Supplemental Material for "Integrating Features and Similarities: Flexible Models for Heterogeneous Multiview Data"

Variational Objective

Define $\boldsymbol{\Theta} = \{\boldsymbol{U}^{(m)}, \boldsymbol{W}^{(m)}, \boldsymbol{\alpha}_m, \gamma_m\}_{m=1}^{M}$. For the multiview classification problem, $\{\boldsymbol{\beta}_c, \{z_{nc}\}_{n=1}^{N}\}_{c=1}^{C}$ also need to be estimated. Besides, for the ordinal views, we denote the cutpoints as $\boldsymbol{G} = \{\boldsymbol{g}^m\}_{m=1}^{M_1}$, and the rotation matrix as $\boldsymbol{Q}$. Throughout the derivation, we discuss the model for the multiclass classification problem (where we are also given the labels which we denote by y).

The goal is to minimize KL divergence $KL(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta}|\boldsymbol{X}, y, \boldsymbol{G}, \boldsymbol{Q}))$, where $q(\boldsymbol{\Theta})$ is a mean field approximation of $p(\boldsymbol{\Theta}|\boldsymbol{X}, y, \boldsymbol{G}, \boldsymbol{Q})$. This is equivalent to maximizing the evidence lower bound (ELBO) $\mathcal{L}(q(\boldsymbol{\Theta}), \boldsymbol{G}, \boldsymbol{Q})$:

$$\begin{aligned}
\mathcal{L}(q(\boldsymbol{\Theta}), \boldsymbol{G}, \boldsymbol{Q}) &= \langle \log p(\boldsymbol{X}, y, \boldsymbol{\Theta}) - \log q(\boldsymbol{\Theta})\rangle_{q(\boldsymbol{\Theta})} \qquad (1)\\
&= \langle \log p(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{V}, \gamma)\rangle + \langle \log \frac{p(\boldsymbol{W}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})p(\boldsymbol{V})p(\gamma)p(\boldsymbol{\beta})}{q(\boldsymbol{W})q(\boldsymbol{\alpha})q(\boldsymbol{V})q(\gamma)q(\boldsymbol{\beta})}\rangle + \langle \log p(y|\boldsymbol{V}, \boldsymbol{\beta})\rangle\\
&= \sum_{m=1}^{M_1} \langle \log p(\boldsymbol{X}^{(m)}|\boldsymbol{V}, \gamma_m)\rangle\\
&\quad + \sum_{m=M_1+1}^{M_1+M_2} \langle \log p(\boldsymbol{X}^{(m)}|\boldsymbol{V}, \gamma_m)\rangle + \sum_{m=M_1+M_2+1}^{M_1+M_2+M_3} \langle \log p(\boldsymbol{X}^{(m)}|\boldsymbol{U}^{(m)}, \tau_m)\rangle\\
&\quad + \sum_{m=M_1+M_2+1}^{M_1+M_2+M_3} \langle \log p(\boldsymbol{U}^{(m)}|\boldsymbol{V}, \gamma_m)\rangle + \sum_{m=1}^{M_1+M_2+M_3} \langle \log p(\boldsymbol{W}^{(m)}|\boldsymbol{\alpha}_m)\rangle + \langle \log p(\boldsymbol{\alpha}_m)\rangle\\
&\quad + \sum_{m=1}^{M_1+M_2+M_3} \langle \log p(\gamma_m)\rangle + \langle \log p(\boldsymbol{V})\rangle + \sum_{c=1}^{C} \langle \log p(\boldsymbol{\beta}_c)\rangle + \langle \log p(\boldsymbol{z}|\boldsymbol{V}, \boldsymbol{\beta})\rangle + \langle \log p(y|\boldsymbol{z})\rangle\\
&\quad - \sum_{m=1}^{M_1+M_2+M3} \langle \log q(\boldsymbol{W}^{(m)})\rangle + \langle \log q(\boldsymbol{\alpha}_m)\rangle - \sum_{m=1}^{M_1+M_2+M3} \langle \log q(\gamma_m)\rangle\\
&\quad - \langle \log q(\boldsymbol{V})\rangle - \sum_{c=1}^{C} \langle \log q(\boldsymbol{\beta}_c)\rangle - \langle \log q(\boldsymbol{z})\rangle - \sum_{m=M_1+M_2+1}^{M_1+M_2+M_3} \langle \log q(\boldsymbol{U}^{(m)})\rangle
\end{aligned}$$

Approximation for ordinal views

Directly maximizing $\mathcal{L}(q(\boldsymbol{\Theta}), \boldsymbol{G}, \boldsymbol{Q})$ is intractable, thus further approximation is needed for the first term of (1). Only the ordinal views are considered in this subsection. For Gaussian (real-valued) and similarity-based views, no such approximation is needed. The approximation for the ordinal views proceeds as follows:

$$\begin{aligned}
\langle \log p(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{V}, \gamma)\rangle_{q(\boldsymbol{\Theta})} &= \sum_{i,j,m} \langle \log \int p(\boldsymbol{X}_{ij}^{(m)}|\boldsymbol{U}_{ij}^{(m)})p(\boldsymbol{U}_{ij}^{(m)}|\boldsymbol{V}_i, \boldsymbol{W}_{:j}^{(m)}, \gamma_m) d\boldsymbol{U}_{ij}^{(m)}\rangle \qquad (2)\\
&= \sum_{i,j,m} \langle \log \int_{\boldsymbol{g}^m_{\boldsymbol{X}_{ij}^{(m)}-1}}^{\boldsymbol{g}^m_{\boldsymbol{X}_{ij}^{(m)}}} \mathcal{N}(\boldsymbol{U}_{ij}^{(m)}; \boldsymbol{V}_i\boldsymbol{W}_{:j}^{(m)}, \gamma_m^{-1}) d\boldsymbol{U}_{ij}^{(m)}\rangle\\
&= \sum_{i,j,m} \langle \log[\Phi(\beta_{i,j,m}) - \Phi(\alpha_{i,j,m})]\rangle\\
&= \text{const} + \sum_{i,j,m} \langle \log \int_{\alpha_{i,j,m}}^{\beta_{i,j,m}} \exp(-\frac{u^2}{2}) du\rangle\\
&\geq \sum_{i,j,m} \frac{1}{\langle \beta_{i,j,m} - \alpha_{i,j,m}\rangle} \langle \int_{\alpha_{i,j,m}}^{\beta_{i,j,m}} \log(\beta_{i,j,m} - \alpha_{i,j,m}) - \frac{1}{2}u^2 du\rangle + \text{const}
\end{aligned}$$

$$\begin{aligned} = & \sum_{i,j,m} \log(\boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}} - \boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}-1}) + \frac{1}{2}\langle \log \gamma_m \rangle - \frac{1}{2}\langle \gamma_m \rangle \langle (\boldsymbol{V}_i \boldsymbol{W}^{(m)}_{:j})^2 \rangle \\ & + \frac{1}{2}\langle \gamma_m \rangle \langle \boldsymbol{V}_i \boldsymbol{W}^{(m)}_{:j} \rangle (\boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}} + \boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}-1}) \\ & - \frac{1}{6}\langle \gamma_m \rangle ((\boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}})^2 + (\boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}-1})^2 + \boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}} \boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}-1}) + \text{const} \end{aligned} \tag{3}$$

In the above, $\beta_{i,j,m} = (\boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}} - \boldsymbol{V}_i \boldsymbol{W}^{(m)}_{:j})\gamma_m^{-\frac{1}{2}}$, $\alpha_{i,j,m} = (\boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}-1} - \boldsymbol{V}_i \boldsymbol{W}^{(m)}_{:j})\gamma_m^{-\frac{1}{2}}$, and $\Phi(.)$ is c.d.f. of the normal distribution. (3) is obtained using Jensen's inequality, but it can be also derived from Taylor's expansion, showing the conditions of the bound's tightness. As below, (3) can be equivalently expressed using the erf fucntion.

$$\begin{aligned} & \sum_{i,j,m} \langle \log[\Phi(\beta_{i,j,m}) - \Phi(\alpha_{i,j,m})] \rangle \\ = & \sum_{i,j,t} \langle \log[\text{erf}(\frac{\beta_{i,j,m}}{\sqrt{2}}) - \text{erf}(\frac{\alpha_{i,j,m}}{\sqrt{2}})] \rangle + \text{const} \\ \approx & \sum_{i,j,m} \langle \log \frac{2}{\sqrt{\pi}} [\frac{\beta_{i,j,m}}{\sqrt{2}} - \frac{1}{3}(\frac{\beta_{i,j,m}}{\sqrt{2}})^3 - \frac{\alpha_{i,j,m}}{\sqrt{2}} + \frac{1}{3}(\frac{\alpha_{i,j,m}}{\sqrt{2}})^3] \rangle + \text{const} \end{aligned} \tag{4}$$

$$= \sum_{i,j,m} \langle \log(\beta_{i,j,m} - \alpha_{i,j,m}) \rangle + \langle \log[1 - \frac{1}{6}(\beta^2_{i,j,m} + \alpha^2_{i,j,m} + \beta_{i,j,m}\alpha_{i,j,m})] \rangle + \text{const}$$

$$\approx \sum_{i,j,m} \langle \log(\beta_{i,j,m} - \alpha_{i,j,m}) \rangle - \frac{1}{6}\langle (\beta^2_{i,j,m} + \alpha^2_{i,j,m} + \beta_{i,j,m}\alpha_{i,j,m}) \rangle + \text{const} \tag{5}$$

$$\begin{aligned} = & \sum_{i,j,m} \log(\boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}} - \boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}-1}) + \frac{1}{2}\langle \log \gamma_m \rangle - \frac{1}{2}\langle \gamma_m \rangle \langle (\boldsymbol{V}_i \boldsymbol{W}^{(m)}_{:j})^2 \rangle \\ & + \frac{1}{2}\langle \gamma_m \rangle \langle \boldsymbol{V}_i \boldsymbol{W}^{t}_{:j} \rangle (\boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}} + \boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}-1}) \\ & - \frac{1}{6}\langle \gamma_m \rangle ((\boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}})^2 + (\boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}-1})^2 + \boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}} \boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}-1}) + \text{const} \end{aligned} \tag{6}$$

In the above derivation, an approximation using Taylor's expansion is used. Ignoring the higher order terms of $\mathcal{O}(x^5)$ for erf function, we have the approximation in (4); while ignoring the higher order terms of $\mathcal{O}(x^2)$ for logarithm function, we achieve the approximation in (5). The final lower bound (6) provides analytical updates of variational parameters for $q(\boldsymbol{\Theta})$. We evaluated this variational approximation on synthetic data where the cutpoints are available, and we can recover the true cutpoints. Markov chain Monte Carlo (MCMC) is also used as comparison for ordinal matrix completion problems, and identical performance are observed. Figure 2(a) in this supplementary material shows the results for the ordinal matrix completion task (questionnaires responses) on cognitive neuroscience data. We notice that the results (in terms of mean absolute error) based on MCMC and VB algorithms are pretty similar.

Learning the cutpoints

With the variational objective derived in (1) and (6), we can use Variational EM to learn the variational distribution $q(\boldsymbol{\Theta})$ (varational E-step), and the point estimates of cutpoints $\boldsymbol{G}$ and rotation matrix $\boldsymbol{Q}$ (varational M-step). Ignoring the constant terms w.r.t. $\boldsymbol{G}$, we have the following objective function for the cutpoints.

$$\mathcal{L}(\boldsymbol{G}) = \sum_{m=1}^{M_1} \tilde{\mathcal{L}}^m(\boldsymbol{g}^m) + \text{const} \tag{7}$$

$$\tilde{\mathcal{L}}^m(\boldsymbol{g}^m) = \sum_{l=1}^{L_m} \tilde{\mathcal{L}}^m_l \tag{8}$$

$$\begin{aligned} \tilde{\mathcal{L}}^m_l = & N^m_l [\log(\boldsymbol{g}^m_l - \boldsymbol{g}^m_{l-1}) - \frac{1}{6}\langle \gamma_m \rangle (\boldsymbol{g}^{m2}_l + {\boldsymbol{g}^m_{l-1}}^2 + \boldsymbol{g}^m_l \boldsymbol{g}^m_{l-1})] \\ & + \frac{1}{2}\langle \gamma_t \rangle (\boldsymbol{g}^m_l + \boldsymbol{g}^m_{l-1}) \sum_{i,j:\boldsymbol{X}^m_{ij}=l} \langle \boldsymbol{V}_i \rangle \langle \boldsymbol{W}^{(m)}_{:j} \rangle \end{aligned} \tag{9}$$

In above, L_m is the number of possible ordinal outcomes, and N_l^m is the number of data points having value l in m–th view. The gradients of $\tilde{\mathcal{L}}_l^m$ are also analytically available. Because $\boldsymbol{g}_0^m$ and $\boldsymbol{g}_{L_m}^m$ are fixed to achieve identifiablity, only the gradients with respect to $\boldsymbol{g}_l^m, l = 1, \cdots, L_m - 1$ are required. Note that the objective fucntion in (9) is concave w.r.t. $\boldsymbol{g}^m$; therefore, in each variational M-step, the solution $\hat{\boldsymbol{g}}^m$ given the variational distributions $q(\boldsymbol{\Theta})$ is global optimal. This constrained optimization problem (with ordering constraints $\boldsymbol{g}_l^m \le \boldsymbol{g}_{l'}^m$, for $l < l'$) can be solved efficiently using Newton's method, with the gradient provided below:

$$\begin{aligned} \nabla_{\boldsymbol{g}_l^m} \tilde{\mathcal{L}}^m(\boldsymbol{g}^m) \quad = \quad & N_l^m[\frac{1}{\boldsymbol{g}_l^m - \boldsymbol{g}_{l-1}^m} - \frac{1}{6}\langle\gamma_m\rangle(2\boldsymbol{g}_l^m + \boldsymbol{g}_{l-1}^m)] + \frac{1}{2}\langle\gamma_m\rangle \sum_{i,j:\boldsymbol{X}_{ij}^{(m)}=l} \langle\boldsymbol{V}_i\rangle\langle\boldsymbol{W}_{:j}^{(m)}\rangle \\ & + N_{l+1}^m[\frac{-1}{\boldsymbol{g}_{l+1}^m - \boldsymbol{g}_l^m} - \frac{1}{6}\langle\gamma_m\rangle(2\boldsymbol{g}_l^m + \boldsymbol{g}_{l+1}^m)] + \frac{1}{2}\langle\gamma_m\rangle \sum_{i,j:\boldsymbol{X}_{ij}^{m}=l+1} \langle\boldsymbol{V}_i\rangle\langle\boldsymbol{W}_{:j}^{(m)}\rangle \end{aligned} \tag{10}$$

Learning the rotation matrix

At each variational M-step, an unconstrained optimization problem to learn $\boldsymbol{Q}$ is solved to achieve faster convergence. After rotation, the variational distributions for $\boldsymbol{V}_i, \boldsymbol{W}_{:j}^{(m)}, \alpha_{mr}, \boldsymbol{\beta}_c$ are updated as follows.

$$\tilde{\boldsymbol{V}}_i \quad = \quad \boldsymbol{V}_i\boldsymbol{Q}^{-1} \sim \mathcal{N}(\boldsymbol{\mu}_{v,old}\boldsymbol{Q}^{-1}, \boldsymbol{Q}^{-T}\boldsymbol{\Sigma}_{v,old}\boldsymbol{Q}^{-1}) \tag{11}$$

$$\tilde{\boldsymbol{W}}_{:j}^{(m)} \quad = \quad \boldsymbol{Q}\boldsymbol{W}_{:j}^{(m)} \sim \mathcal{N}(\boldsymbol{Q}\boldsymbol{\mu}_{w,old}, \boldsymbol{Q}\boldsymbol{\Sigma}_{w,old}\boldsymbol{Q}^T) \tag{12}$$

$$\tilde{\alpha}_{mr} \quad \sim \quad \text{Ga}(a_\alpha + \frac{1}{2}K_m, b_\alpha + \frac{1}{2}\boldsymbol{Q}_{:r}^T\langle\boldsymbol{W}^{(m)}\boldsymbol{W}^{(m)\top}\rangle\boldsymbol{Q}_{:r}) \tag{13}$$

$$\tilde{\boldsymbol{\beta}}_c \quad \sim \quad \mathcal{N}(\tilde{\boldsymbol{\Sigma}}_\beta \sum_{i=1}^{N} \langle\boldsymbol{z}_{ic}\rangle\langle\boldsymbol{V}_i^T\rangle\boldsymbol{Q}^{-1}, \tilde{\boldsymbol{\Sigma}}_\beta) \tag{14}$$

$$\tilde{\boldsymbol{\Sigma}}_\beta \quad = \quad (\rho\boldsymbol{I}_R + \boldsymbol{Q}^{-T}\sum_{i=1}^{N}\langle\boldsymbol{V}_i^T\boldsymbol{V}_i\rangle\boldsymbol{Q}^{-1})^{-1}$$

Ignoring the terms that are constant w.r.t. $\boldsymbol{Q}$ in the variational lower bound, we have the following objective function w.r.t. $\boldsymbol{Q}$:

$$\begin{aligned} \tilde{\mathcal{L}}'(\boldsymbol{Q}) \quad & = \quad \langle\log\frac{p(\tilde{\boldsymbol{W}}, \tilde{\alpha})p(\tilde{\boldsymbol{V}})p(\tilde{\boldsymbol{\beta}})}{q(\tilde{\boldsymbol{W}})q(\tilde{\alpha})q(\tilde{\boldsymbol{V}})q(\tilde{\boldsymbol{\beta}})}\rangle \\ & = \quad \langle\log\frac{p(\tilde{\boldsymbol{V}})}{q(\tilde{\boldsymbol{V}})}\rangle + \langle\log\frac{p(\tilde{\boldsymbol{W}}|\tilde{\alpha})p(\tilde{\alpha})}{q(\tilde{\boldsymbol{W}})q(\tilde{\alpha})}\rangle + \langle\log\frac{p(\tilde{\boldsymbol{\beta}})}{q(\tilde{\boldsymbol{\beta}})}\rangle \end{aligned} \tag{15}$$

Inspecting (15) term by term, we have the analytical form as follows.

$$\langle\log\frac{p(\tilde{\boldsymbol{V}})}{q(\tilde{\boldsymbol{V}})}\rangle \quad = \quad -N\log|\boldsymbol{Q}| + \sum_{i=1}^{N}\log|\boldsymbol{\Sigma}_{\boldsymbol{V}_i}| - \frac{1}{2}\text{tr}(\boldsymbol{Q}^{-1}\langle\boldsymbol{V}^T\boldsymbol{V}\rangle\boldsymbol{Q}^{-T}) \tag{16}$$

$$\langle\log\frac{p(\tilde{\boldsymbol{W}}|\tilde{\alpha})p(\tilde{\alpha})}{q(\tilde{\boldsymbol{W}})q(\tilde{\alpha})}\rangle \quad = \quad \sum_{m=1}^{M_1+M_2+M_3} K_m\log|\boldsymbol{Q}| - \frac{K_m}{2}\sum_{r=1}^{R}\log\text{tr}(\boldsymbol{Q}_{:r}^T\langle\boldsymbol{W}^{(m)}\boldsymbol{W}^{(m)\top}\rangle\boldsymbol{Q}_{:r}) \tag{17}$$

$$\langle\log\frac{p(\tilde{\boldsymbol{\beta}})}{q(\tilde{\boldsymbol{\beta}})}\rangle \quad \overset{\rho\to 0}{\approx} \quad -C|\boldsymbol{Q}| + \frac{1}{2}C\log|\sum_{i=1}^{N}\langle\boldsymbol{V}_i^T\boldsymbol{V}_i\rangle| \tag{18}$$

Further, we have the gradients w.r.t. $\boldsymbol{Q}$ available in analytical form. If $\boldsymbol{Q} = \boldsymbol{I}_R$, no rotation is added; with rotation, $\boldsymbol{Q}$ draws $q(\boldsymbol{\Theta})$ towards the prior $p(\boldsymbol{\Theta})$ because (15) effective minimizes $KL(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta}))$ while not affecting the likelihood term $p(\boldsymbol{X}|\boldsymbol{\Theta})$. The solution of this unconstrained optimization problem is guaranteed to increase the variational lower bound.

Updating variational distributions

We use a mean field approximation to learn the variational distributions $q(\boldsymbol{\Theta})$:

$$q(\boldsymbol{\Theta}) = \prod_{m=M_1+M_2+1}^{M}\prod_{i=1}^{N} q(\boldsymbol{U}_i^{(m)})\prod_{i=1}^{N} q(\boldsymbol{V}_i)\prod_{m=1}^{M}\prod_{r=1}^{R} q(\boldsymbol{W}_r^{(m)})\prod_{m=1}^{M}\prod_{r=1}^{R} q(\alpha_{mr})\prod_{m=1}^{M} q(\gamma_m) \tag{19}$$

Update $\boldsymbol{V}_i$, for $i = 1, \cdots, N$ Multiview classification:

$$\begin{aligned} q(\boldsymbol{V}_i) &= \mathcal{N}(\boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v) \qquad (20) \\ \boldsymbol{\Sigma}_v &= (\boldsymbol{I} + \sum_{m=1}^{M} \langle\gamma_m\rangle \langle \boldsymbol{W}^{(m)} {\boldsymbol{W}^{(m)}}^\top \rangle + \sum_{c=1}^{C} \langle \boldsymbol{\beta}_c \boldsymbol{\beta}_c^\top \rangle)^{-1} \\ \boldsymbol{\mu}_v &= (\sum_{m=1}^{M_1} \langle\gamma_m\rangle \sum_{j=1}^{K_m} \langle {\boldsymbol{W}_{:j}^{(m)}}^\top \rangle \frac{\boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{i,j}} + \boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{i,j}-1}}{2} + \sum_{m=M_1+1}^{M_1+M_2} \langle\gamma_m\rangle \boldsymbol{X}_i^{(m)} \langle {\boldsymbol{W}^{(m)}}^\top \rangle \\ &\quad + \sum_{m=M_1+M_2+1}^{M_1+M_2+M_3} \langle\gamma_m\rangle \langle \boldsymbol{U}_i^{(m)} \rangle \langle {\boldsymbol{W}^{(m)}}^\top \rangle + \sum_{c=1}^{C} \langle z_{ic} \rangle \langle \boldsymbol{\beta}_c \rangle) \boldsymbol{\Sigma}_v \end{aligned}$$

$$\begin{aligned} \langle \boldsymbol{V}_i \rangle &= \boldsymbol{\mu}_v \qquad (21) \\ \langle \boldsymbol{V}_i^T \boldsymbol{V}_i \rangle &= \boldsymbol{\mu}_v^T \boldsymbol{\mu}_v + \boldsymbol{\Sigma}_v \qquad (22) \end{aligned}$$

Update α_{mr}

$$\begin{aligned} q(\alpha_{mr}) &= \text{Ga}(\tilde{a}_\alpha, \tilde{b}_\alpha) \qquad (23) \\ \tilde{a}_\alpha &= a_\alpha + \frac{K_m}{2} \\ \tilde{b}_\alpha &= b_\alpha + \frac{1}{2} \langle \boldsymbol{W}_r^{(m)} {\boldsymbol{W}_r^{(m)}}^\top \rangle \end{aligned}$$

$$\begin{aligned} \langle \alpha_{mr} \rangle &= \frac{\tilde{a}_\alpha}{\tilde{b}_\alpha} \qquad (24) \\ \langle \log \alpha_{mr} \rangle &= \psi(\tilde{a}_\alpha) - \log(\tilde{b}_\alpha) \qquad (25) \end{aligned}$$

Update γ_m

$$\begin{aligned} q(\gamma_m) &= \text{Ga}(\tilde{a}_\gamma, \tilde{b}_\gamma) \qquad (26) \\ \tilde{a}_\gamma &= a_\gamma + \frac{NK_m}{2} \end{aligned}$$

For $m = 1, \cdots, M_1$,

$$\tilde{b}_\gamma = b_\gamma + \frac{1}{2} \sum_{i,j} \langle (\boldsymbol{V}_i \boldsymbol{W}_{:j}^{(m)})^2 \rangle - \langle \boldsymbol{V}_i \boldsymbol{W}_{:j}^{(m)} \rangle (\boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}} + \boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}-1}) + \frac{1}{3} ({\boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}}}^2 + {\boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}-1}}^2 + \boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}} \boldsymbol{g}^m_{\boldsymbol{X}^{(m)}_{ij}-1})$$

For $m = M_1 + 1, \cdots, M_1 + M_2$,

$$\tilde{b}_\gamma = b_\gamma + \frac{1}{2} \text{tr}(\langle \boldsymbol{W}^{(m)} {\boldsymbol{W}^{(m)}}^\top \rangle \sum_{i=1}^{N} \langle \boldsymbol{V}_i^T \boldsymbol{V}_i \rangle) - \text{tr}(\sum_{i=1}^{N} \langle {\boldsymbol{U}_i^{(m)}}^\top \rangle \langle \boldsymbol{V}_i \rangle \langle \boldsymbol{W}^{(m)} \rangle) + \frac{1}{2} \text{tr}(\sum_{i=1}^{N} {\boldsymbol{X}_i^{(m)}}^\top \boldsymbol{X}_i^{(m)})$$

For $m = M_1 + M_2 + 1, \cdots, M_1 + M_2 + M_3$,

$$\tilde{b}_\gamma = b_\gamma + \frac{1}{2} \text{tr}(\langle \boldsymbol{W}^{(m)} {\boldsymbol{W}^{(m)}}^\top \rangle \sum_{i=1}^{N} \langle \boldsymbol{V}_i^T \boldsymbol{V}_i \rangle) - \text{tr}(\sum_{i=1}^{N} {\boldsymbol{X}_i^{(m)}}^\top \langle \boldsymbol{V}_i \rangle \langle \boldsymbol{W}^{(m)} \rangle) + \frac{1}{2} \text{tr}(\sum_{i=1}^{N} \langle {\boldsymbol{U}_i^{(m)}}^\top \boldsymbol{U}_i^{(m)} \rangle)$$

$$\begin{aligned} \langle \gamma_m \rangle &= \frac{\tilde{a}_\gamma}{\tilde{b}_\gamma} \qquad (27) \\ \langle \log \gamma_m \rangle &= \psi(\tilde{a}_\gamma) - \log(\tilde{b}_\gamma) \qquad (28) \end{aligned}$$

Update τ_m, for $m = M_1 + M_2 + 1, \cdots, M_1 + M_2 + M_3$

$$\begin{aligned} q(\tau_m) &= \text{Ga}(\tilde{a}_\tau, \tilde{b}_\tau) \qquad (29) \\ \tilde{a}_\tau &= a_\tau + \frac{N(N-1)}{4} \end{aligned}$$

$$\tilde{b}_\tau = b_\tau + \frac{1}{2} \sum_{i=1}^{N} \sum_{j>i} ((\boldsymbol{X}_{ij}^{(m)})^2 - 2\boldsymbol{X}_{ij}^{(m)} \langle \boldsymbol{U}_i^{(m)} \rangle \langle {\boldsymbol{U}_j^{(m)}}^\top \rangle + \text{tr}(\langle {\boldsymbol{U}_i^{(m)}}^\top \boldsymbol{U}_i^{(m)} \rangle \langle \boldsymbol{U}_j^{(m)} {\boldsymbol{U}_j^{(m)}}^\top \rangle))$$

$$\langle \tau_m \rangle = \frac{\tilde{a}_\tau}{\tilde{b}_\tau} \tag{30}$$

$$\langle \log \tau_m \rangle = \psi(\tilde{a}_\tau) - \log(\tilde{b}_\tau) \tag{31}$$

Update $\boldsymbol{\beta}_c$ and $\boldsymbol{z}_i$ for classification task.

$$q(\boldsymbol{\beta}_c) = \mathcal{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \tag{32}$$

$$\boldsymbol{\Sigma}_\beta = (\rho \boldsymbol{I}_R + \sum_{i=1}^{N} \langle \boldsymbol{V}_i^T \boldsymbol{V}_i \rangle)^{-1}$$

$$\boldsymbol{\mu}_\beta = \boldsymbol{\Sigma}_\beta \sum_{i=1}^{N} \langle \boldsymbol{z}_{ic} \rangle \langle \boldsymbol{V}_i^T \rangle$$

$$\langle \boldsymbol{\beta}_c \rangle = \boldsymbol{\mu}_\beta \tag{33}$$

$$\langle \boldsymbol{\beta}_c^T \boldsymbol{\beta}_c \rangle = \boldsymbol{\mu}_\beta^T \boldsymbol{\mu}_\beta + \boldsymbol{\Sigma}_\beta \tag{34}$$

$$q(\boldsymbol{z}_i) = \mathcal{TN}_{y_i}(\boldsymbol{\xi}, \boldsymbol{I}_C) \tag{35}$$

$$\xi_c = \langle \boldsymbol{V}_i \rangle \langle \boldsymbol{\beta}_c^T \rangle$$

In above, $\mathcal{TN}_{y_i}(\boldsymbol{z}_i)$ means $\boldsymbol{z}_{iy_i} = \max_c \boldsymbol{z}_{ic}$

$$\langle \boldsymbol{z}_{ic} \rangle_{c \neq y_i} = \xi_c - \frac{\mathbb{E}_{p(u)}[\phi(u + \xi_{y_i} - \xi_c)\Phi_u^{i,c}]}{\mathbb{E}_{p(u)}[\Phi(u + \xi_{y_i} - \xi_c)\Phi_u^{i,c}]} \tag{36}$$

$$\Phi_u^{i,c} = \prod_{j \neq y_i, c} \Phi(u + \xi_{y_i} - \xi_c)$$

$$\langle \boldsymbol{z}_{y_i c} \rangle = \xi_{y_i} - \sum_{j \neq y_i} (\langle \boldsymbol{z}_{ij} \rangle - \xi_j) \tag{37}$$

In above, $u \sim \mathcal{N}(0, 1)$. $\phi(.)$ and $\Phi(.)$ denote the p.d.f. and c.d.f. for normal distribution respectively.

Out-of-sample prediction

For out-of-sample data point(s) $\boldsymbol{X}_*$, we would like to infer $q(\boldsymbol{V}_*) \approx p(\boldsymbol{V}_* | \boldsymbol{X}_*^{(1)}, \cdots, \boldsymbol{X}_*^{(M_1+M_2+M_3)})$. Based on the chain rule, we have

$$q(\boldsymbol{V}_*) \propto p(\boldsymbol{V}_*) \prod_m \int p(\boldsymbol{X}_*^{(m)} | \boldsymbol{U}_*^{(m)}) p(\boldsymbol{U}_*^{(m)} | \boldsymbol{V}_*) d\boldsymbol{U}_*^{(m)} = p(\boldsymbol{V}_*) \prod_m p(\boldsymbol{X}_*^{(m)} | \boldsymbol{V}_*) \tag{38}$$

$$\propto \int p(\boldsymbol{V}_* | \boldsymbol{U}_*^{(m)}, \cdots, \boldsymbol{U}_*^{(m)}) \prod_m p(\boldsymbol{U}_*^{(m)} | \boldsymbol{X}_*^{(m)}) d\boldsymbol{U}_*^{(m)} \tag{39}$$

For ordinal feature views, (38) is used, where the likelihood term $p(\boldsymbol{X}_*^{(m)} | \boldsymbol{V}_*)$ is approximated by (3). Thus the $q(\boldsymbol{V}_*) \approx p(\boldsymbol{V}_* | \boldsymbol{X}_*^{(1)}, \cdots, \boldsymbol{X}_*^{(M_1)})$ is accordingly approximated by a Gaussian distribution. For Gaussian feature views, because $p(\boldsymbol{X}_*^{(m)} | \boldsymbol{V}_*), m = M_1 + 1, \cdots, M_1 + M_2$ is a Gaussian distribution, we can directly apply (38) and obtain the Gaussian posterior $q(\boldsymbol{V}_*) \approx p(\boldsymbol{V}_* | \boldsymbol{X}_*^{(M_1+1)}, \cdots, \boldsymbol{X}_*^{(M_1+M_2)})$.

For similarity-based views, because $\boldsymbol{U}_{*:}^{(m)}$ cannot be integrated out, (39) is used. We need to estimate $q(\boldsymbol{U}_*^{(m)})$ and $q(\boldsymbol{V}_*)$ in two stages as follows.

$$p(\boldsymbol{U}_*^{(m)}) = \int p(\boldsymbol{U}_*^{(m)} | \boldsymbol{V}_*) p(\boldsymbol{V}_*) d\boldsymbol{V}_*$$

$$= \mathcal{N}(\boldsymbol{0}, \langle {\boldsymbol{W}^{(m)}}^\top \boldsymbol{W}^{(m)} \rangle + \langle \gamma_m \rangle^{-1} \boldsymbol{I}_{K_m}) \tag{40}$$

$$p(\boldsymbol{X}_*^{(m)} | \boldsymbol{U}_*^{(m)}) = \prod_{j=1}^{N} \mathcal{N}(\boldsymbol{U}_*^{(m)} \langle {\boldsymbol{U}_j^{(m)}}^\top \rangle, \langle \tau_m \rangle^{-1}) \tag{41}$$

Therefore we have the following posterior for $p(\boldsymbol{U}_*^{(m)}|\boldsymbol{X}_*^{(m)})$:

$$\begin{aligned} p(\boldsymbol{U}_*^{(m)}|\boldsymbol{X}_*^{(m)}) &= \mathcal{N}(\langle\tau_m\rangle \sum_{j=1}^{N} \boldsymbol{X}_{*j}^{(m)} \langle \boldsymbol{U}_j^{(m)}\rangle \boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_u) \qquad (42) \\ \boldsymbol{\Sigma}_u &= ((\langle {\boldsymbol{W}^{(m)}}^\top \boldsymbol{W}^{(m)}\rangle + \langle\gamma_m\rangle^{-1}\boldsymbol{I}_{K_m})^{-1} + \langle\tau_m\rangle \sum_{j=1}^{N} \langle {\boldsymbol{U}_j^{(m)}}^\top \boldsymbol{U}_j^{(m)}\rangle)^{-1} \end{aligned}$$

We also observed that

$$\begin{aligned} p(\boldsymbol{V}_*|\boldsymbol{U}_*) &\propto p(\boldsymbol{V}_*) \prod_{m=M_1+M_2+1}^{M_1+M_2+M_3} p(\boldsymbol{U}_*^{(m)}|\boldsymbol{V}_*) \\ &= \mathcal{N}(\sum_{m=M_1+M_2+1}^{M_1+M_2+M_3} \langle\gamma_m\rangle \langle \boldsymbol{U}_*^{(m)}\rangle \langle {\boldsymbol{W}^{(m)}}^\top\rangle \boldsymbol{\Sigma}_{v|u}, \boldsymbol{\Sigma}_{v|u}) \qquad (43) \\ \boldsymbol{\Sigma}_{v|u} &= (\boldsymbol{I}_R + \sum_{m=M_1+M_2+1}^{M_1+M_2+M_3} \langle\gamma_m\rangle \langle \boldsymbol{W}^{(m)} {\boldsymbol{W}^{(m)}}^\top\rangle)^{-1} \end{aligned}$$

Combining (42) and (43), we have $q(\boldsymbol{V}_*) \approx p(\boldsymbol{V}_*|\boldsymbol{X}_*^{(M_1+M_2+1)}, \cdots, \boldsymbol{X}_*^{(M_1+M_2+M_3)})$, which is a Gaussian distribution. Finally, combing the ordinal, real (Gaussian), and similarity-based views in a sequential manner (dealing with feature views first and using this posterior as the prior for the similarity-based views), we get the overall out-of-sample prediction for $\boldsymbol{V}_*$:

$$\begin{aligned} q(\boldsymbol{V}_*) &= \mathcal{N}(\boldsymbol{\mu}_v, \boldsymbol{\Sigma}_v) \qquad (44) \\ \boldsymbol{\mu}_v &= (\sum_{m=1}^{M_1} \sum_{j} \langle\gamma_m\rangle \frac{\boldsymbol{g}^m_{\boldsymbol{X}_{ij}^{(m)}} + \boldsymbol{g}^m_{\boldsymbol{X}_{ij}^{(m)}-1}}{2} \langle {\boldsymbol{W}_{:j}^{(m)}}^\top\rangle + \sum_{m=M_1+1}^{M_1+M_2} \langle\gamma_m\rangle \boldsymbol{X}_*^{(m)} \langle {\boldsymbol{W}^{(m)}}^\top\rangle \\ &\quad + \sum_{m=M_1+M_2+1}^{M_1+M_2+M_3} \langle\gamma_t\rangle \langle \boldsymbol{U}_*^{(m)}\rangle \langle {\boldsymbol{W}^{(m)}}^\top\rangle)\boldsymbol{\Sigma}_v \\ \boldsymbol{\Sigma}_v &= (\boldsymbol{I}_R + \sum_{m=1}^{M_1+M_2+M_3} \langle\gamma_t\rangle \langle \boldsymbol{W}^{(m)} {\boldsymbol{W}^{(m)}}^\top\rangle)^{-1} \end{aligned}$$

Streaming extension

In the setting where data points are observed in a streaming fashion, we need to update local variables for newly observed data $\{\boldsymbol{V}_*, \boldsymbol{U}_*^{(m)}, \boldsymbol{z}_*\}$, and global variables $\{\boldsymbol{W}^{(m)}, \alpha_{mr}, \boldsymbol{\beta}_c\}$. The hyperparameter γ_m is fixed at a reasonable estimate for simplicity, which can also be updated similarly to the batch setting.

Local variables

As derived in Section , treating each newly observed example $\boldsymbol{X}_*$ (having some or all the views) as an *out-of-sample* point, we have the variational estimate $q(\boldsymbol{U}_*^{(m)})$ for $m = M_1 + M_2 + 1, \cdots, M_1 + M_2 + M_3$, and $q(\boldsymbol{V}_*)$, provided in (42) and (44). Further, We can natually update $\boldsymbol{z}_*$ following (35).

Global variables

Once the local variable distributions are learned, we can update the global variables $\boldsymbol{\Theta}^g = \{\boldsymbol{W}^{(m)}, \alpha_{mr}, \boldsymbol{\beta}_c\}$, based on $q(\boldsymbol{\Theta}^g_{n+1}) \propto q(\boldsymbol{\Theta}^g_n) p(\boldsymbol{X}_*|\boldsymbol{\Theta}^g_n)$.

Specifically, we can update $\boldsymbol{W}_{:j}^{(m)}$ for similarity-based views as follows (updates for feature-based views have a similar form by replacing $\boldsymbol{U}_*^{(m)}$, as in the previous section.

$$\begin{aligned} q(\boldsymbol{W}_{:j}^{(m)}) &= \mathcal{N}(\boldsymbol{\mu}_{w,n}, \boldsymbol{\Sigma}_{w,n}) \to \mathcal{N}(\boldsymbol{\mu}_{w,n+1}, \boldsymbol{\Sigma}_{w,n+1}) \qquad (45) \\ \boldsymbol{\Sigma}_{w,n+1} &= (\boldsymbol{\Sigma}_{w,n}^{-1} + \langle\gamma_m\rangle \langle \boldsymbol{V}_*^T \boldsymbol{V}_*\rangle)^{-1} \\ \boldsymbol{\mu}_{w,n+1} &= \boldsymbol{\Sigma}_{w,n+1}(\boldsymbol{\Sigma}_{w,n}^{-1}\boldsymbol{\mu}_{w,n} + \langle\gamma_m\rangle \langle \boldsymbol{V}_*^T\rangle \langle \boldsymbol{U}_{*j}^{(m)}\rangle) \end{aligned}$$

Updating α_{mr} is the same as batch setting in (23) because $q(\alpha_{mr})$ does not directly depend on local variables.

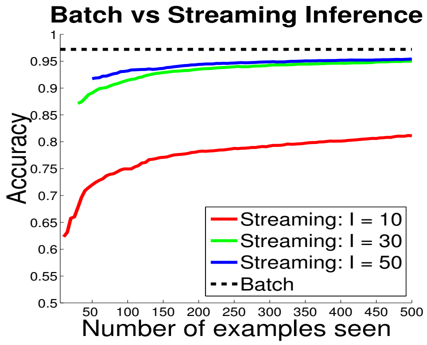

Figure 1: Steaming Extension

We can also update the classifier $\boldsymbol{\beta}_c$. Computational speed-up details are not discussed here, which includes avoiding matrix inversions for each new observation.

$$\begin{aligned} q(\boldsymbol{\beta}_c) &= \mathcal{N}(\boldsymbol{\mu}_{\beta,n}, \boldsymbol{\Sigma}_{\beta,n}) \rightarrow \mathcal{N}(\boldsymbol{\mu}_{\beta,n+1}, \boldsymbol{\Sigma}_{\beta,n+1}) \\ \boldsymbol{\Sigma}_{\beta,n+1} &= (\boldsymbol{\Sigma}_{\beta,n}^{-1} + \langle \boldsymbol{V}_*^T \boldsymbol{V}_* \rangle)^{-1} \\ \boldsymbol{\mu}_{\beta,n+1} &= \boldsymbol{\Sigma}_{\beta,n+1}(\boldsymbol{\Sigma}_{\beta,n}^{-1} \boldsymbol{\mu}_{\beta,n} + \langle \boldsymbol{z}_{*c} \rangle \langle \boldsymbol{V}_*^T \rangle) \end{aligned} \tag{46}$$

Experiments

We demonstrate MLFS in a streaming setting, revisiting the Digits data classification experiment for: (i) MLFS with batch inference with a training set of 500 examples, (ii) MLFS with streaming inference, processing one example at a time (for various choices of the initial pool size I) and doing only a *single pass* over the data. Figure 1 shows the average accuracy changing with number of visited examples increasing, run with 10 data splits. While it is unreasonable to expect that a truly streaming algorithm (seeing each example just once) will outperform its batch counterpart, it attains reasonably competitive accuracies even when running with very small initial pool sizes.

Additional results for cognitive neuroscience data

Here, we include some additional results on the cognitive neuroscience data for two tasks: matrix completion of ordinal responses (questionnaires data), and prediction of fMRI responses. For the first task, Figure 2 (a)shows the average mean absolute error (MAE) for different percentage of missingness over 10 runs considering three scenarios: (i) MLFS fitted using the proposed VB algorithm, (ii) MLFS fitted using MCMC, and (iii) MLFS fitted using the proposed VB algorithm but considering all the ordinal views concatenated as a single ordinal matrix. We used an unoptimized MATLAB implementation and our VB based inference method converged in about 10 iterations (in terms of variational lower bound). As we can see, for MLFS with all the views, our VB result is competitive to MCMC; moreover, both outperform the baseline of concatenating all the ordinal views. For the second task to predict fMRI responses leveraging information from other views, Figure 2(b) shows a plot of observed vs. predicted values. The points roughly follow a straight line indicating good predictions.

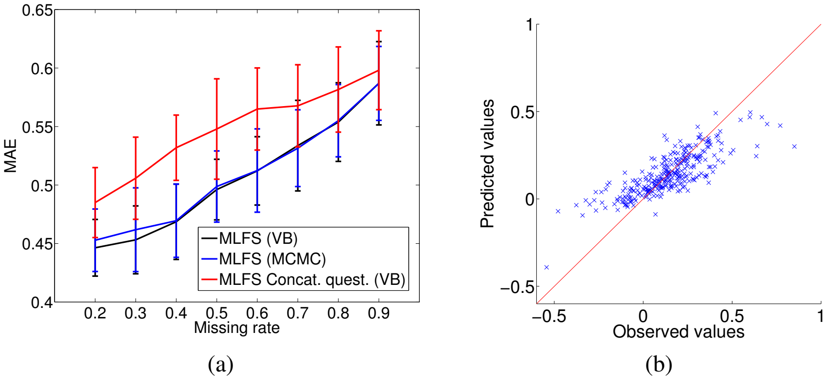

Figure 2: (a) Average mean absolute error (MAE) for the ordinal responses over 10 runs as a function of the fraction of missing data. Error bars indicate the standard deviation around the mean. (b) Observed vs. predicted fMRI values (amygdala and VS) from 30% of the subjects.