsymbol{\xi}_k^{\intercal}\left(\|\boldsymbol{x}_k\|^{-2} E\left[\boldsymbol{\Lambda}\right]^{-1}\boldsymbol{V}_t^{-1} E\left[\boldsymbol{\Lambda}\right]^{-1}\|\boldsymbol{x}_k\|^{-2} + \|\boldsymbol{x}_k\|^{-2} E\left[\boldsymbol{\Lambda}\right]^{-1}\right)^{-1}\boldsymbol{\xi}_k\right\} && (43)\\
&= \exp\left\{\frac{1}{2}\boldsymbol{\xi}_k^{\intercal}\left(\|\boldsymbol{x}_k\|^{-2} E\left[\boldsymbol{\Lambda}\right]^{-1} - \left(\boldsymbol{V}_t + E\left[\boldsymbol{\Lambda}\right]^{-1}\|\boldsymbol{x}_k\|^{-2}\right)^{-1}\right)\boldsymbol{\xi}_k\right\} && (44)\\
&\propto \exp\left\{-\frac{1}{2}\boldsymbol{\xi}_k^{\intercal}\left(\boldsymbol{V}_t + E\left[\boldsymbol{\Lambda}\right]^{-1}\|\boldsymbol{x}_k\|^{-2}\right)^{-1}\boldsymbol{\xi}_k\right\}, && (45)
\end{aligned}$$

where we used Woodbury for (44). Combining (39) and (45), we have that

$$\rho_{kt} \propto \pi_t \, \mathrm{N}(\boldsymbol{\xi}_k|\boldsymbol{0}, \boldsymbol{V}_t + E\left[\boldsymbol{\Lambda}\right]^{-1}\|\boldsymbol{x}_k\|^{-2}). \tag{46}$$

We'll just re-define ρ_{kt} by (46). The mixing proportions are thus

$$\begin{aligned}
\gamma_{kt} &:= \frac{\rho_{kt}}{\sum_{t=0}^{T}\rho_{kt}} && (47)\\
&= \frac{\pi_t \, \mathrm{N}(\boldsymbol{\xi}_k|\boldsymbol{0}, \boldsymbol{V}_t + E\left[\boldsymbol{\Lambda}\right]^{-1}\|\boldsymbol{x}_k\|^{-2})}{\sum_{t=0}^{T}\pi_t \, \mathrm{N}(\boldsymbol{\xi}_k|\boldsymbol{0}, \boldsymbol{V}_t + E\left[\boldsymbol{\Lambda}\right]^{-1}\|\boldsymbol{x}_k\|^{-2})}. && (48)
\end{aligned}$$

You could actually derive this result by first taking expectations and completing the square in (26), then using standard normal arguments for each mixing component. We would be doing the same thing as what we did above, but using standard Bayesian arguments to circumvent a lot of the algebra. That is, we first define $\boldsymbol{\xi}_k$ by

$$\boldsymbol{\xi}_k := E\left[\boldsymbol{R}_{-k}\right]^\intercal \boldsymbol{x}_k \|\boldsymbol{x}_k\|^{-2} \tag{49}$$

then note that

$$(26) \propto (\boldsymbol{\xi}_k - \boldsymbol{b}_k)^\intercal \|\boldsymbol{x}_k\|^2 E[\boldsymbol{\Lambda}](\boldsymbol{\xi}_k - \boldsymbol{b}_k), \tag{50}$$

which is proportional to a log-density of a $\mathrm{N}(\boldsymbol{\xi}_k|\boldsymbol{b}_k, \|\boldsymbol{x}_k\|^{-2}E[\boldsymbol{\Lambda}]^{-1})$. Since each mixture is a $\mathrm{N}(\boldsymbol{b}_k|\boldsymbol{0}, \boldsymbol{V}_t)$, we have that for each new mixture we are calculating

$$p(\boldsymbol{\xi}_k|\boldsymbol{b}_k)p(\boldsymbol{b}_k) = p(\boldsymbol{\xi}_k)p(\boldsymbol{b}_k|\boldsymbol{\xi}_k), \tag{51}$$

where $p(\boldsymbol{\xi}_k) = \mathrm{N}(\boldsymbol{\xi}_k|\boldsymbol{0}, \boldsymbol{V}_t + E\left[\boldsymbol{\Lambda}\right]^{-1}\|\boldsymbol{x}_k\|^{-2})$ and $p(\boldsymbol{b}_k|\boldsymbol{\xi}_k) = \mathrm{N}(\boldsymbol{b}_k|\boldsymbol{\mu}_{kt}, \boldsymbol{\Sigma}_{kt})$.

Important note: If $\boldsymbol{V}_t = \boldsymbol{0}$, then we have that $\boldsymbol{\Sigma}_{kt} = \boldsymbol{0}$ and $\boldsymbol{\mu}_{kt} = 0$ for all $k = 1, \ldots, K$. Thus, the variational density is gauranteed to have a pointmass at zero. This is super important.

2.2 Update $\boldsymbol{Z}$

Again, by a general result in mean-field variational inference, we have

$$\begin{aligned}
\log q(\boldsymbol{Z}) &\propto E_{-\boldsymbol{Z}}\left\{\log p(\boldsymbol{Y}, \boldsymbol{B}, \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Lambda}|\boldsymbol{\pi}, \boldsymbol{A})\right\} && (52)\\
&\propto E_{-\boldsymbol{Z}}\left\{\log p(\boldsymbol{Y}|\boldsymbol{B}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{\Lambda}) + \log p(\boldsymbol{Z})\right\} && (53)\\
&\propto E_{-\boldsymbol{Z}}\left\{-\frac{1}{2}\operatorname{tr}\left[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B} - \boldsymbol{Z}\boldsymbol{A})\boldsymbol{\Lambda}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B} - \boldsymbol{Z}\boldsymbol{A})^\intercal\right] - \frac{1}{2}\operatorname{tr}[\boldsymbol{Z}\boldsymbol{Z}^\intercal]\right\} && (54)\\
&\propto -\frac{1}{2}E_{-\boldsymbol{Z}}\left\{\operatorname{tr}\left[\boldsymbol{Z}\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}^\intercal\boldsymbol{Z}^\intercal - 2\boldsymbol{Z}\boldsymbol{A}\boldsymbol{\Lambda}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})^\intercal + \boldsymbol{Z}\boldsymbol{Z}^\intercal\right]\right\} && (55)\\
&\propto -\frac{1}{2}\operatorname{tr}\left[\boldsymbol{Z}\left(\boldsymbol{A}E[\boldsymbol{\Lambda}]\boldsymbol{A}^\intercal + \boldsymbol{I}_R\right)\boldsymbol{Z}^\intercal - 2\boldsymbol{Z}\boldsymbol{A}E[\boldsymbol{\Lambda}](\boldsymbol{Y} - \boldsymbol{X}E[\boldsymbol{B}])^\intercal\right]. && (56)
\end{aligned}$$

Equation (56) is the log-kernal of a matrix-normal. Hence we have

$$q(\boldsymbol{Z}) = \underset{N\times R}{\mathrm{N}}(\boldsymbol{Z}|\boldsymbol{\mu}_{\boldsymbol{Z}}, \boldsymbol{\Sigma}_{\boldsymbol{Z}} \otimes \boldsymbol{I}_N), \tag{57}$$

where

$$\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{Z}} &:= \left(\boldsymbol{A}E[\boldsymbol{\Lambda}]\boldsymbol{A}^\intercal + \boldsymbol{I}_R\right)^{-1}, \text{ and} && (58)\\
\boldsymbol{\mu}_{\boldsymbol{Z}} &:= (\boldsymbol{Y} - \boldsymbol{X}E[\boldsymbol{B}])E[\boldsymbol{\Lambda}]\boldsymbol{A}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{Z}} && (59)\\
&= (\boldsymbol{Y} - \boldsymbol{X}E[\boldsymbol{B}])E[\boldsymbol{\Lambda}]\boldsymbol{A}^\intercal\left(\boldsymbol{A}E[\boldsymbol{\Lambda}]\boldsymbol{A}^\intercal + \boldsymbol{I}_R\right)^{-1}. && (60)
\end{aligned}$$

2.3 Update $\boldsymbol{\Lambda}$

Again, by a general result in mean-field variational inference, we have

$$\begin{aligned}
\log q(\lambda_m) \propto& E_{-\boldsymbol{\Lambda}}\left\{\log p(\boldsymbol{Y}, \boldsymbol{B}, \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Lambda}|\boldsymbol{\pi}, \boldsymbol{A})\right\} && (61)\\
\propto& E_{-\boldsymbol{\Lambda}}\left\{\log p(\boldsymbol{Y}|\boldsymbol{B}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{\Lambda}) + \log p(\boldsymbol{\Lambda})\right\} && (62)
\end{aligned}$$

$$\begin{aligned}\propto E_{-\boldsymbol{\Lambda}}\Bigg\{&\sum_{m=1}^{M}\frac{N}{2}\log(\lambda_m)-\frac{1}{2}\operatorname{tr}\left[(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}-\boldsymbol{Z}\boldsymbol{A})\boldsymbol{\Lambda}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}-\boldsymbol{Z}\boldsymbol{A})^{\intercal}\right]\\&+\sum_{m=1}^{M}(\alpha-1)\log(\lambda_m)-\sum_{m=1}^{M}\beta\lambda_m\Bigg\}\end{aligned}\tag{63}$$

Let

$$\boldsymbol{\delta}:=\operatorname{diag}\left\{E\left[(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}-\boldsymbol{Z}\boldsymbol{A})^{\intercal}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}-\boldsymbol{Z}\boldsymbol{A})\right]\right\}\in\mathbb{R}^{M}.\tag{64}$$

Then

$$(63)=\sum_{m=1}^{M}\left[(\frac{N}{2}+\alpha-1)\log(\lambda_m)-(\delta_m/2+\beta)\lambda_m\right].\tag{65}$$

Thus,

$$q(\lambda_m)=\operatorname{Gamma}\left(\lambda_m|\frac{N}{2}+\alpha,\delta_m/2+\beta\right).\tag{66}$$

If you don't want to put a prior on λ, but want to optimize it directly, then this this value would simply be

$$\lambda_m=N/\delta_m.\tag{67}$$

However, λ_m only ever shows up through $E[\lambda_m]$, which is

$$\frac{N/2+\alpha}{\delta_m/2+\beta}.\tag{68}$$

So doing maximum likelihood is equivalent to setting $\alpha=\beta=0$ as your prior.

2.4 Update $\boldsymbol{A}$

We need to maximize over $\boldsymbol{A}$

$$E\left\{\log p(\boldsymbol{Y},\boldsymbol{B},\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\Lambda}|\boldsymbol{\pi},\boldsymbol{A})\right\},\tag{69}$$

where the expectation is with respect to q. This is equivalent to

$$\arg\max_{\boldsymbol{A}}E\left\{\log p(\boldsymbol{Y}|\boldsymbol{B},\boldsymbol{Z},\boldsymbol{A},\boldsymbol{\Lambda})\right\}=\arg\max_{\boldsymbol{A}}E\left\{-\frac{1}{2}\operatorname{tr}\left[(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}-\boldsymbol{Z}\boldsymbol{A})\boldsymbol{\Lambda}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}-\boldsymbol{Z}\boldsymbol{A})^{\intercal}\right]\right\}\tag{70}$$

$$=\arg\max_{\boldsymbol{A}}E\left\{-\frac{1}{2}\operatorname{tr}\left[\boldsymbol{Z}\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}^{\intercal}\boldsymbol{Z}^{\intercal}-2\boldsymbol{Z}\boldsymbol{A}\boldsymbol{\Lambda}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})^{\intercal}\right]\right\}\tag{71}$$

$$=\arg\max_{\boldsymbol{A}}\left\{-\frac{1}{2}\operatorname{tr}\left[\boldsymbol{A}E\left[\boldsymbol{\Lambda}\right]\boldsymbol{A}^{\intercal}E\left[\boldsymbol{Z}^{\intercal}\boldsymbol{Z}\right]\right]+\operatorname{tr}\left[\boldsymbol{A}E\left[\boldsymbol{\Lambda}\right](\boldsymbol{Y}-\boldsymbol{X}E\left[\boldsymbol{B}\right])^{\intercal}E\left[\boldsymbol{Z}\right]\right]\right\}.\tag{72}$$

Taking derivatives of $\boldsymbol{A}$ and setting to $\mathbf{0}$, we have

$$-E\left[\boldsymbol{\Lambda}\right]\boldsymbol{A}^{\intercal}E\left[\boldsymbol{Z}^{\intercal}\boldsymbol{Z}\right]+E\left[\boldsymbol{\Lambda}\right](\boldsymbol{Y}-\boldsymbol{X}E\left[\boldsymbol{B}\right])^{\intercal}E\left[\boldsymbol{Z}\right]=\mathbf{0}. \tag{73}$$

Solving for $\boldsymbol{A}$, we get

$$\boldsymbol{A}=E\left[\boldsymbol{Z}^{\intercal}\boldsymbol{Z}\right]^{-1}E\left[\boldsymbol{Z}\right]^{\intercal}(\boldsymbol{Y}-\boldsymbol{X}E\left[\boldsymbol{B}\right]). \tag{74}$$

2.5 Update $\boldsymbol{\pi}$

We need to maximize over $\boldsymbol{\pi}$

$$E\left\{\log p(\boldsymbol{Y},\boldsymbol{B},\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\Lambda}|\boldsymbol{\pi},\boldsymbol{A})\right\}, \tag{75}$$

where the expectation is with respect to q. This is equivalent to

$$\begin{aligned}
&\arg\max_{\boldsymbol{\pi}} E\left\{\log p(\boldsymbol{w}_k|\boldsymbol{\pi})\right\} && (76)\\
&=\arg\max_{\boldsymbol{\pi}} E\left\{\sum_{k=1}^{K}\sum_{t=0}^{T}w_{kt}\log(\pi_t)\right\} && (77)\\
&=\arg\max_{\boldsymbol{\pi}} \sum_{t=0}^{T}\left(\sum_{k=1}^{K}E[w_{kt}]\right)\log(\pi_t). && (78)
\end{aligned}$$

We have

$$\pi_t=\frac{\sum_{k=1}^{K}E[w_{kt}]}{\sum_{t=0}^{T}\sum_{k=1}^{K}E[w_{kt}]}, \tag{79}$$

where $E[w_{kt}]=\gamma_{kt}$ from (47).

2.6 Expectations for VEM

There are a lot of expectations we need to calculate for our VEM. I collect them all here.

First, we review the variational densities

$$\begin{aligned}
q(\boldsymbol{Z}) &= \underset{N\times R}{\mathrm{N}}(\boldsymbol{Z}|\boldsymbol{\mu}_{\boldsymbol{Z}},\boldsymbol{\Sigma}_{\boldsymbol{Z}}\otimes\boldsymbol{I}_N) && (80)\\
q(\boldsymbol{b}_k|\boldsymbol{w}_k) &= \prod_{t=0}^{T}\left[\underset{M}{\mathrm{N}}(\boldsymbol{b}_k|\boldsymbol{\mu}_{kt},\boldsymbol{\Sigma}_{kt})\right]^{w_{kt}} && (81)\\
q(\boldsymbol{w}_k) &= \prod_{t=0}^{T}\gamma_{kt}^{w_{kt}} && (82)\\
q(\lambda_m) &= \mathrm{Gamma}(\lambda_m|\alpha_m,\beta_m). && (83)
\end{aligned}$$

$$\begin{aligned}
E[\boldsymbol{b}_k] &= \sum_{t=0}^{T} \gamma_{kt} \boldsymbol{\mu}_{kt}, && (84)\\
\boldsymbol{\mu}_{\boldsymbol{B}} := E[\boldsymbol{B}] &= \begin{pmatrix} E[\boldsymbol{b}_1]^\intercal \\ \vdots \\ E[\boldsymbol{b}_K]^\intercal \end{pmatrix}, && (85)\\
E[\boldsymbol{b}_k \boldsymbol{b}_k^\intercal] &= \sum_{t=0}^{T} \gamma_{kt} \left(\boldsymbol{\mu}_{kt} \boldsymbol{\mu}_{kt}^\intercal + \boldsymbol{\Sigma}_{kt} \right), && (86)\\
E[\boldsymbol{Z}] &= \boldsymbol{\mu}_{\boldsymbol{Z}}, && (87)\\
E[\boldsymbol{Z}^\intercal \boldsymbol{Z}] &= \sum_{r=1}^{R} E[\boldsymbol{z}_r \boldsymbol{z}_r^\intercal] = \sum_{r=1}^{R} \left[\boldsymbol{\mu}_{\boldsymbol{Z}r} \boldsymbol{\mu}_{\boldsymbol{Z}r}^\intercal + \boldsymbol{\Sigma}_{\boldsymbol{Z}} \right] = \boldsymbol{\mu}_{\boldsymbol{Z}}^\intercal \boldsymbol{\mu}_{\boldsymbol{Z}} + R \boldsymbol{\Sigma}_{\boldsymbol{Z}}, && (88)\\
E[\boldsymbol{R}_{-k}] &= \boldsymbol{Y} - E[\boldsymbol{Z}]A - \boldsymbol{X} E[\boldsymbol{B}] + \boldsymbol{x}_k E[\boldsymbol{b}_k]^\intercal && (89)\\
E[w_{kt}] &= \gamma_{kt}, && (90)\\
E[\lambda_m] &= \frac{N/2 + \alpha}{\delta_m/2 + \beta}, && (91)\\
& && (92)
\end{aligned}$$

where in (88) $\boldsymbol{z}_r$ is the rth row of $\boldsymbol{Z}$ and $\boldsymbol{\mu}_{\boldsymbol{Z}r}$ is the rth row of $\boldsymbol{\mu}_{\boldsymbol{Z}}$.

The last expectation we need is for $\boldsymbol{\delta}$ in (102), which we reproduce here for convenience

$$\begin{aligned}
\boldsymbol{\delta} =& E\left[\operatorname{diag}\left\{ (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B} - \boldsymbol{Z}\boldsymbol{A})^\intercal (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B} - \boldsymbol{Z}\boldsymbol{A}) \right\} \right] && (93)\\
=& \operatorname{diag}(\boldsymbol{Y}^\intercal \boldsymbol{Y}) - 2 \operatorname{diag}(\boldsymbol{Y}^\intercal \boldsymbol{X} E[\boldsymbol{B}]) - 2 \operatorname{diag}(\boldsymbol{Y}^\intercal E[\boldsymbol{Z}] \boldsymbol{A}) + 2 \operatorname{diag}(E[\boldsymbol{B}]^\intercal \boldsymbol{X}^\intercal E[\boldsymbol{Z}] \boldsymbol{A}) \\
&+ E[\operatorname{diag}(\boldsymbol{B}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X} \boldsymbol{B})] + \operatorname{diag}(\boldsymbol{A}^\intercal E[\boldsymbol{Z}^\intercal \boldsymbol{Z}] \boldsymbol{A}) && (94)\\
=& \operatorname{diag}(\boldsymbol{Y}^\intercal \boldsymbol{Y}) - 2 \operatorname{diag}(\boldsymbol{Y}^\intercal \boldsymbol{X} \boldsymbol{\mu}_{\boldsymbol{B}}) - 2 \operatorname{diag}(\boldsymbol{Y}^\intercal \boldsymbol{\mu}_{\boldsymbol{Z}} \boldsymbol{A}) + 2 \operatorname{diag}(\boldsymbol{\mu}_{\boldsymbol{B}}^\intercal \boldsymbol{X}^\intercal \boldsymbol{\mu}_{\boldsymbol{Z}} \boldsymbol{A}) \\
&+ E[\operatorname{diag}(\boldsymbol{B}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X} \boldsymbol{B})] + \operatorname{diag}(\boldsymbol{A}^\intercal \{ \boldsymbol{\mu}_{\boldsymbol{Z}}^\intercal \boldsymbol{\mu}_{\boldsymbol{Z}} + R \boldsymbol{\Sigma}_{\boldsymbol{Z}} \} \boldsymbol{A}) && (95)
\end{aligned}$$

Let $\boldsymbol{S} := \boldsymbol{X}^\intercal \boldsymbol{X}$. We focus on

$$\begin{aligned}
E[\operatorname{diag}(\boldsymbol{B}^\intercal \boldsymbol{X}^\intercal \boldsymbol{X} \boldsymbol{B})]_{ii} &= E[\operatorname{diag}(\boldsymbol{B}^\intercal \boldsymbol{S} \boldsymbol{B})]_{ii} && (96)\\
&= E\left[\sum_{k=1}^{K} \sum_{\ell=1}^{K} b_{ki} s_{k\ell} b_{\ell i} \right] && (97)\\
&= \sum_{k=1}^{K} \sum_{\ell=1}^{K} s_{k\ell} E\left[b_{\ell i} b_{ki} \right] && (98)\\
&= \sum_{k=1}^{K} \sum_{\ell=1}^{K} s_{k\ell} E\left[b_{\ell i} \right] E\left[b_{ki} \right] + \sum_{k=1}^{K} s_{kk} \operatorname{var}(b_{ki}). && (99)
\end{aligned}$$

Let

$$\boldsymbol{H} := \operatorname{diag}\left(\left[\sum_{k=1}^{K} s_{kk}\operatorname{var}(b_{km})\right]_{m=1}^{M}\right). \tag{100}$$

Then

$$E[\operatorname{diag}(\boldsymbol{B}^{\intercal}\boldsymbol{X}^{\intercal}\boldsymbol{X}\boldsymbol{B})] = E\left[\operatorname{diag}\left(\boldsymbol{\mu_B}\boldsymbol{S}\boldsymbol{\mu_B}\right)\right] + \operatorname{diag}(\boldsymbol{H}). \tag{101}$$

Hence,

$$\boldsymbol{\delta} = \operatorname{diag}\left\{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu_B} - \boldsymbol{\mu_Z}\boldsymbol{A})^{\intercal}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu_B} - \boldsymbol{\mu_Z}\boldsymbol{A})\right\} + R\operatorname{diag}(\boldsymbol{A}^{\intercal}\boldsymbol{\Sigma_Z}\boldsymbol{A}) + \operatorname{diag}(\boldsymbol{H}). \tag{102}$$

We can find $\operatorname{var}(b_{km})$ by noting that marginally each b_{km} is a mixture of univariate normals with mixing weights γ_{kt} mixing means μ_{ktm} and mixing variances $\boldsymbol{\Sigma}_{kt[m,m]}$. Hence, we have

$$\operatorname{var}(b_{km}) = \sum_{t=0}^{T}\gamma_{kt}\left[\left(\mu_{ktm} - E[b_{km}]\right)^2 + \boldsymbol{\Sigma}_{kt[m,m]}\right], \tag{103}$$

and $E[b_{km}]$ is the (k,m)th element of $\boldsymbol{\mu_B}$. Hence

$$\boldsymbol{H} := \operatorname{diag}\left(\left[\sum_{k=1}^{K} s_{kk}\sum_{t=0}^{T}\gamma_{kt}\left[\left(\mu_{ktm} - E[b_{km}]\right)^2 + \boldsymbol{\Sigma}_{kt[m,m]}\right]\right]_{m=1}^{M}\right). \tag{104}$$

3 The Evidence Lower Bound

In this section, we drive the formula for the Evidence Lower BOund (ELBO).

$$\begin{aligned}\mathcal{L}(q,\boldsymbol{\pi},\boldsymbol{A}) =& E\left[\log\left\{\frac{p(\boldsymbol{Y},\boldsymbol{B},\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\Lambda}|\boldsymbol{\pi},\boldsymbol{A})}{q(\boldsymbol{B},\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\Lambda})}\right\}\right], &(105)\\ =& E\left[\log p(\boldsymbol{Y}|\boldsymbol{B},\boldsymbol{Z},\boldsymbol{A},\boldsymbol{\Lambda}) + \log p(\boldsymbol{B},\boldsymbol{W}|\boldsymbol{\pi}) + \log p(\boldsymbol{Z}) + \log p(\boldsymbol{\Lambda})\right.\\ &\left.- \log q(\boldsymbol{B},\boldsymbol{W}) - \log q(\boldsymbol{Z}) - \log q(\boldsymbol{\Lambda}))\right], &(106)\end{aligned}$$

where the expectation is with respect to the density $q(\boldsymbol{B},\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\Lambda})$. Hence, we have seven expectations we need to calculate. We'll ignore any constants that do not depend on the parameters that index $q(\boldsymbol{B},\boldsymbol{W},\boldsymbol{Z},\boldsymbol{\Lambda})$.

1. $E[\log p(\boldsymbol{Z})]$

$$\begin{aligned}E[\log p(\boldsymbol{Z})] &\propto -\frac{1}{2}E\left[\operatorname{tr}(\boldsymbol{Z}^{\intercal}\boldsymbol{Z})\right] &(107)\\ &= -\frac{1}{2}\operatorname{tr}(\boldsymbol{\mu}_{\boldsymbol{Z}}^{\intercal}\boldsymbol{\mu_Z}) - \frac{R}{2}\operatorname{tr}(\boldsymbol{\Sigma_Z}) &(108)\end{aligned}$$

2. $E[\log p(\mathbf{\Lambda})]$

$$E[\log p(\mathbf{\Lambda})] \propto \sum_{m=1}^{M} (\alpha - 1) E[\log(\lambda_m)] - \sum_{m=1}^{M} \beta E[\lambda_m] \tag{109}$$

$$= \sum_{m=1}^{M} (\alpha - 1)[\psi(\alpha_m) - \log(\beta_m)] - \sum_{m=1}^{M} \beta \alpha_m / \beta_m, \tag{110}$$

where $\psi(\cdot)$ is the digamma function.

3. $E[\log p(\boldsymbol{B}, \boldsymbol{W} | \boldsymbol{\pi})]$

$$E[\log p(\boldsymbol{B}, \boldsymbol{W} | \boldsymbol{\pi})] = E\left[\sum_{k=1}^{K}\sum_{t=0}^{T} w_{kt} \left\{\log(\pi_t) - \frac{M}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{V}_t) - \frac{1}{2}\boldsymbol{b}_k^{\intercal}\boldsymbol{V}_t^{-1}\boldsymbol{b}_k\right\}\right] \tag{111}$$

$$= \sum_{k=1}^{K}\sum_{t=0}^{T} \left(E\left[w_{kt}\right] \left\{\log(\pi_t) - \frac{M}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{V}_t)\right\} - \frac{1}{2} E\left[w_{kt}\boldsymbol{b}_k^{\intercal}\boldsymbol{V}_t^{-1}\boldsymbol{b}_k\right]\right) \tag{112}$$

$$= \sum_{k=1}^{K}\sum_{t=0}^{T} \left(\gamma_{kt} \left\{\log(\pi_t) - \frac{M}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{V}_t)\right\} - \frac{1}{2} E\left[w_{kt}\boldsymbol{b}_k^{\intercal}\boldsymbol{V}_t^{-1}\boldsymbol{b}_k\right]\right) \tag{113}$$

We work with

$$E\left[w_{kt}\boldsymbol{b}_k^{\intercal}\boldsymbol{V}_t^{-1}\boldsymbol{b}_k\right] = E\left[E\left[w_{kt}\boldsymbol{b}_k^{\intercal}\boldsymbol{V}_t^{-1}\boldsymbol{b}_k | w_{kt}\right]\right] \tag{114}$$

$$= E\left[E\left[w_{kt}\,\mathrm{tr}\left(\boldsymbol{b}_k^{\intercal}\boldsymbol{V}_t^{-1}\boldsymbol{b}_k\right) | w_{kt}\right]\right] \tag{115}$$

$$= E\left[E\left[w_{kt}\,\mathrm{tr}\left(\boldsymbol{V}_t^{-1}\boldsymbol{b}_k\boldsymbol{b}_k^{\intercal}\right) | w_{kt}\right]\right] \tag{116}$$

$$= \mathrm{tr}\left(E\left[\boldsymbol{V}_t^{-1} E\left[w_{kt}\boldsymbol{b}_k\boldsymbol{b}_k^{\intercal} | w_{kt}\right]\right]\right) \tag{117}$$

$$= \mathrm{tr}\left(E\left[\boldsymbol{V}_t^{-1} w_{kt}(\boldsymbol{\mu}_{kt}\boldsymbol{\mu}_{kt}^{\intercal} + \boldsymbol{\Sigma}_{kt})\right]\right) \tag{118}$$

$$= \gamma_{kt}\,\mathrm{tr}\left(\boldsymbol{V}_t^{-1}(\boldsymbol{\mu}_{kt}\boldsymbol{\mu}_{kt}^{\intercal} + \boldsymbol{\Sigma}_{kt})\right). \tag{119}$$

Hence

$$(113) = \sum_{k=1}^{K}\sum_{t=0}^{T} \gamma_{kt} \left(\log(\pi_t) - \frac{M}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{V}_t) - \frac{1}{2}\mathrm{tr}\left[\boldsymbol{V}_t^{-1}(\boldsymbol{\mu}_{kt}\boldsymbol{\mu}_{kt}^{\intercal} + \boldsymbol{\Sigma}_{kt})\right]\right) \tag{120}$$

4. $E\left[\log p(\boldsymbol{Y}|\boldsymbol{B},\boldsymbol{Z},\boldsymbol{A},\boldsymbol{\Lambda})\right]$

$$E\left[\log p(\boldsymbol{Y}|\boldsymbol{B},\boldsymbol{Z},\boldsymbol{A},\boldsymbol{\Lambda})\right] \propto \sum_{m=1}^{M} \frac{N}{2} E\left[\log(\lambda_m)\right] - \frac{1}{2} E\left\{\operatorname{tr}\left[(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}-\boldsymbol{Z}\boldsymbol{A})\boldsymbol{\Lambda}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}-\boldsymbol{Z}\boldsymbol{A})^{\mathsf{T}}\right]\right\} \tag{121}$$

We focus on

$$E\left\{\operatorname{tr}\left[(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}-\boldsymbol{Z}\boldsymbol{A})\boldsymbol{\Lambda}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}-\boldsymbol{Z}\boldsymbol{A})^{\mathsf{T}}\right]\right\} \tag{122}$$

$$\begin{aligned} &= \operatorname{tr}(\boldsymbol{Y}E[\boldsymbol{\Lambda}]\boldsymbol{Y}) - 2\operatorname{tr}(\boldsymbol{Y}E[\boldsymbol{\Lambda}]E[\boldsymbol{B}]^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}}) - 2\operatorname{tr}(\boldsymbol{Y}E[\boldsymbol{\Lambda}]\boldsymbol{A}^{\mathsf{T}}E[\boldsymbol{Z}^{\mathsf{T}}]) \\ &+ 2\operatorname{tr}(\boldsymbol{X}E[\boldsymbol{B}]E[\boldsymbol{\Lambda}]\boldsymbol{A}^{\mathsf{T}}E[\boldsymbol{Z}]^{\mathsf{T}}) + E[\operatorname{tr}(\boldsymbol{X}\boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}})] + E[\operatorname{tr}(\boldsymbol{Z}\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{Z}^{\mathsf{T}})] \end{aligned} \tag{123}$$

All of these expectations are easy except

$$E[\operatorname{tr}(\boldsymbol{X}\boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}})] = E[\operatorname{tr}(\boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})] \tag{124}$$

$$= E[\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{S}\boldsymbol{B})] \tag{125}$$

$$= E\left[\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{\ell=1}^{K} \lambda_m b_{km} s_{k\ell} b_{\ell m}\right] \tag{126}$$

$$= \sum_{k=1}^{K}\sum_{\ell=1}^{K} s_{k\ell} E\left[\sum_{m=1}^{M} \lambda_m b_{\ell m} b_{km}\right] \tag{127}$$

$$= \sum_{k=1}^{K}\sum_{\ell=1}^{K} s_{k\ell} E\left[\boldsymbol{b}_\ell^{\mathsf{T}} E[\boldsymbol{\Lambda}]\boldsymbol{b}_k\right] \tag{128}$$

$$= \sum_{k=1}^{K}\sum_{\ell=1}^{K} s_{k\ell} \operatorname{tr}\left(E[\boldsymbol{\Lambda}]E\left[\boldsymbol{b}_k\boldsymbol{b}_\ell^{\mathsf{T}}\right]\right) \tag{129}$$

$$\begin{aligned} &= \sum_{k=1}^{K}\sum_{\ell=1}^{K} s_{k\ell} \operatorname{tr}\left(E[\boldsymbol{\Lambda}]E\left[\boldsymbol{b}_k\right]E\left[\boldsymbol{b}_\ell\right]^{\mathsf{T}}\right) - \sum_{k=1}^{K} s_{kk} \operatorname{tr}\left(E[\boldsymbol{\Lambda}]E\left[\boldsymbol{b}_k\right]E\left[\boldsymbol{b}_k\right]^{\mathsf{T}}\right) \\ &+ \sum_{k=1}^{K} s_{kk} \operatorname{tr}\left(E[\boldsymbol{\Lambda}]E\left[\boldsymbol{b}_k\boldsymbol{b}_k^{\mathsf{T}}\right]\right) \end{aligned} \tag{130}$$

$$\begin{aligned} &= \sum_{k=1}^{K}\sum_{\ell=1}^{K} s_{k\ell} E\left[\boldsymbol{b}_\ell\right]^{\mathsf{T}} E[\boldsymbol{\Lambda}]E\left[\boldsymbol{b}_k\right] - \sum_{k=1}^{K} s_{kk} E\left[\boldsymbol{b}_k\right]^{\mathsf{T}} E[\boldsymbol{\Lambda}]E\left[\boldsymbol{b}_k\right] \\ &+ \sum_{k=1}^{K} s_{kk} \operatorname{tr}\left(E[\boldsymbol{\Lambda}]E\left[\boldsymbol{b}_k\boldsymbol{b}_k^{\mathsf{T}}\right]\right). \end{aligned} \tag{131}$$

We already calculated $E\left[\boldsymbol{b}_k\right]$ from (84) and $E\left[\boldsymbol{b}_k\boldsymbol{b}_k^{\mathsf{T}}\right]$ from (86).

We go through

$$E[\operatorname{tr}(\boldsymbol{Z}\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{Z}^{\mathsf{T}})] = \operatorname{tr}(\boldsymbol{A}E[\boldsymbol{\Lambda}]\boldsymbol{A}^{\mathsf{T}}E[\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}]) \tag{132}$$

$$= \operatorname{tr}\left(\boldsymbol{A}E[\boldsymbol{\Lambda}]\boldsymbol{A}^{\intercal}\left[\boldsymbol{\mu}_{\boldsymbol{Z}}^{\intercal}\boldsymbol{\mu}_{\boldsymbol{Z}} + R\boldsymbol{\Sigma}_{\boldsymbol{Z}}\right]\right) \tag{133}$$

$$= \operatorname{tr}\left(\boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A}E[\boldsymbol{\Lambda}]\boldsymbol{A}^{\intercal}\boldsymbol{\mu}_{\boldsymbol{Z}}^{\intercal}\right) + \operatorname{tr}\left(\boldsymbol{A}E[\boldsymbol{\Lambda}]\boldsymbol{A}^{\intercal}\boldsymbol{\Sigma}_{\boldsymbol{Z}}\right) \tag{134}$$

Hence

$$E\left\{\operatorname{tr}\left[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B} - \boldsymbol{Z}\boldsymbol{A})\boldsymbol{\Lambda}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B} - \boldsymbol{Z}\boldsymbol{A})^{\intercal}\right]\right\} \tag{135}$$

$$= \operatorname{tr}\left[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}} - \boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A})E[\boldsymbol{\Lambda}](\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}} - \boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A})^{\intercal}\right] + \operatorname{tr}\left(\boldsymbol{A}E[\boldsymbol{\Lambda}]\boldsymbol{A}^{\intercal}\boldsymbol{\Sigma}_{\boldsymbol{Z}}\right)$$
$$- \sum_{k=1}^{K} s_{kk}E\left[\boldsymbol{b}_k\right]^{\intercal}E[\boldsymbol{\Lambda}]E\left[\boldsymbol{b}_k\right] + \sum_{k=1}^{K} s_{kk}\operatorname{tr}\left(E[\boldsymbol{\Lambda}]E\left[\boldsymbol{b}_k\boldsymbol{b}_k^{\intercal}\right]\right) \tag{136}$$

5. $E[\log q(\boldsymbol{B}, \boldsymbol{W})]$

$$E[\log q(\boldsymbol{B}, \boldsymbol{W})] = E\left[\sum_{k=1}^{K}\sum_{t=0}^{T} w_{kt}\left\{\log(\gamma_{kt}) - \frac{M}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{\Sigma}_{kt}) - \frac{1}{2}\left(\boldsymbol{b}_k - \boldsymbol{\mu}_{kt}\right)^{\intercal}\boldsymbol{\Sigma}_{kt}^{-1}\left(\boldsymbol{b}_k - \boldsymbol{\mu}_{kt}\right)\right\}\right] \tag{137}$$

$$= \sum_{k=1}^{K}\sum_{t=0}^{T}\left(E\left[w_{kt}\right]\left\{\log(\gamma_{kt}) - \frac{M}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{\Sigma}_{kt})\right\} - \frac{1}{2}E\left[w_{kt}\left(\boldsymbol{b}_k - \boldsymbol{\mu}_{kt}\right)^{\intercal}\boldsymbol{\Sigma}_{kt}^{-1}\left(\boldsymbol{b}_k - \boldsymbol{\mu}_{kt}\right)\right]\right) \tag{138}$$

$$= \sum_{k=1}^{K}\sum_{t=0}^{T}\left(\gamma_{kt}\left\{\log(\gamma_{kt}) - \frac{M}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{\Sigma}_{kt})\right\} - \frac{1}{2}E\left[w_{kt}\left(\boldsymbol{b}_k - \boldsymbol{\mu}_{kt}\right)^{\intercal}\boldsymbol{\Sigma}_{kt}^{-1}\left(\boldsymbol{b}_k - \boldsymbol{\mu}_{kt}\right)\right]\right). \tag{139}$$

We focus on

$$E\left[w_{kt}\left(\boldsymbol{b}_k - \boldsymbol{\mu}_{kt}\right)^{\intercal}\boldsymbol{\Sigma}_{kt}^{-1}\left(\boldsymbol{b}_k - \boldsymbol{\mu}_{kt}\right)\right] = E\left[E\left[w_{kt}\left(\boldsymbol{b}_k - \boldsymbol{\mu}_{kt}\right)^{\intercal}\boldsymbol{\Sigma}_{kt}^{-1}\left(\boldsymbol{b}_k - \boldsymbol{\mu}_{kt}\right)|w_{kt}\right]\right] \tag{140}$$

$$= E\left[E\left[w_{kt}\left(\boldsymbol{b}_k - \boldsymbol{\mu}_{kt}\right)^{\intercal}\boldsymbol{\Sigma}_{kt}^{-1}\left(\boldsymbol{b}_k - \boldsymbol{\mu}_{kt}\right)|w_{kt}\right]\right] \tag{141}$$

$$= E\left[w_{kt}E\left[\left(\boldsymbol{\Sigma}_{kt}^{-1/2}\boldsymbol{b}_k - \boldsymbol{\Sigma}_{kt}^{-1/2}\boldsymbol{\mu}_{kt}\right)^{\intercal}\left(\boldsymbol{\Sigma}_{kt}^{-1/2}\boldsymbol{b}_k - \boldsymbol{\Sigma}_{kt}^{-1/2}\boldsymbol{\mu}_{kt}\right)|w_{kt}\right]\right] \tag{142}$$

$$= E\left[w_{kt}M\right] \tag{143}$$

$$= M\gamma_{kt}. \tag{144}$$

Equation (143) follows because we treat $\boldsymbol{\Sigma}_{kt}^{-1/2}\boldsymbol{b}_k$ as following a $\mathrm{N}(\boldsymbol{\Sigma}_{kt}^{-1/2}\boldsymbol{\mu}_{kt}, \boldsymbol{I}_M)$, then the expectation in (142) becomes the summation of the M variances, each of which is 1. Hence

$$(139) = \sum_{k=1}^{K}\sum_{t=0}^{T}\gamma_{kt}\left(\log(\gamma_{kt}) - \frac{M}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{\Sigma}_{kt}) - \frac{M}{2}\right). \tag{145}$$

6. $E[\log q(\boldsymbol{Z})]$

$$E[\log q(\boldsymbol{Z})] \propto -\frac{N}{2}\log\det(\boldsymbol{\Sigma}_{\boldsymbol{Z}}) - \frac{1}{2}E\left\{\operatorname{tr}\left[(\boldsymbol{Z}-\boldsymbol{\mu}_{\boldsymbol{Z}})\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}(\boldsymbol{Z}-\boldsymbol{\mu}_{\boldsymbol{Z}})^{\intercal}\right]\right\} \tag{146}$$

$$= -\frac{N}{2}\log\det(\boldsymbol{\Sigma}_{\boldsymbol{Z}}) - \frac{1}{2}E\left\{\operatorname{tr}\left[(\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1/2}-\boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1/2})(\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1/2}-\boldsymbol{\mu}_{\boldsymbol{Z}})^{\intercal}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1/2}\right]\right\} \tag{147}$$

$$= -\frac{N}{2}\log\det(\boldsymbol{\Sigma}_{\boldsymbol{Z}}) - \frac{NR}{2} \tag{148}$$

$$= -\frac{N}{2}\log\det(\boldsymbol{\Sigma}_{\boldsymbol{Z}}). \tag{149}$$

Equation (148) follows because $(\boldsymbol{Z}\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1/2}-\boldsymbol{\mu}_{\boldsymbol{Z}})$ is $\mathrm{N}_{N\times R}(\boldsymbol{0}, \boldsymbol{I}_R \otimes \boldsymbol{I}_N)$ and so the trace is the sum of NR variances, all of which are 1.

7. $E[\log q(\boldsymbol{\Lambda}))]$

$$E[\log q(\boldsymbol{\Lambda}))] = \sum_{m=1}^{M}\left(\alpha_m - \log(\beta_m) + \log(\Gamma(\alpha_m)) + (1-\alpha_m)\psi(\alpha_m)\right). \tag{150}$$

This is because we just have M calculations of the entropy of Gamma(α_m, β_m), which is well-known. $\Gamma(\cdot)$ is the gamma function and $\psi(\cdot)$ is the di-gamma function.

ELBO

$$\begin{aligned}
\mathcal{L}(q,\boldsymbol{\pi},\boldsymbol{A}) \propto & \sum_{m=1}^{M}\frac{N}{2}E\left[\log(\lambda_m)\right] - \frac{1}{2}\operatorname{tr}\left[(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}}-\boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A})E[\boldsymbol{\Lambda}](\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}}-\boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A})^{\intercal}\right] \\
& - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{A}E[\boldsymbol{\Lambda}]\boldsymbol{A}^{\intercal}\boldsymbol{\Sigma}_{\boldsymbol{Z}}\right) + \frac{1}{2}\sum_{k=1}^{K}s_{kk}E\left[\boldsymbol{b}_k\right]^{\intercal}E[\boldsymbol{\Lambda}]E\left[\boldsymbol{b}_k\right] - \frac{1}{2}\sum_{k=1}^{K}s_{kk}\operatorname{tr}\left(E[\boldsymbol{\Lambda}]E\left[\boldsymbol{b}_k\boldsymbol{b}_k^{\intercal}\right]\right) \\
& - \frac{1}{2}\operatorname{tr}(\boldsymbol{\mu}_{\boldsymbol{Z}}^{\intercal}\boldsymbol{\mu}_{\boldsymbol{Z}}) - \frac{R}{2}\operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{Z}}) \\
& + \sum_{m=1}^{M}(\alpha-1)[\psi(\alpha_m)-\log(\beta_m)] - \sum_{m=1}^{M}\beta\alpha_m/\beta_m \\
& + \sum_{k=1}^{K}\sum_{t=0}^{T}\gamma_{kt}\left(\log(\pi_t) - \frac{M}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{V}_t) - \frac{1}{2}\operatorname{tr}\left[\boldsymbol{V}_t^{-1}(\boldsymbol{\mu}_{kt}\boldsymbol{\mu}_{kt}^{\intercal}+\boldsymbol{\Sigma}_{kt})\right]\right) \\
& - \sum_{k=1}^{K}\sum_{t=0}^{T}\gamma_{kt}\left(\log(\gamma_{kt}) - \frac{M}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{\Sigma}_{kt}) - \frac{M}{2}\right) \\
& + \frac{N}{2}\log\det(\boldsymbol{\Sigma}_{\boldsymbol{Z}}) \\
& - \sum_{m=1}^{M}\left(\alpha_m - \log(\beta_m) + \log(\Gamma(\alpha_m)) + (1-\alpha_m)\psi(\alpha_m)\right)
\end{aligned} \tag{151}$$

4 Missing Data

Let $\mathcal{C} \subset \{(n, m) : n = 1, \ldots, N, m = 1, \ldots, M\}$ be the set of indicies such that y_{nm} is missing for $(n, m) \in \mathcal{C}$. We treat these y_{nm} as any other parameter and estimate them by mean-field variational Bayes. That is, we assume that

$$q(\boldsymbol{B}, \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Lambda}, \boldsymbol{Y}_{\mathcal{C}}) = q(\boldsymbol{Z})q(\boldsymbol{\Lambda}) \left(\prod_{k=1}^{K} q(\boldsymbol{b}_k, \boldsymbol{w}_k)\right) \left(\prod_{(n,m)\in\mathcal{C}} q(y_{nm})\right), \tag{152}$$

and we maximize the lower bound of the marginal log-likelihood by

$$\mathcal{L}(q, \boldsymbol{\pi}, \boldsymbol{A}) = \int q(\boldsymbol{B}, \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Lambda}, \boldsymbol{Y}_{\mathcal{C}}) \log \left\{ \frac{p(\boldsymbol{Y}, \boldsymbol{B}, \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Lambda}|\boldsymbol{\pi}, \boldsymbol{A})}{q(\boldsymbol{B}, \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Lambda}, \boldsymbol{Y}_{\mathcal{C}})} \right\} \mathrm{d}\boldsymbol{B}\, \mathrm{d}\boldsymbol{W}\, \mathrm{d}\boldsymbol{Z}\, \mathrm{d}\boldsymbol{\Lambda}\, \mathrm{d}\boldsymbol{Y}_{\mathcal{C}}. \tag{153}$$

We update $q(y_{nm})$ using the same standard result for mean-field variational inference that we have been using this whole time

$$\begin{aligned}
q(y_{nm}) &\propto E_{-y_{nm}} \left\{ \log p(\boldsymbol{Y}, \boldsymbol{B}, \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Lambda}|\boldsymbol{\pi}, \boldsymbol{A}) \right\} && (154)\\
&\propto E\left[\log p(\boldsymbol{Y}|\boldsymbol{B}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{\Lambda})\right] && (155)\\
&\propto -\frac{1}{2} E\left\{ \mathrm{tr}\left[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B} - \boldsymbol{Z}\boldsymbol{A})\boldsymbol{\Lambda}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B} - \boldsymbol{Z}\boldsymbol{A})^{\intercal}\right] \right\} && (156)\\
&\propto -\frac{1}{2} E\left\{ \mathrm{tr}\left[\boldsymbol{Y}\boldsymbol{\Lambda}\boldsymbol{Y}^{\intercal}\right] - 2\,\mathrm{tr}\left[\boldsymbol{Y}\boldsymbol{\Lambda}(\boldsymbol{X}\boldsymbol{B} + \boldsymbol{Z}\boldsymbol{A})^{\intercal}\right] \right\} && (157)\\
&\propto -\frac{1}{2} \left\{ E[\lambda_m] y_{nm}^2 - 2E[\lambda_m] y \nu_{nm} \right\}, && (158)
\end{aligned}$$

where ν_{nm} is the (n, m)th element of $(\boldsymbol{X}E[\boldsymbol{B}] + E[\boldsymbol{Z}]\boldsymbol{A})$. Hence

$$q(y_{nm}) = \mathrm{N}(y_{nm}|\nu_{nm}, E[\lambda_m]^{-1}), \tag{159}$$

which is just the marginal distribution of y_{nm}.

Parameterize $q(y_{nm})$ by $N(\nu_{nm}, \tau_{nm}^2)$. The updates for all parameters *except* $\boldsymbol{\Lambda}$ are the exact same as in the non-missing case by replacing $\boldsymbol{Y}$ by $E[\boldsymbol{Y}]$. That is, by setting $y_{nm} < -\nu_{nm}$. For $\boldsymbol{\Lambda}$ we need to update the definition of $\boldsymbol{\delta}$. Specifically, we need to look at

$$\begin{aligned}
E[\mathrm{diag}(\boldsymbol{Y}^{\intercal}\boldsymbol{Y})]_m &= \sum_{n=1}^{N} E[y_{nm}^2] && (160)\\
&= \sum_{n:(n,m)\notin\mathcal{C}} y_{nm}^2 + \sum_{n:(n,m)\in\mathcal{C}} (\nu_{nm}^2 + \tau_{nm}^2). && (161)
\end{aligned}$$

Hence, $\boldsymbol{\delta}$ becomes

$$\boldsymbol{\delta} := \mathrm{diag}\left\{ (E[\boldsymbol{Y}] - \boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}} - \boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A})^{\intercal}(E[\boldsymbol{Y}] - \boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}} - \boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A}) \right\} + R\,\mathrm{diag}(\boldsymbol{A}^{\intercal}\boldsymbol{\Sigma}_{\boldsymbol{Z}}\boldsymbol{A}) + \mathrm{diag}(\boldsymbol{H}) + \boldsymbol{g} \tag{162}$$

where

$$g_m = \sum_{n:(n,m)\in\mathcal{C}} \tau_{nm}^2. \tag{163}$$

Only one term in the ELBO needs to be updated in the presence of missing y_{nm}.

$$E\left[\log p(\boldsymbol{Y}|\boldsymbol{B},\boldsymbol{Z},\boldsymbol{A},\boldsymbol{\Lambda})\right] \propto \sum_{m=1}^{M} \frac{N}{2} E\left[\log(\lambda_m)\right] - \frac{1}{2} E\left\{\operatorname{tr}\left[(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}-\boldsymbol{Z}\boldsymbol{A})\boldsymbol{\Lambda}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}-\boldsymbol{Z}\boldsymbol{A})^\intercal\right]\right\}. \tag{164}$$

Of this term, we need only modify

$$\begin{aligned} E[\operatorname{tr}(\boldsymbol{Y}\boldsymbol{\Lambda}\boldsymbol{Y})] &= \sum_{n=1}^{N}\sum_{m=1}^{M} E\left[\lambda_m y_{nm}^2\right] &(165)\\ &= \sum_{(n,m)\notin\mathcal{C}} E\left[\lambda_m\right] y_{nm}^2 + \sum_{(n,m)\in\mathcal{C}} E\left[\lambda_m\right] E\left[y_{nm}^2\right] &(166)\\ &= \sum_{(n,m)\notin\mathcal{C}} E\left[\lambda_m\right] y_{nm}^2 + \sum_{(n,m)\in\mathcal{C}} E\left[\lambda_m\right] E\left[\nu_{nm}^2 + \tau_{nm}^2\right]. &(167)\end{aligned}$$

Hence, we have that this term in the ELBO is equal to

$$\begin{aligned} &E\left\{\operatorname{tr}\left[(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}-\boldsymbol{Z}\boldsymbol{A})\boldsymbol{\Lambda}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B}-\boldsymbol{Z}\boldsymbol{A})^\intercal\right]\right\} &(168)\\ &= \operatorname{tr}\left[(E[\boldsymbol{Y}]-\boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}}-\boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A})E[\boldsymbol{\Lambda}](E[\boldsymbol{Y}]-\boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}}-\boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A})^\intercal\right] + \operatorname{tr}\left(\boldsymbol{A}E[\boldsymbol{\Lambda}]\boldsymbol{A}^\intercal\boldsymbol{\Sigma}_{\boldsymbol{Z}}\right)\\ &- \sum_{k=1}^{K} s_{kk} E\left[\boldsymbol{b}_k\right]^\intercal E[\boldsymbol{\Lambda}] E\left[\boldsymbol{b}_k\right] + \sum_{k=1}^{K} s_{kk} \operatorname{tr}\left(E[\boldsymbol{\Lambda}]E\left[\boldsymbol{b}_k\boldsymbol{b}_k^\intercal\right]\right) + \sum_{(n,m)\in\mathcal{C}} E\left[\lambda_m\right]\tau_{nm}^2. &(169)\end{aligned}$$

5 Initialization of Parameters

Since the VEM is only gauranteed to converge to a local optimum, it's important to choose good starting locations. In conjunction with Gao, we decided on the following path to deciding the initial values.

We first estimate $\boldsymbol{\mu}_{\boldsymbol{B}}$ by multivariate multiple regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$. That is,

$$\boldsymbol{\mu}_{\boldsymbol{B}}^{(init)} = (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal\boldsymbol{Y}. \tag{170}$$

We then initialize $\boldsymbol{\mu}_{\boldsymbol{Z}}$, $\boldsymbol{A}$, and $\boldsymbol{\Lambda}$ using the residuals of this regression.

$$\boldsymbol{E} := (\boldsymbol{I}_K - \boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal)\boldsymbol{Y}. \tag{171}$$

Let $\boldsymbol{E} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\intercal$ be the singular value decomposition of $\boldsymbol{E}$. Then we initialize

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{Z}}^{(init)} &= \sqrt{N}\boldsymbol{U}_{[,1:K]} &(172)\\ \boldsymbol{A}^{(init)} &= \boldsymbol{D}_{[1:K,1:K]}\boldsymbol{V}_{[,1:K]}^\intercal/\sqrt{N} &(173)\end{aligned}$$

$$(\lambda_1^{(init)}, \ldots, \lambda_m^{(init)}) = \text{colSums}\left[(\boldsymbol{E} - \boldsymbol{\mu}_{\boldsymbol{Z}}^{(init)}\boldsymbol{A}^{(init)})^2\right]/(N-K-R). \tag{174}$$

As a final step for the initialization of $\boldsymbol{\Lambda}$, we can apply `squeezeVar` from the `limma` package to these $(\lambda_1^{(init)}, \ldots, \lambda_m^{(init)})$ to initialize them. This way, none of them will be too small.

The intuition with multiplying $\boldsymbol{U}_{[,1:K]}$ by $\sqrt{n}$ is so that the initial value is on a similar order to that of a normal random variate, but it shouldn't really matter.

We can also use the residuals to estimate R by using, for example, parallel factor

Finally, we initialize $\boldsymbol{\pi}$ using the same approach as in ASH.

The rest of the parameters do not need to be initalized because we can estimate them using just the current initializations with one pass of Algorithm 1.

In the case of missing data, we can first use the `softImpute` R package to impute the missing data, initialize $\boldsymbol{\mu}_{\boldsymbol{B}}$, $\boldsymbol{\mu}_{\boldsymbol{Z}}$, $\boldsymbol{A}$, and $\boldsymbol{\Lambda}$, then set $\boldsymbol{Y}_{\mathcal{C}} = [\boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}} + \boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A}]_{\mathcal{C}}$ before starting the optimization program.

6 Penalize $\boldsymbol{\pi}$

Suppose we want to include the same penalization on $\boldsymbol{\pi}$ as in ASH. Let $\eta_t \geq 1$ for $t = 1, \ldots, T$. Let

$$h(\boldsymbol{\pi}|\boldsymbol{\eta}) = \prod_{t=0}^{T} \pi_t^{\eta_t - 1}. \tag{175}$$

Then the objective that we wish to maximize is

$$\mathcal{L}(q, \boldsymbol{\pi}, \boldsymbol{A}) + \log h(\boldsymbol{\pi}|\boldsymbol{\eta}). \tag{176}$$

The updates for all parameters except $\boldsymbol{\pi}$ are the exact same as before. To update $\boldsymbol{\pi}$ we need to maximize over $\boldsymbol{\pi}$

$$\begin{aligned}
&E\left\{\log p(\boldsymbol{Y}, \boldsymbol{B}, \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Lambda}|\boldsymbol{\pi}, \boldsymbol{A})\right\} + \log h(\boldsymbol{\pi}|\boldsymbol{\eta}) && (177)\\
&= \arg\max_{\boldsymbol{\pi}} \Big(E\left\{\log p(\boldsymbol{w}_k|\boldsymbol{\pi})\right\} + \log h(\boldsymbol{\pi}|\boldsymbol{\eta})\Big) && (178)\\
&= \arg\max_{\boldsymbol{\pi}} \left(E\left\{\sum_{k=1}^{K}\sum_{t=0}^{T} w_{kt}\log(\pi_t)\right\} + \sum_{t=0}^{T}(\eta_t - 1)\log(\pi_t)\right) && (179)\\
&= \arg\max_{\boldsymbol{\pi}} \sum_{t=0}^{T}\left\{\left(\sum_{k=1}^{K} E[w_{kt}]\right) + \eta_t - 1\right\}\log(\pi_t) && (180)\\
&= \arg\max_{\boldsymbol{\pi}} \sum_{t=0}^{T}\left\{\left(\sum_{k=1}^{K} \gamma_{kt}\right) + \eta_t - 1\right\}\log(\pi_t). && (181)
\end{aligned}$$

We have

$$\pi_t = \frac{\left(\sum_{k=1}^{K} \gamma_{kt}\right) + \eta_t - 1}{\sum_{t=0}^{T} \sum_{k=1}^{K} \gamma_{kt} + \sum_{t=0}^{T} \eta_t - T}. \tag{182}$$

We use the same default penalty as in ASH with $\eta_0 = 10$ and $\eta_t = 1$ for all $t \neq 0$.

7 The VEM Algorithm

The full VEM algorithm is presented in Algorithm 1. I present the details there assuming that we are estimating $\boldsymbol{\Lambda}$ by ML. A version that allows for missing y_{nm} is presented in Algorithm 2. For comparing Algorithm 2 and Section 4, we are basically using placing ν_{nm} into the $\boldsymbol{Y}$ matrix and using $\tau_{nm}^2 = 1/\lambda_m$.

8 Ideas to speed up VEM

The slowest part of the VEM will be calculating the inverse of $\boldsymbol{V}_t^{-1} + \|\boldsymbol{x}_k\|^2\boldsymbol{\Lambda}$ at every iteration. This is because $M \sim 10000$. We can significanly speed this up if we take $\boldsymbol{V}_t$ to be low rank. From what I recall, I think a lot of these $\boldsymbol{V}_t$'s are derived by empirical covariances. We can take the first 10, 20, or 50 eigenvectors of these empirical covariances and set these as the $\boldsymbol{V}_t$'s. We would then use Woodbury at every iteration. Though, if $\boldsymbol{V}_t$ is low-rank, then its inverse does not exist. Even its density does not exist (with respect to Lebesgue measure in $\mathbb{R}^M$). Sarah re-writes

$$\left[\boldsymbol{V}_t^{-1} + \|\boldsymbol{x}_k\|^2\boldsymbol{\Lambda}\right]^{-1} = \left[\boldsymbol{V}_t^{-1}\left(\boldsymbol{I}_M + \boldsymbol{V}_t\|\boldsymbol{x}_k\|^2\boldsymbol{\Lambda}\right)\right]^{-1} \tag{183}$$

$$= \left[\boldsymbol{I}_M + \boldsymbol{V}_t\|\boldsymbol{x}_k\|^2\boldsymbol{\Lambda}\right]^{-1}\boldsymbol{V}_t, \tag{184}$$

then just uses (184) as the posterior covariance. After thinking about it, I think this is correct, but the derivation is very incorrect because you can't even write the inverse of a low-rank matrix. Thus, you would need to make some limiting argument where you begin with a full rank matrix then take the limit to a low-rank matrix and use the continuity of the matrix inverse. To make computation much faster, we assume that

$$\boldsymbol{V}_t = \boldsymbol{U}_t\boldsymbol{U}_t^\intercal, \tag{185}$$

where $\boldsymbol{U}_t \in \mathbb{R}^{M\times L}$, where $L \ll M$. Then

$$(184) = \|\boldsymbol{x}_k\|^{-2}\boldsymbol{\Lambda}^{-1}\left[\|\boldsymbol{x}_k\|^{-2}\boldsymbol{\Lambda}^{-1} + \boldsymbol{V}_t\right]^{-1}\boldsymbol{V}_t \tag{186}$$

$$= \|\boldsymbol{x}_k\|^{-2}\boldsymbol{\Lambda}^{-1}\left[\|\boldsymbol{x}_k\|^{-2}\boldsymbol{\Lambda}^{-1} + \boldsymbol{U}_t\boldsymbol{U}_t^\intercal\right]^{-1}\boldsymbol{U}_t\boldsymbol{U}_t^\intercal \tag{187}$$

$$= \|\boldsymbol{x}_k\|^{-2}\boldsymbol{\Lambda}^{-1}\left[\|\boldsymbol{x}_k\|^2\boldsymbol{\Lambda} - \|\boldsymbol{x}_k\|^2\boldsymbol{\Lambda}\boldsymbol{U}_t\left(\boldsymbol{I}_L + \boldsymbol{U}_t^\intercal\|\boldsymbol{x}_k\|^2\boldsymbol{\Lambda}\boldsymbol{U}_t\right)^{-1}\boldsymbol{U}_t^\intercal\|\boldsymbol{x}_k\|^2\boldsymbol{\Lambda}\right]\boldsymbol{U}_t\boldsymbol{U}_t^\intercal \tag{188}$$

$$= \left[\boldsymbol{I}_M - \boldsymbol{U}_t\left(\boldsymbol{I}_L + \boldsymbol{U}_t^\intercal\|\boldsymbol{x}_k\|^2\boldsymbol{\Lambda}\boldsymbol{U}_t\right)^{-1}\boldsymbol{U}_t^\intercal\|\boldsymbol{x}_k\|^2\boldsymbol{\Lambda}\right]\boldsymbol{U}_t\boldsymbol{U}_t^\intercal \tag{189}$$

Algorithm 1 Variational EM algorithm for M&MASH when assuming low-rank column covariance and complete data.

```
Initialize parameters:
  $\boldsymbol{\mu}_{\boldsymbol{B}} = E[\boldsymbol{B}]$ by multivariate multiple regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$
  $\boldsymbol{\mu}_{\boldsymbol{Z}} \in \mathbb{R}^{N \times R}$ and $\boldsymbol{A}$ by a low-rank SVD on $\boldsymbol{Y} - E[\boldsymbol{B}]$.
  $\boldsymbol{\Lambda}$ by residual sums squares of $\boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{B}}\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A}$ divided by $N - K - R$. Optionally limma-shrink
      these variances.
  $\pi_t \in [0, 1]$ s.t. $\sum_{t=0}^{T} \pi_t = 1$
repeat
  Update $q(\boldsymbol{B}, \boldsymbol{W})$:
  for $k = 1, \ldots, K$ do
    Update $q(\boldsymbol{b}_k, \boldsymbol{w}_k)$:
      Let $E[\boldsymbol{R}_{-k}] := \boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A} - \boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}} + \boldsymbol{x}_k \boldsymbol{\mu}_{\boldsymbol{B}[k,]}^{\intercal}$
      Let $\boldsymbol{\xi}_k = E\left[\boldsymbol{R}_{-k}\right]^{\intercal} \boldsymbol{x}_k \|\boldsymbol{x}_k\|^{-2}$
      for $t = 1, \ldots, T$ do
            Set $\boldsymbol{\Sigma}_{kt} = \left(\boldsymbol{V}_t^{-1} + \|\boldsymbol{x}_k\|^2 \boldsymbol{\Lambda}\right)^{-1}$
            Set $\boldsymbol{\mu}_{kt} = \boldsymbol{\Sigma}_{kt} \boldsymbol{\Lambda} E\left[\boldsymbol{R}_{-k}\right]^{\intercal} \boldsymbol{x}_k$
            Set $\gamma_{kt} = \frac{\pi_t \, \mathrm{N}(\boldsymbol{\xi}_k | \boldsymbol{0}, \boldsymbol{V}_t + \boldsymbol{\Lambda}^{-1} \|\boldsymbol{x}_k\|^{-2})}{\sum_{t=0}^{T} \pi_t \, \mathrm{N}(\boldsymbol{\xi}_k | \boldsymbol{0}, \boldsymbol{V}_t + \boldsymbol{\Lambda}^{-1} \|\boldsymbol{x}_k\|^{-2})}$
      end for
      Set $\boldsymbol{\mu}_{\boldsymbol{B}[k,]} = E[\boldsymbol{b}_k] = \sum_{t=0}^{T} \gamma_{kt} \boldsymbol{\mu}_{kt}$
  end for
  Recalculate $\boldsymbol{h} = \left[\sum_{k=1}^{K} \|\boldsymbol{x}_k\|^2 \sum_{t=0}^{T} \gamma_{kt} \left[\left(\mu_{ktm} - E[b_{km}]\right)^2 + \boldsymbol{\Sigma}_{kt[m,m]}\right]\right]_{m=1}^{M}$
  Update $\boldsymbol{\pi}$:
   Set $\pi_t = \frac{\left(\sum_{k=1}^{K} \gamma_{kt}\right) + \eta_t - 1}{\sum_{t=0}^{T} \sum_{k=1}^{K} \gamma_{kt} + \sum_{t=0}^{T} \eta_t - T}$ for $t = 1, \ldots, T$
  Update $q(\boldsymbol{Z})$:
   Set $\boldsymbol{\Sigma}_{\boldsymbol{Z}} = (\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}^{\intercal} + \boldsymbol{I}_R)^{-1}$
   Set $\boldsymbol{\mu}_{\boldsymbol{Z}} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}})\boldsymbol{\Lambda}\boldsymbol{A}^{\intercal}\boldsymbol{\Sigma}_{\boldsymbol{Z}}$
  Update $\boldsymbol{\Lambda}$:
   Set $\boldsymbol{\delta} = \mathrm{diag}\left\{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}} - \boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A})^{\intercal}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}} - \boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A})\right\} + R\,\mathrm{diag}(\boldsymbol{A}^{\intercal}\boldsymbol{\Sigma}_{\boldsymbol{Z}}\boldsymbol{A}) + \boldsymbol{h}$
   Set $\lambda_m = N/\delta_m$ for $m = 1, \ldots, M$
  Update $\boldsymbol{A}$:
   Set $\boldsymbol{A} = \left(\boldsymbol{\mu}_{\boldsymbol{Z}}^{\intercal}\boldsymbol{\mu}_{\boldsymbol{Z}} + R\boldsymbol{\Sigma}_{\boldsymbol{Z}}\right)^{-1} \boldsymbol{\mu}_{\boldsymbol{Z}}^{\intercal}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}})$
until convergence
```

Algorithm 2 Variational EM algorithm for M&MASH when assuming low-rank column covariance and some missing data.

```
Initialize parameters:
  Set $\boldsymbol{Y}_{\mathcal{C}}$ to either be the overall mean of the observed $y_{nm}$'s or use softImpute to impute missing values.
  $\boldsymbol{\mu}_{\boldsymbol{B}} = E[\boldsymbol{B}]$ by multivariate multiple regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$
  $\boldsymbol{\mu}_{\boldsymbol{Z}} \in \mathbb{R}^{N \times R}$ and $\boldsymbol{A}$ by a low-rank SVD on $\boldsymbol{Y} - E[\boldsymbol{B}]$.
  $\boldsymbol{\Lambda}$ by residual sums squares of $\boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{B}}\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A}$ divided by $N-K-R$. Optionally limma-shrink these variances.
  $\pi_t \in [0,1]$ s.t. $\sum_{t=0}^{T} \pi_t = 1$
Set $\boldsymbol{Y}_{\mathcal{C}} = [\boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}} + \boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A}]_{\mathcal{C}}$
Set $\boldsymbol{g} = \left[\sum_{n:(n,m)\in\mathcal{C}} 1/\lambda_m\right]_{m=1}^{M}$
repeat
  Update $q(\boldsymbol{B}, \boldsymbol{W})$:
  for $k = 1, \ldots, K$ do
    Update $q(\boldsymbol{b}_k, \boldsymbol{w}_k)$:
      Let $E[\boldsymbol{R}_{-k}] := \boldsymbol{Y} - \boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A} - \boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}} + \boldsymbol{x}_k\boldsymbol{\mu}_{\boldsymbol{B}[k,]}^{\intercal}$
      Let $\boldsymbol{\xi}_k = E\left[\boldsymbol{R}_{-k}\right]^{\intercal}\boldsymbol{x}_k\|\boldsymbol{x}_k\|^{-2}$
      for $t = 1, \ldots, T$ do
            Set $\boldsymbol{\Sigma}_{kt} = \left(\boldsymbol{V}_t^{-1} + \|\boldsymbol{x}_k\|^2\boldsymbol{\Lambda}\right)^{-1}$
            Set $\boldsymbol{\mu}_{kt} = \boldsymbol{\Sigma}_{kt}\boldsymbol{\Lambda}E\left[\boldsymbol{R}_{-k}\right]^{\intercal}\boldsymbol{x}_k$
            Set $\gamma_{kt} = \frac{\pi_t\,\mathrm{N}(\boldsymbol{\xi}_k|\boldsymbol{0},\boldsymbol{V}_t+\boldsymbol{\Lambda}^{-1}\|\boldsymbol{x}_k\|^{-2})}{\sum_{t=0}^{T}\pi_t\,\mathrm{N}(\boldsymbol{\xi}_k|\boldsymbol{0},\boldsymbol{V}_t+\boldsymbol{\Lambda}^{-1}\|\boldsymbol{x}_k\|^{-2})}$
      end for
      Set $\boldsymbol{\mu}_{\boldsymbol{B}[k,]} = E[\boldsymbol{b}_k] = \sum_{t=0}^{T}\gamma_{kt}\boldsymbol{\mu}_{kt}$
  end for
  Recalculate $\boldsymbol{h} = \left[\sum_{k=1}^{K}\|\boldsymbol{x}_k\|^2\sum_{t=0}^{T}\gamma_{kt}\left[\left(\mu_{ktm} - E[b_{km}]\right)^2 + \boldsymbol{\Sigma}_{kt[m,m]}\right]\right]_{m=1}^{M}$
  Update $\boldsymbol{\pi}$:
    Set $\pi_t = \frac{\left(\sum_{k=1}^{K}\gamma_{kt}\right)+\eta_t-1}{\sum_{t=0}^{T}\sum_{k=1}^{K}\gamma_{kt}+\sum_{t=0}^{T}\eta_t-T}$ for $t = 1, \ldots, T$
  Update $q(\boldsymbol{Z})$:
    Set $\boldsymbol{\Sigma}_{\boldsymbol{Z}} = (\boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}^{\intercal} + \boldsymbol{I}_R)^{-1}$
    Set $\boldsymbol{\mu}_{\boldsymbol{Z}} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}})\boldsymbol{\Lambda}\boldsymbol{A}^{\intercal}\boldsymbol{\Sigma}_{\boldsymbol{Z}}$
  Update $\boldsymbol{\Lambda}$:
    $\boldsymbol{\delta} := \mathrm{diag}\left\{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}} - \boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A})^{\intercal}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}} - \boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A})\right\} + R\,\mathrm{diag}(\boldsymbol{A}^{\intercal}\boldsymbol{\Sigma}_{\boldsymbol{Z}}\boldsymbol{A}) + \boldsymbol{h} + \boldsymbol{g}$
    Set $\lambda_m = N/\delta_m$ for $m = 1, \ldots, M$.
  Recalculate $\boldsymbol{g} = \left[\sum_{n:(n,m)\in\mathcal{C}} 1/\lambda_m\right]_{m=1}^{M}$
  Update $\boldsymbol{A}$:
    Set $\boldsymbol{A} = \left(\boldsymbol{\mu}_{\boldsymbol{Z}}^{\intercal}\boldsymbol{\mu}_{\boldsymbol{Z}} + R\boldsymbol{\Sigma}_{\boldsymbol{Z}}\right)^{-1}\boldsymbol{\mu}_{\boldsymbol{Z}}^{\intercal}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}})$
  Update missing values:
    Set $\boldsymbol{Y}_{\mathcal{C}} = [\boldsymbol{X}\boldsymbol{\mu}_{\boldsymbol{B}} + \boldsymbol{\mu}_{\boldsymbol{Z}}\boldsymbol{A}]_{\mathcal{C}}$
until convergence
```

$$
\begin{aligned}
&= \boldsymbol{U}_t\boldsymbol{U}_t^\intercal - \boldsymbol{U}_t\left(\boldsymbol{I}_L + \boldsymbol{U}_t^\intercal\|\boldsymbol{x}_k\|^2\boldsymbol{\Lambda}\boldsymbol{U}_t\right)^{-1}\boldsymbol{U}_t^\intercal\|\boldsymbol{x}_k\|^2\boldsymbol{\Lambda}\boldsymbol{U}_t\boldsymbol{U}_t^\intercal && (190)\\
&= \boldsymbol{V}_t - \boldsymbol{U}_t\left(\boldsymbol{I}_L + \boldsymbol{Q}_t\right)^{-1}\boldsymbol{Q}_t\boldsymbol{U}_t^\intercal, && (191)
\end{aligned}
$$

where $\boldsymbol{Q}_t := \boldsymbol{U}_t^\intercal\|\boldsymbol{x}_k\|^2\boldsymbol{\Lambda}\boldsymbol{U}_t$. This requires much less computation than calculating $\left[\boldsymbol{I}_M + \boldsymbol{V}_t\|\boldsymbol{x}_k\|^2\boldsymbol{\Lambda}\right]^{-1}\boldsymbol{V}_t$ directly. To get these low rank $\boldsymbol{V}_t$'s, you can use the `irlba` R package to efficiently calculate the top SV's without having to calculate all of the SV's, though you should only need to do this once.

In calculating $\boldsymbol{\delta}$, we note that diag$\left\{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu_B} - \boldsymbol{\mu_Z}\boldsymbol{A})^\intercal(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu_B} - \boldsymbol{\mu_Z}\boldsymbol{A})\right\}$ is equal to the vector that contains the columns sums of squares of $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\mu_B} - \boldsymbol{\mu_Z}\boldsymbol{A})$.

If calculating the ELBO is taking too long, then you don't really need to calculate it. Choose a different stopping criterion. Just use the ELBO as an implementation check on small datasets.

9 All genes at once

All of these derivations assume that we are only operating on one gene. However, we are going to assume that all genes share the same $\boldsymbol{\pi}$. Suppose we have P genes. Then the model becomes

$$
\boldsymbol{Y}_{p_{N_p\times M_p}} = \boldsymbol{X}_{p_{N_p\times K_p}}\boldsymbol{B}_{p_{K_p\times M_p}} + \boldsymbol{Z}_{p_{N_p\times R_p}}\boldsymbol{A}_{p_{R_p\times M_p}} + \boldsymbol{E}_{p_{N_p\times M_p}}, \tag{192}
$$

where

$$
\begin{aligned}
\boldsymbol{E}_p &\sim \underset{N_p\times M_p}{\mathrm{N}}(\boldsymbol{0}, \boldsymbol{\Lambda}_p^{-1}\otimes\boldsymbol{I}_{N_p}), && (193)\\
\boldsymbol{\Lambda}_p &= \mathrm{diag}(\lambda_{1p},\ldots,\lambda_{M_pp}), && (194)\\
\boldsymbol{Z}_p &\sim \underset{N_p\times R_p}{\mathrm{N}}(\boldsymbol{0}, \boldsymbol{I}_{R_p}\otimes\boldsymbol{I}_{N_p}). && (195)
\end{aligned}
$$

Again, we place iid mixtures of normals prior on the rows of $\boldsymbol{B}_p$.

$$
\begin{aligned}
\boldsymbol{B}_p &= \begin{pmatrix} \boldsymbol{b}_{1p}^\intercal \\ \vdots \\ \boldsymbol{b}_{K_p}^\intercal \end{pmatrix}, && (196)\\
\boldsymbol{b}_{kp} \text{ i.i.d. s.t. } p(\boldsymbol{b}_{kp}) &= \sum_{t=0}^{T}\pi_t\underset{M}{\mathrm{N}}(\boldsymbol{b}_{kp}|\boldsymbol{0},\boldsymbol{V}_t). && (197)
\end{aligned}
$$

Notice that we have the same π_t's and the same $\boldsymbol{V}_t$'s for all p. We use the same augmented parameterization as in Section 1. Let

$$
\begin{aligned}
\mathcal{Y} &= \{\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_p\}, && (198)\\
\mathcal{X} &= \{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_p\}, && (199)\\
\mathcal{B} &= \{\boldsymbol{B}_1,\ldots,\boldsymbol{B}_p\}, && (200)\\
\mathcal{W} &= \{\boldsymbol{W}_1,\ldots,\boldsymbol{W}_p\}, && (201)\\
\mathcal{Z} &= \{\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_p\}, && (202)\\
\mathscr{L} &= \{\boldsymbol{\Lambda}_1,\ldots,\boldsymbol{\Lambda}_p\}, \text{ and} && (203)
\end{aligned}
$$

$$\mathcal{A} = \{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_p\}. \tag{204}$$

We assume all genes are independent, hence we have that

$$p(\mathcal{Y}, \mathcal{B}, \mathcal{W}, \mathcal{Z} | \mathscr{L}, \mathcal{A}, \boldsymbol{\pi}) = \prod_{p=1}^{P} p(\boldsymbol{Y}_p, \boldsymbol{B}_p, \boldsymbol{W}_p, \boldsymbol{Z}_p | \boldsymbol{\Lambda}_p, \boldsymbol{\pi}, \boldsymbol{A}_p). \tag{205}$$

We perform variational Bayes and try to optimize the following lower bounds on the marginal log-likelihood over q, $\boldsymbol{\pi}$, $\mathcal{A}$, and $\mathscr{L}$:

$$\mathcal{L}(q, \boldsymbol{\pi}, \mathcal{A}, \mathscr{L}) = \int q(\mathcal{B}, \mathcal{W}, \mathcal{Z}) \log \left\{ \frac{p(\mathcal{Y}, \mathcal{B}, \mathcal{W}, \mathcal{Z} |, \boldsymbol{\pi}, \mathcal{A})}{q(\mathcal{B}, \mathcal{W}, \mathcal{Z})} \right\} \mathrm{d}\mathcal{B}\, \mathrm{d}\mathcal{W}\, \mathrm{d}\mathcal{Z}. \tag{206}$$

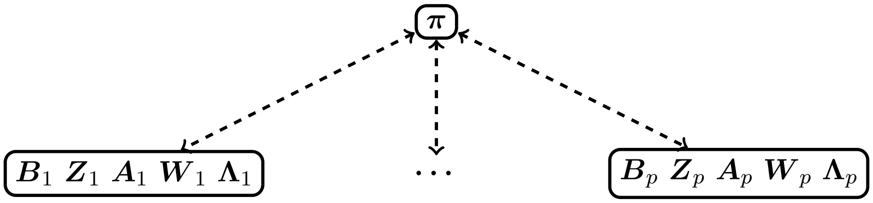
Here, we assume a separable q, thus performing mean-field variational inference:

$$q(\mathcal{B}, \mathcal{W}, \mathcal{Z}) = \prod_{p=1}^{P} \left(q(\boldsymbol{Z}_p) \prod_{k=1}^{K} q(\boldsymbol{b}_{kp}, \boldsymbol{w}_{kp}) \right). \tag{207}$$

It turns out that all of the updates for q, the $\boldsymbol{\Lambda}_p$'s, and the $\boldsymbol{A}_p$'s are the exact same as in the case of one gene. That is, one looks individually at each gene. The only change then is the update for $\boldsymbol{\pi}$. In this case, we use all of the current values of γ_{kpt} and the update becomes

$$\pi_t = \frac{\left(\sum_{p=1}^{P} \sum_{k=1}^{K} \gamma_{ktp} \right) + \eta_t - 1}{\sum_{t=0}^{T} \sum_{p=1}^{P} \sum_{k=1}^{K} \gamma_{ktp} + \sum_{t=0}^{T} \eta_t - T}. \tag{208}$$

To do this on real data, we would probably need to parallelize the updates for all parameters except $\boldsymbol{\pi}$ then update $\boldsymbol{\pi}$ then repeat.


$\boldsymbol{\pi}$
$\boldsymbol{B}_1\ \boldsymbol{Z}_1\ \boldsymbol{A}_1\ \boldsymbol{W}_1\ \boldsymbol{\Lambda}_1$
$\cdots$
$\boldsymbol{B}_p\ \boldsymbol{Z}_p\ \boldsymbol{A}_p\ \boldsymbol{W}_p\ \boldsymbol{\Lambda}_p$

References

Christopher M Bishop. Pattern recognition. *Machine Learning*, 128, 2006.