Probabilistic Low-Rank Subspace Clustering Supplementary Material

S. Derin Babacan
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
`dbabacan@gmail.com`

Shinichi Nakajima
Nikon Corporation
Tokyo, 140-8601, Japan
`nakajima.s@nikon.co.jp`

Minh N. Do
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
`minhdo@illinois.edu`

In this supplementary material, we provide the derivation of the global solution of the expectation-maximization method in Sec. 2.2 and the required statistics in the variational Bayesian methods in Secs. 3 and 4. Equation numbers are denoted with preceding "S-", and the ones without "S-" refer to the main text.

1 Global Solution of the EM method

The log-likelihood is given by

$$\begin{aligned}\mathcal{L} &= \sum_{i=1}^{N} \log \mathrm{p}(\mathbf{d}_i, \mathbf{y}_i|\mathbf{D}, \mathbf{A}) \qquad \text{(S-1)}\\ &= -\frac{N}{2}\left(M\log(\sigma_y^2) + \log|\mathbf{K}| - \frac{1}{N}\,\mathrm{tr}\left(\mathbf{K}^{-1}\mathbf{D}\mathbf{D}^T\right)\right) - \frac{1}{2\sigma_y^2}\,\mathrm{tr}\left((\mathbf{Y}-\mathbf{D})^T(\mathbf{Y}-\mathbf{D})\right) + \mathrm{const}\,,\end{aligned}$$

with $\mathbf{K} = \sigma_d^2\mathbf{I} + \mathbf{D}\mathbf{A}\mathbf{A}^T\mathbf{D}^T$. To maximize the log-likelihood w.r.t. $\mathbf{A}$, we take its gradient w.r.t. $\mathbf{A}$ using matrix differentiation identities [2] and set it equal to zero, which yields

$$\mathbf{D}^T\mathbf{K}^{-1}\mathbf{D}\mathbf{A} = \frac{1}{N}\mathbf{D}^T\mathbf{K}^{-1}\mathbf{D}\mathbf{D}^T\mathbf{K}^{-1}\mathbf{D}\mathbf{A}\,. \qquad \text{(S-2)}$$

This has three possible solutions: (i) $\mathbf{D}\mathbf{A} = \mathbf{0}$, (ii) $\mathbf{K} = \frac{1}{N}\mathbf{D}\mathbf{D}^T$, and (iii) $\mathbf{D}\mathbf{A} \neq \mathbf{0}$ and $\mathbf{K} \neq \frac{1}{N}\mathbf{D}\mathbf{D}^T$. We consider the latter two cases, as the first one is not interesting for subspace clustering. In the last case, assuming $\sigma_d^2 > 0$ and thus $\mathbf{K}^{-1}$ exists, we have

$$\mathbf{D}\mathbf{A} = \frac{1}{N}\mathbf{D}\mathbf{D}^T\mathbf{K}^{-1}\mathbf{D}\mathbf{A}\,. \qquad \text{(S-3)}$$

We first solve this system w.r.t. $\mathbf{D}\mathbf{A}$. Let the SVDs of $\mathbf{D}$ and $\mathbf{D}\mathbf{A}$ be[1] $\mathbf{D} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ and $\mathbf{D}\mathbf{A} = \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{V}}^T$, respectively, such that we have

$$\begin{aligned}\mathbf{K}^{-1}\mathbf{D}\mathbf{A} &= \left(\sigma_d^2\mathbf{I} + \mathbf{D}\mathbf{A}\mathbf{A}^T\mathbf{D}^T\right)^{-1}\mathbf{D}\mathbf{A}\,, && \text{(S-4)}\\ &= \mathbf{D}\mathbf{A}\left(\sigma_d^2\mathbf{I} + \mathbf{A}^T\mathbf{D}^T\mathbf{D}\mathbf{A}\right)^{-1}\,, && \text{(S-5)}\\ &= \hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\left(\sigma_d^2\mathbf{I} + \hat{\mathbf{\Lambda}}^2\right)^{-1}\hat{\mathbf{V}}^T\,. && \text{(S-6)}\end{aligned}$$

---

[1] At this point, we do not know if the singular vectors of $\mathbf{D}\mathbf{A}$ and $\mathbf{D}$ are related.

Plugging this in (S-2), we have at the stationary points

$$\hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\hat{\mathbf{V}}^T = \frac{1}{N}\mathbf{D}\mathbf{D}^T\hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\left(\sigma_d^2\mathbf{I}+\hat{\mathbf{\Lambda}}^2\right)^{-1}\hat{\mathbf{V}}^T\,, \tag{S-7}$$

$$\hat{\mathbf{U}}\left(\sigma_d^2\mathbf{I}+\hat{\mathbf{\Lambda}}^2\right)\hat{\mathbf{\Lambda}} = \frac{1}{N}\mathbf{D}\mathbf{D}^T\hat{\mathbf{U}}\hat{\mathbf{\Lambda}}\,, \tag{S-8}$$

from which it can be observed that $\hat{\mathbf{U}}$ contains the eigenvectors of $\mathbf{D}\mathbf{D}^T$ and hence the left singular vectors of $\mathbf{D}$, such that $\hat{\mathbf{U}} = \mathbf{U}$. Moreover, $\sigma_d^2\mathbf{I}+\hat{\mathbf{\Lambda}}^2$ contains the eigenvalues of $\frac{1}{N}\mathbf{D}\mathbf{D}^T$. Therefore, similarly to [3], we have the solution

$$\mathbf{D}\mathbf{A} = \mathbf{U}_q\left(\frac{1}{N}\mathbf{\Lambda}_q^2 - \sigma_d^2\mathbf{I}\right)^{1/2}\mathbf{R}\,, \tag{S-9}$$

where $\mathbf{R}$ is an arbitrary orthogonal rotation matrix, and $\mathbf{U}_q$ is a $M \times q$ matrix consisting of q left singular vectors of $\mathbf{D}$ with corresponding singular values that are larger than $\sqrt{N}\sigma_d$. Therefore, the singular values of $\mathbf{D}\mathbf{A}$ satisfy $l_i = (\frac{\lambda_i^2}{N} - \sigma_d^2)^{1/2}$.

In the case (ii), we have the same solution (S-9) where the last $M-q$ smallest singular values of $\mathbf{D}$ are equal to $\sqrt{N}\sigma_d$. This is an unrealistic case and is analyzed also in PPCA [3].

Using the solution (S-9), we can solve for the optimal $\mathbf{B}$ using (9) as

$$\langle\mathbf{B}\rangle = \mathbf{\Sigma}_{\mathbf{B}}\,\frac{1}{\sigma_d^2}\,\mathbf{A}^T\mathbf{D}^T\mathbf{D}\,, \tag{S-10}$$

$$= \left(\sigma_d^2\mathbf{I} + \mathbf{R}^T(\frac{1}{N}\mathbf{\Lambda}_q^2 - \sigma_d^2\mathbf{I})\mathbf{R}\right)^{-1}\mathbf{R}^T(\frac{1}{N}\mathbf{\Lambda}_q^2 - \sigma_d^2\mathbf{I})^{1/2}\mathbf{U}_q^T\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T\,, \tag{S-11}$$

$$= \mathbf{R}^T\sigma_d^{-2}(\frac{1}{N}\mathbf{\Lambda}_q^2 - \sigma_d^2\mathbf{I})^{1/2}\left(\mathbf{I} + \sigma_d^{-2}(\frac{1}{N}\mathbf{\Lambda}_q^2 - \sigma_d^2\mathbf{I})\mathbf{R}\mathbf{R}^T\right)^{-1}\mathbf{\Lambda}_q\mathbf{V}_q^T\,, \tag{S-12}$$

$$= \mathbf{R}^T(\frac{1}{N}\mathbf{\Lambda}_q^2 - \sigma_d^2\mathbf{I})^{1/2}\mathbf{\Lambda}_q^{-1}N\mathbf{V}_q^T\,. \tag{S-13}$$

Now we have an expression for $\mathbf{D}\mathbf{A}$ and $\langle\mathbf{B}\rangle$. Combining,

$$\mathbf{D}\mathbf{A}\langle\mathbf{B}\rangle = \mathbf{U}_q(\mathbf{\Lambda}_q^2 - N\sigma_d^2\mathbf{I})\mathbf{\Lambda}_q^{-1}\mathbf{V}_q^T\,. \tag{S-14}$$

Plugging $\mathbf{D} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ in (S-14) yields the final solution

$$\mathbf{A}\langle\mathbf{B}\rangle = \mathbf{V}_q(\mathbf{\Lambda}_q^2 - N\sigma_d^2\mathbf{I})\mathbf{\Lambda}_q^{-2}\mathbf{V}_q^T = \mathbf{V}_q\tilde{\mathbf{\Lambda}}_q\mathbf{V}_q^T\,, \tag{S-15}$$

with $\tilde{\mathbf{\Lambda}}_q$ is a diagonal matrix with $1 - \frac{N\sigma_d^2}{\lambda_j^2}$ on the diagonal. The optimal solution for $\mathbf{A}$ can easily be extracted from this expression.

Finally, using this expression for $\mathbf{A}\langle\mathbf{B}\rangle$ in (10), we solve for $\mathbf{D}$ as

$$\mathbf{Y} = \mathbf{D}\left[\mathbf{I} + \frac{\sigma_y^2}{\sigma_d^2}\langle\,(\mathbf{I}-\mathbf{A}\mathbf{B})\,(\mathbf{I}-\mathbf{A}\mathbf{B})^T\,\rangle\right]\,, \tag{S-16}$$

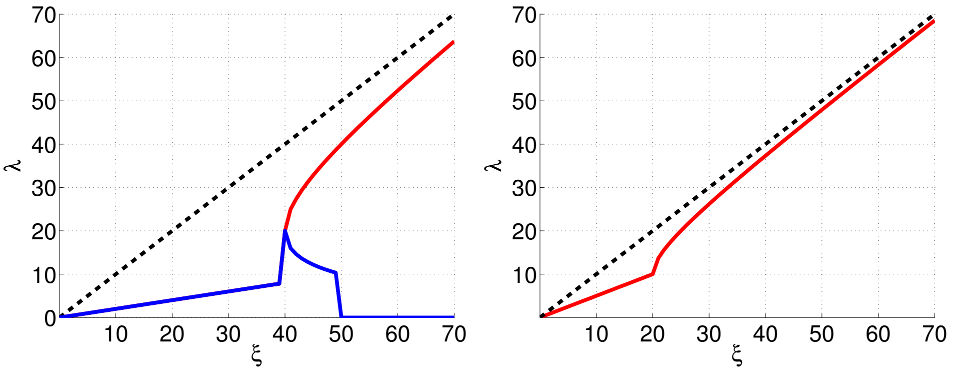
$$= \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T\left[\mathbf{I} + N\sigma_y^2\mathbf{V}_q\mathbf{\Lambda}_q^{-2}\mathbf{V}_q^T\right]\,, \tag{S-17}$$

Using the partitioning $\mathbf{D} = [\mathbf{U}_q\,,\,\mathbf{U}_{N-q}]\,\mathrm{diag}(\mathbf{\Lambda}_q, \mathbf{\Lambda}_{N-q})\,[\mathbf{V}_q\,,\,\mathbf{V}_{N-q}]^T$, we have the final solution

$$\mathbf{Y} = [\mathbf{U}_q, \mathbf{U}_{N-q}]\left[\begin{array}{cc} \mathbf{\Lambda}_q + N\sigma_y^2\mathbf{\Lambda}_q^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{\sigma_y^2+\sigma_d^2}{\sigma_d^2}\mathbf{\Lambda}_{N-q} \end{array}\right][\mathbf{V}_q\mathbf{V}_{N-q}]^T\,. \tag{S-18}$$

Therefore, the eigenvectors of $\mathbf{D}$ and $\mathbf{Y}$ are the same, but the eigenvalues are related via

$$\xi_j = \begin{cases} \lambda_j + N\sigma_y^2\;\lambda_j^{-1}\,, & \text{if} \quad \lambda_j > \sqrt{N}\sigma_d \\ \lambda_j\frac{\sigma_y^2+\sigma_d^2}{\sigma_d^2}\,, & \text{if} \quad \lambda_j \leq \sqrt{N}\sigma_d \end{cases} \tag{S-19}$$



(a) (b)

Figure 1: Estimates of singular values λ of $\mathbf{D}$ given singular values ξ of $\mathbf{Y}$ (N = 100). The dashed line is $\lambda = \xi$. In (a), $\sigma_d = 1, \sigma_y = 2$, in (b), $\sigma_d = \sigma_y = 1$.

The explicit solutions for λ_j are given by

$$\lambda_j = \begin{cases} \xi_j \frac{\sigma_d^2}{\sigma_y^2 + \sigma_d^2}\,, & \xi_j < 2\sqrt{N}\sigma_y \\ \frac{\xi_j}{2} + \frac{1}{2}\sqrt{\xi_j^2 - 4N\sigma_y^2} & \xi_j \geq 2\sqrt{N}\sigma_y,\, \xi_j \geq \min\left(2\sqrt{N}\sigma_d,\, \frac{\sqrt{N}}{\sigma_d}(\sigma_d^2 + \sigma_y^2)\right) \\ \frac{\xi_j}{2} - \frac{1}{2}\sqrt{\xi_j^2 - 4N\sigma_y^2} & \sigma_y \geq \sigma_d,\, 2\sqrt{N}\sigma_y \leq \xi_j \leq \frac{\sqrt{N}}{\sigma_d}(\sigma_d^2 + \sigma_y^2) \end{cases} \tag{S-20}$$

The solution for λ_j is unique except when $\sigma_y \geq \sigma_d$ and $2\sqrt{N}\sigma_y \leq \xi_j \leq \frac{\sqrt{N}}{\sigma_d}(\sigma_d^2 + \sigma_y^2)$, where we have the latter two cases as solutions. As shown in Fig. 1(a), the last solution is only valid in a comparably small region. To achieve continuity in the solutions, we always choose the first two solutions (S-20).

As can be observed from Fig. 1, the solution (S-20) is a combination of two operations: a down-scaling when $\xi_j < 2\sqrt{N}\sigma_y$ and a polynomial thresholding operation for larger singular values. The polynomial thresholding preserves the larger singular values as the shrinkage amount gets smaller: ξ_j gets larger compared to $2N\sigma_y$, and for very large values $\lambda_j \approx \xi_j$. On the other hand, small singular values get shrunk via down-scaling. Obviously, when $\sigma_d = 0$, no shrinkage is applied and $\mathbf{D} = \mathbf{Y}$.

2 Derivation of the Variational Bayesian Methods

The explicit form of the variational free energy in (17) is given by

$$\begin{aligned} \mathcal{F} &= \langle \log \mathrm{q}(\mathbf{D}, \mathbf{A}, \mathbf{B}, \sigma_d^2, \sigma_y^2) - \log \mathrm{p}(\mathbf{D}, \mathbf{A}, \mathbf{B}, \sigma_d^2, \sigma_y^2) \rangle_{q(\mathbf{D}, \mathbf{A}, \mathbf{B}, \sigma_d^2, \sigma_y^2)} \\ &= \langle \log \mathrm{q}(\mathbf{D})\, \mathrm{q}(\mathbf{A})\, \mathrm{q}(\mathbf{B})\, \mathrm{q}(\sigma_d^2)\, \mathrm{q}(\sigma_y^2) \rangle \\ &+ \frac{MN}{2} \langle \log \sigma_d^2 \rangle + \frac{MN}{2} \langle \log \sigma_y^2 \rangle + \frac{1}{2} \mathrm{tr}(\langle \mathbf{A} \mathbf{C}_{\mathbf{A}}^{-1} \mathbf{A}^T \rangle) + \frac{1}{2} \mathrm{tr}(\langle \mathbf{C}_{\mathbf{B}}^{-1} \mathbf{B} \mathbf{B}^T \rangle) + \frac{1}{2} \mathrm{tr}(\langle \mathbf{D} \mathbf{D}^T \rangle) \\ &+ \left(\frac{1}{2\langle \sigma_y^2 \rangle} + \frac{1}{2\langle \sigma_d^2 \rangle} \right) \mathrm{tr}(\langle \mathbf{D} \mathbf{D}^T \rangle) + \frac{1}{2\langle \sigma_y^2 \rangle} \|\mathbf{Y}\|_{\mathrm{F}}^2 - \frac{1}{\langle \sigma_y^2 \rangle} \mathrm{tr}(\langle \mathbf{D} \rangle^T \mathbf{Y}) \\ &- \frac{1}{\langle \sigma_d^2 \rangle} \mathrm{tr}(\langle \mathbf{B}^T \mathbf{A}^T \mathbf{D}^T \mathbf{D} \rangle) + \frac{1}{2\langle \sigma_d^2 \rangle} \mathrm{tr}(\langle \mathbf{A}^T \mathbf{D}^T \mathbf{D} \mathbf{A} \mathbf{B} \mathbf{B}^T \rangle) \\ &+ \frac{N}{2} \log |\mathbf{C}_{\mathbf{A}}| + \frac{N}{2} \log |\mathbf{C}_{\mathbf{B}}| + \mathrm{const}\,. \end{aligned} \tag{S-21}$$

The optimal forms of $\mathrm{q}(\mathbf{D})$ and $\mathrm{q}(\mathbf{B})$ can be found as matrix-variate normal distributions by inspection. The optimal $\mathrm{q}(\mathbf{A})$ does not have a matrix-variate normal form. The optimal distribution is found in terms of $\mathbf{a} = \mathrm{vec}(\mathbf{A})$, by rewriting the terms involving $\mathbf{A}$ in (S-23) as

$$\begin{aligned}
-\log \mathrm{q}(\mathbf{a}) &= \mathrm{tr}\left(\langle\sigma_d^{-2}\rangle \, \langle\|\mathbf{D}-\mathbf{DAB}\|_F\rangle^2 + \mathbf{AC}_\mathbf{A}^{-1}\mathbf{A}^T\right) + \frac{N}{2}\log|\mathbf{C}_\mathbf{A}| + \mathrm{const} \\
&= \langle\sigma_d^{-2}\rangle \, \langle\|\mathbf{d}-(\mathbf{B}^T\otimes\mathbf{D})\mathbf{a}\|_2^2\rangle + \mathbf{a}^T(\mathbf{C}_\mathbf{A}^{-1}\otimes\mathbf{I})\mathbf{a} + \frac{N}{2}\log|\mathbf{C}_\mathbf{A}| + \mathrm{const} \\
&= \langle\sigma_d^{-2}\rangle \, \langle \left(\mathbf{d}^T\mathbf{d} + \mathbf{a}^T(\mathbf{B}^T\otimes\mathbf{D})^T(\mathbf{B}^T\otimes\mathbf{D})\mathbf{a} - 2\mathbf{a}^T(\mathbf{B}^T\otimes\mathbf{D})^T\mathbf{d}\right) \rangle + \mathbf{a}^T(\mathbf{C}_\mathbf{A}^{-1}\otimes\mathbf{I})\mathbf{a} + \frac{N}{2}\log|\mathbf{C}_\mathbf{A}| + \mathrm{const} \\
&= \mathbf{a}^T\left[\langle\sigma_d^{-2}\rangle\langle(\mathbf{B}^T\otimes\mathbf{D})^T(\mathbf{B}^T\otimes\mathbf{D})\rangle + \mathbf{C}_\mathbf{A}^{-1}\otimes\mathbf{I}\right]\mathbf{a} - 2\mathbf{a}^T(\langle\mathbf{B}\rangle^T\otimes\langle\mathbf{D}\rangle)^T\langle\mathbf{d}\rangle + \frac{N}{2}\log|\mathbf{C}_\mathbf{A}| + \mathrm{const} \\
&= \mathbf{a}^T\left[\langle\sigma_d^{-2}\rangle(\langle\mathbf{B}^T\mathbf{B}\rangle\otimes\langle\mathbf{D}^T\mathbf{D}\rangle) + \mathbf{C}_\mathbf{A}^{-1}\otimes\mathbf{I}\right]\mathbf{a} - 2\mathbf{a}^T(\langle\mathbf{B}\rangle^T\otimes\langle\mathbf{D}\rangle)^T\langle\mathbf{d}\rangle + \frac{N}{2}\log|\mathbf{C}_\mathbf{A}| + \mathrm{const}
\end{aligned} \tag{S-22}$$

where we used $\mathrm{vec}(\mathbf{DAB}) = (\mathbf{B}^T\otimes\mathbf{D})\,\mathrm{vec}(\mathbf{A})$, and $\mathbf{d} = \mathrm{vec}(\mathbf{D})$, $\mathbf{b} = \mathrm{vec}(\mathbf{B})$. It can be derived from here that $\mathrm{q}(\mathbf{a})$ has a multivariate normal distribution with mean $\mathbf{\Sigma}_\mathbf{a}\,(\langle\mathbf{B}\rangle^T\otimes\langle\mathbf{D}\rangle)^T\langle\mathbf{d}\rangle$ and covariance $\mathbf{\Sigma}_\mathbf{a} = \left[\langle\sigma_d^{-2}\rangle(\langle\mathbf{B}^T\mathbf{B}\rangle\otimes\langle\mathbf{D}^T\mathbf{D}\rangle) + \mathbf{C}_\mathbf{A}^{-1}\otimes\mathbf{I}\right]^{-1}$. However, computing $\mathbf{A}$ in this manner can be very inefficient, as $\mathbf{\Sigma}_\mathbf{A}$ might get extremely big ($MN\times MN$ for $\mathbf{A}$ of size $N\times N$ and $\mathbf{D}$ of size $M\times N$).

Therefore, we force $\mathrm{q}(\mathbf{A})$ to have a matrix-variate form $\mathcal{N}(\langle\mathbf{A}\rangle, \mathbf{\Sigma}_\mathbf{A}, \mathbf{\Omega}_\mathbf{A})$, which leads to an efficient algorithm. Under this constraint, the variational free energy can be rewritten as (treating all terms not involving $\mathbf{A}$ as constant)

$$\begin{aligned}
\mathcal{F} = \frac{1}{2}\,\mathrm{tr}(\langle\mathbf{AC}_\mathbf{A}^{-1}\mathbf{A}^T\rangle) - \frac{1}{\langle\sigma_d^2\rangle}\,\mathrm{tr}(\langle\mathbf{B}^T\mathbf{A}^T\mathbf{D}^T\mathbf{D}\rangle) + \frac{1}{2\langle\sigma_d^2\rangle}\,\mathrm{tr}(\langle\mathbf{A}^T\mathbf{D}^T\mathbf{DABB}^T\rangle) \\
- \frac{N}{2}\log|\mathbf{\Sigma}_\mathbf{A}| - \frac{N}{2}\log|\mathbf{\Omega}_\mathbf{A}| + \frac{N}{2}\log|\mathbf{C}_\mathbf{A}| + \frac{N}{2}\log|\mathbf{C}_\mathbf{B}| + \mathrm{const}\,.
\end{aligned} \tag{S-23}$$

Evaluating the expectations using the matrix-variate normal form for $\mathrm{q}(\mathbf{A})$ (see the next section), we minimize $\mathcal{F}$ with respect to $\mathbf{\Sigma}_\mathbf{A}$, resulting in

$$\mathbf{\Sigma}_\mathbf{A}^{-1} = \frac{1}{N}\,\mathrm{tr}(\mathbf{C}_\mathbf{A}^{-1}\mathbf{\Omega}_\mathbf{A})\,\mathbf{I} + \frac{1}{N\sigma_d^2}\,\mathrm{tr}(\mathbf{\Omega}_\mathbf{A}\langle\mathbf{BB}^T\rangle)\,\langle\mathbf{D}^T\mathbf{D}\rangle \tag{S-24}$$

Similarly, minimization with respect to $\mathbf{\Omega}_\mathbf{A}$ yields

$$\mathbf{\Omega}_\mathbf{A}^{-1} = \frac{1}{N}\,\mathrm{tr}(\mathbf{\Sigma}_\mathbf{A})\mathbf{C}_\mathbf{A}^{-1} + \frac{1}{N\sigma_d^2}\,\mathrm{tr}(\mathbf{\Sigma}_\mathbf{A}\langle\mathbf{D}^T\mathbf{D}\rangle)\,\langle\mathbf{BB}^T\rangle\,. \tag{S-25}$$

Finally, the update of $\langle\mathbf{A}\rangle$ is given by

$$\langle\mathbf{A}\rangle\mathbf{C}_\mathbf{A}^{-1} + \frac{1}{\sigma_d^2}\langle\mathbf{D}^T\mathbf{D}\rangle\langle\mathbf{A}\rangle\langle\mathbf{BB}^T\rangle = \frac{1}{\sigma_d^2}\langle\mathbf{D}^T\mathbf{D}\rangle\langle\mathbf{B}\rangle^T \tag{S-26}$$

The closed form solution for $\langle\mathbf{A}\rangle$ cannot be found, but it can be solved using a fixed-point iteration starting from an initial estimate.

2.1 Required Statistics for the Variational Bayesian Methods

For a general matrix-variate Gaussian distribution $\mathrm{p}(\mathbf{X}|\mathbf{M},\mathbf{\Omega},\mathbf{\Sigma}) = \mathcal{N}(\mathbf{X}|\mathbf{M},\mathbf{\Sigma},\mathbf{\Omega})$, we have [1]

$$\langle\mathbf{X}^T\mathbf{KX}\rangle = \mathrm{tr}(\mathbf{\Sigma K}^T)\mathbf{\Omega} + \mathbf{M}^T\mathbf{KM}\,, \tag{S-27}$$

$$\langle\mathbf{XKX}^T\rangle = \mathrm{tr}(\mathbf{K}^T\mathbf{\Omega})\mathbf{\Sigma} + \mathbf{MKM}^T\,. \tag{S-28}$$

Thus, for $\mathrm{q}(\mathbf{D}) = \mathcal{N}(\langle\mathbf{D}\rangle, \mathbf{I}, \mathbf{\Omega_D})$, $\mathrm{q}(\mathbf{A}) = \mathcal{N}(\langle\mathbf{A}\rangle, \mathbf{\Sigma_A}, \mathbf{\Omega_A})$, and $\mathrm{q}(\mathbf{B}) = \mathcal{N}(\langle\mathbf{B}\rangle, \mathbf{I}, \mathbf{\Sigma_B})$, we have

$$\begin{aligned}
\langle\mathbf{D}^T\mathbf{D}\rangle &= \mathrm{tr}(\mathbf{I}_M)\mathbf{\Omega_D} + \langle\mathbf{D}\rangle^T\langle\mathbf{D}\rangle && \text{(S-29)}\\
&= M\mathbf{\Omega_D} + \langle\mathbf{D}\rangle^T\langle\mathbf{D}\rangle && \text{(S-30)}\\
\langle\mathbf{A}\mathbf{A}^T\rangle &= \mathrm{tr}(\mathbf{\Omega_A})\mathbf{\Sigma_A} + \langle\mathbf{A}\rangle\langle\mathbf{A}\rangle^T && \text{(S-31)}\\
\langle\mathbf{A}^T\mathbf{A}\rangle &= \mathrm{tr}(\mathbf{\Sigma_A})\mathbf{\Omega_A} + \langle\mathbf{A}\rangle^T\langle\mathbf{A}\rangle && \text{(S-32)}\\
\langle\mathbf{B}\mathbf{B}^T\rangle &= \mathrm{tr}(\mathbf{\Omega_B})\mathbf{\Sigma_B} + \langle\mathbf{B}\rangle\langle\mathbf{B}\rangle^T && \text{(S-33)}\\
&= N\mathbf{\Sigma_B} + \langle\mathbf{B}\rangle\langle\mathbf{B}\rangle^T && \text{(S-34)}\\
\langle\mathbf{B}^T\mathbf{B}\rangle &= \mathrm{tr}(\mathbf{\Sigma_B})\mathbf{I}_N + \langle\mathbf{B}\rangle^T\langle\mathbf{B}\rangle && \text{(S-35)}
\end{aligned}$$

Combining, we obtain

$$\begin{aligned}
\langle\mathbf{A}^T\mathbf{D}^T\mathbf{D}\mathbf{A}\rangle &= \mathrm{tr}(\mathbf{\Sigma_A}\langle\mathbf{D}^T\mathbf{D}\rangle)\mathbf{\Omega_A} + \langle\mathbf{A}\rangle^T\langle\mathbf{D}^T\mathbf{D}\rangle\langle\mathbf{A}\rangle && \text{(S-36)}\\
\langle\mathbf{B}^T\mathbf{A}^T\mathbf{A}\mathbf{B}\rangle &= \mathrm{tr}(\mathbf{\Sigma_B}\langle\mathbf{A}^T\mathbf{A}\rangle)\mathbf{I}_N + \langle\mathbf{B}\rangle^T\langle\mathbf{A}^T\mathbf{A}\rangle\langle\mathbf{B}\rangle && \text{(S-37)}\\
\langle\mathbf{A}\mathbf{B}\mathbf{B}^T\mathbf{A}^T\rangle &= \mathrm{tr}(\langle\mathbf{B}\mathbf{B}^T\rangle\mathbf{\Omega_A})\mathbf{\Sigma_A} + \langle\mathbf{A}\rangle\langle\mathbf{B}\mathbf{B}^T\rangle\langle\mathbf{A}\rangle^T && \text{(S-38)}\\
\langle\mathbf{B}^T\mathbf{A}^T\mathbf{D}^T\mathbf{D}\mathbf{A}\mathbf{B}\rangle &= \mathrm{tr}(\mathbf{\Sigma_B}\langle\mathbf{A}^T\mathbf{D}^T\mathbf{D}\mathbf{A}\rangle)\mathbf{I}_N + \mathbf{B}^T\langle\mathbf{A}^T\mathbf{D}^T\mathbf{D}\mathbf{A}\rangle\mathbf{B} && \text{(S-39)}
\end{aligned}$$

References

[1] A. K. Gupta and D. K. Nagar. *Matrix Variate Distributions*. Chapman & Hall/CRC, New York, 2000.

[2] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate analysis*. Academic Press; New York, 1979.

[3] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Comput.*, 11(2):443–482, February 1999.