Supplementary Material for: "A Bayesian Framework for Multi-Modality Analysis of Mental Health"

This document presents supplementary materials to the submitted paper. It contains: (*i*) Details about inference procedures; (*ii*) Details on computational time; (*iii*) Comparison between VB and MCMC results; and (*iv*) Additional results on neuroscience data.

1 Details about posterior inference

1.1 Variational Bayes derivation

Without lost of generality, throughout this derivation we only consider inference for the model without the regression component. Thus, we infer the variational distribution for the latent variables, collectively referred to as $\boldsymbol{\Theta}$, such that $\boldsymbol{\Theta} = \{\tilde{\boldsymbol{W}}^{(m)}, \boldsymbol{\alpha}_m, \gamma_m, \tau_m\}_{m=1}^{M}, \boldsymbol{V}\}$ along with $\{\{\boldsymbol{\mu}_j\}_{j=1}^{J}, \boldsymbol{z}\}$ for clustering. For the ordinal views, we denote the cut-points as $\boldsymbol{G} = \{\boldsymbol{g}^m\}_{m=1}^{M_1}$ and the rotation matrix as $\boldsymbol{Q}$. Sometimes, for brevity, we will use $\{\tilde{\boldsymbol{W}}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\tau}\}$ for $\{\tilde{\boldsymbol{W}}^{(m)}, \boldsymbol{\alpha}_m, \gamma_m, \tau_m\}_{m=1}^{M}$ and $\{\boldsymbol{\mu}, \boldsymbol{z}\}$ for $\{\{\boldsymbol{\mu}_j\}_{j=1}^{J}, \boldsymbol{z}\}$, respectively. The data from all views are collectively referred to as $\boldsymbol{\mathcal{Y}}$. We approximate the true posterior $p(\boldsymbol{\Theta}|\boldsymbol{\mathcal{Y}}, \boldsymbol{G}, \boldsymbol{Q}))$ by its mean-field approximation:

$$q(\boldsymbol{\Theta}) = \prod_{i=1}^{N} q(\boldsymbol{v}_i) \prod_{m=1}^{M} \left(\prod_{k=1}^{K} q(\tilde{\boldsymbol{w}}_k^{(m)}) \prod_{k=1}^{K} q(\alpha_{mk}) q(\gamma_m) q(\tau_m) \right). \tag{1}$$

The goal here is to minimize the KL-divergence $KL(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta}|\boldsymbol{\mathcal{Y}}, \boldsymbol{G}, \boldsymbol{Q}))$, which is equivalent to maximizing the evidence lower bound (ELBO) given by

$$\begin{aligned} \mathcal{L}(q(\boldsymbol{\Theta}), \boldsymbol{G}, \boldsymbol{Q}) &= \mathbb{E}_{q(\boldsymbol{\Theta})}[\log p(\boldsymbol{\mathcal{Y}}, \boldsymbol{\Theta}|\boldsymbol{G}, \boldsymbol{Q}) - \log(q(\boldsymbol{\Theta}))] \qquad (2) \\ &= \langle \log p(\boldsymbol{\mathcal{Y}}, \boldsymbol{\Theta}) - \log q(\boldsymbol{\Theta}) \rangle_{q(\boldsymbol{\Theta})} \\ &= \langle \log p(\boldsymbol{\mathcal{Y}}|\boldsymbol{W}, \boldsymbol{V}, \boldsymbol{\gamma}, \boldsymbol{\tau}) \rangle + \langle \log \frac{p(\boldsymbol{W}|\boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\boldsymbol{V}) p(\boldsymbol{\gamma}) p(\boldsymbol{\tau})}{q(\boldsymbol{W}) q(\boldsymbol{\alpha}) q(\boldsymbol{V}) q(\boldsymbol{\gamma}) q(\boldsymbol{\tau})} \rangle \end{aligned}$$

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 (2). Only the ordinal views are considered in this subsection. For real-valued views no such approximation is needed. The approximation for the ordinal views proceeds as follows:

$$
\begin{aligned}
\langle \log p(\mathcal{Y}|\boldsymbol{W},\boldsymbol{v},\boldsymbol{\gamma},\boldsymbol{\tau})\rangle_{q(\boldsymbol{\Theta})} &= \sum_{i,j,m} \langle \log \int p(y_{ij}^{(m)}|x_{ij}^{(m)}) p(x_{ij}^{(m)}|\boldsymbol{v}_i, \boldsymbol{W}_{:j}^{(m)}, \gamma_m, \tau_m)\, dx_{ij}^{(m)} \rangle \\
&= \sum_{i,j,m} \langle \log \int\limits_{\boldsymbol{g}^m_{y_{ij}^{(m)}-1}}^{\boldsymbol{g}^m_{y_{ij}^{(m)}}} \mathcal{N}(x_{ij}^{(m)}; \boldsymbol{v}_i \tilde{\boldsymbol{W}}_{:j}^{(m)}, \gamma_m^{-1})\, dx_{ij}^{(m)} \rangle \\
&= \sum_{i,j,m} \langle \log[\Phi(\beta_{i,j,m}) - \Phi(\alpha_{i,j,m})] \rangle \\
&= \text{const} + \sum_{i,j,m} \langle \log \int\limits_{\alpha_{i,j,m}}^{\beta_{i,j,m}} \exp(-\frac{u^2}{2}) du \rangle \\
&\geq \sum_{i,j,m} \frac{1}{\langle \beta_{i,j,m} - \alpha_{i,j,m} \rangle} \langle \int\limits_{\alpha_{i,j,m}}^{\beta_{i,j,m}} \log(\beta_{i,j,m} - \alpha_{i,j,m}) - \frac{1}{2} u^2 du \rangle + \text{const} \\
&= \sum_{i,j,m} \log(\boldsymbol{g}^m_{y_{ij}^{(m)}} - \boldsymbol{g}^m_{y_{ij}^{(m)}-1}) + \frac{1}{2}\langle \log \gamma_m \rangle - \frac{1}{2}\langle \gamma_m \rangle \langle (\boldsymbol{v}_i \tilde{\boldsymbol{W}}_{:j}^{(m)})^2 \rangle \\
&\quad + \frac{1}{2}\langle \gamma_m \rangle \langle \boldsymbol{v}_i \tilde{\boldsymbol{W}}_{:j}^{(m)} \rangle (\boldsymbol{g}^m_{y_{ij}^{(m)}} + \boldsymbol{g}^m_{y_{ij}^{(m)}-1}) \\
&\quad - \frac{1}{6}\langle \gamma_m \rangle ((\boldsymbol{g}^m_{y_{ij}^{(m)}})^2 + (\boldsymbol{g}^m_{y_{ij}^{(m)}-1})^2 + \boldsymbol{g}^m_{y_{ij}^{(m)}} \boldsymbol{g}^m_{y_{ij}^{(m)}-1}) + \text{const}. \qquad (3)
\end{aligned}
$$

In the above, $\tilde{\boldsymbol{W}}_{:j}^{(m)} = \sqrt{1+\tau_m^{-1}}\boldsymbol{W}_{:j}^{(m)}$, $\beta_{i,j,m} = (\boldsymbol{g}^m_{y_{ij}^{(m)}} - \boldsymbol{V}_i \tilde{\boldsymbol{W}}_{:j}^{(m)})\gamma_m^{-\frac{1}{2}}$, $\alpha_{i,j,m} = (\boldsymbol{g}^m_{y_{ij}^{(m)}-1} - \boldsymbol{V}_i \tilde{\boldsymbol{W}}_{:j}^{(m)})\gamma_m^{-\frac{1}{2}}$, 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. The final lower bound (3) provides analytical updates of variational parameters for $q(\boldsymbol{\Theta})$. We evaluated this variational approximation on synthetic data where the cut-points are available, and we can recover the true cut-points. Markov chain Monte Carlo (MCMC) is also used as comparison for ordinal matrix completion problems, and identical performance are observed. Figure 1, in this supplementary material, shows the results for the ordinal matrix completion task on the NEO questionnaire. We notice that the results (in terms of mean absolute error) based on MCMC and VB algorithms are pretty similar.

Learning the cut-points

With the variational objective derived in (2) and (3), we can use Variational EM to learn the variational distribution $q(\boldsymbol{\Theta})$ (varational E-step), and the point estimates of cut-points $\boldsymbol{G}$ 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 cut-points.

$$\begin{aligned}
\mathcal{L}(\boldsymbol{G}) &= \sum_{m=1}^{M_1} \tilde{\mathcal{L}}^m(\boldsymbol{g}^m) + \text{const} \\
\tilde{\mathcal{L}}^m(\boldsymbol{g}^m) &= \sum_{l=1}^{L_m} \tilde{\mathcal{L}}_l^m \\
\tilde{\mathcal{L}}_l^m &= N_l^m[\log(\boldsymbol{g}_l^m - \boldsymbol{g}_{l-1}^m) - \frac{1}{6}\langle\gamma_m\rangle(\boldsymbol{g}_l^{m2} + {\boldsymbol{g}_{l-1}^m}^2 + \boldsymbol{g}_l^m \boldsymbol{g}_{l-1}^m)] \\
&\quad + \frac{1}{2}\langle\gamma_t\rangle(\boldsymbol{g}_l^m + \boldsymbol{g}_{l-1}^m) \sum_{i,j:y_{ij}^m = l} \langle\boldsymbol{v}_i\rangle\langle\tilde{\boldsymbol{W}}_{:j}^{(m)}\rangle \qquad (4)
\end{aligned}$$

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 $\boldsymbol{g}_0^m = -G$ and $\boldsymbol{g}_{L_m}^m = +G$ are fixed to achieve identifiability, only the gradients w.r.t. $\boldsymbol{g}_l^m, l = 1, \cdots, L_m - 1$ are required. Note that the objective function in (4) 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 $\boldsymbol{g}_l^m \leq \boldsymbol{g}_{l'}^m$, for $l < l'$) can be solved efficiently using the Newton's method, with the gradient provided below:

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

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 $\boldsymbol{v}_i, \tilde{\boldsymbol{W}}_{:j}^{(m)}$ and

α_{mr} are updated as follows:

$$\begin{aligned}
\check{\boldsymbol{v}}_i &= \boldsymbol{v}_i \boldsymbol{Q}^{-1} \sim \mathcal{N}(\boldsymbol{\mu}_{v,old}\boldsymbol{Q}^{-1}, \boldsymbol{Q}^{-T}\boldsymbol{\Sigma}_{v,old}\boldsymbol{Q}^{-1}) \\
\check{\boldsymbol{W}}^{(m)}_{:j} &= \boldsymbol{Q}\tilde{\boldsymbol{W}}^{(m)}_{:j} \sim \mathcal{N}(\boldsymbol{Q}\boldsymbol{\mu}_{w,old}, \boldsymbol{Q}\boldsymbol{\Sigma}_{w,old}\boldsymbol{Q}^{T}) \\
\check{\alpha}_{mr} &\sim \mathrm{Ga}(a_\alpha + \frac{1}{2}P_m, b_\alpha + \frac{1}{2}\boldsymbol{Q}^T_{:r}\langle \tilde{\boldsymbol{W}}^{(m)}\tilde{\boldsymbol{W}}^{(m)\top}\rangle \boldsymbol{Q}_{:r}).
\end{aligned}$$

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. $\boldsymbol{Q}$:

$$\begin{aligned}
\tilde{\mathcal{L}}'(\boldsymbol{Q}) &= \langle \log \frac{p(\check{\boldsymbol{W}}, \check{\alpha})p(\check{\boldsymbol{V}})}{q(\check{\boldsymbol{W}})q(\check{\alpha})q(\check{\boldsymbol{V}})} \rangle \\
&= \langle \log \frac{p(\check{\boldsymbol{V}})}{q(\check{\boldsymbol{V}})} \rangle + \langle \log \frac{p(\check{\boldsymbol{W}}|\check{\alpha})p(\check{\alpha})}{q(\check{\boldsymbol{W}})q(\check{\alpha})} \rangle. \qquad (5)
\end{aligned}$$

Inspecting (5) term by term, we have the analytical form as follows

$$\begin{aligned}
\langle \log \frac{p(\check{\boldsymbol{V}})}{q(\check{\boldsymbol{V}})} \rangle &= -N \log |\boldsymbol{Q}| + \sum_{i=1}^{N} \log |\boldsymbol{\Sigma}_{\boldsymbol{V}_i}| - \frac{1}{2}\mathrm{tr}(\boldsymbol{Q}^{-1}\langle \boldsymbol{V}^T\boldsymbol{V}\rangle \boldsymbol{Q}^{-T}) \\
\langle \log \frac{p(\check{\boldsymbol{W}}|\check{\alpha})p(\check{\alpha})}{q(\check{\boldsymbol{W}})q(\check{\alpha})} \rangle &= \sum_{m=1}^{M} P_m \log |\boldsymbol{Q}| - \frac{P_m}{2} \sum_{r=1}^{K} \log \mathrm{tr}(\boldsymbol{Q}^T_{:r}\langle \tilde{\boldsymbol{W}}^{(m)}\tilde{\boldsymbol{W}}^{(m)\top}\rangle \boldsymbol{Q}_{:r}).
\end{aligned}$$

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 (5) effective minimizes $KL(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta}))$ while not affecting the likelihood term $p(\mathcal{Y}|\boldsymbol{\Theta})$. The solution of this unconstrained optimization problem is guaranteed to increase the variational lower bound.

For the model adapted to the multi-view clustering problem, with J clusters, (5) is modified as below, which is also analytically available

$$\begin{aligned}
\tilde{\mathcal{L}}'(\boldsymbol{Q}) &= \sum_{i=1}^{N}\sum_{c=1}^{J} \log \frac{(p(\boldsymbol{v}_i|z_i, \boldsymbol{\mu}_{z_i})p(z_i))^{z_i=c}}{(q(\boldsymbol{v}_i|z_i, \boldsymbol{\mu}_{z_i})q(z_i))^{z_i=c}} + \sum_{c=1}^{J}\sum_{r=1}^{K} \log \frac{p(\boldsymbol{\mu}_{cr})}{p(\boldsymbol{\mu}_{cr})} + \langle \log \frac{p(\check{\boldsymbol{W}}|\tilde{\alpha})p(\check{\alpha})}{q(\check{\boldsymbol{W}})q(\check{\alpha})} \rangle \\
&= N \log |\boldsymbol{Q}| - \frac{1}{2}\sum_{i=1}^{N}\sum_{c=1}^{J} q(z_i = c)\mathrm{tr}[\boldsymbol{Q}^{-1}(\langle \boldsymbol{v}_i^T\boldsymbol{v}_i\rangle + \langle \boldsymbol{\mu}_c\boldsymbol{\mu}_c^T\rangle - 2\langle \boldsymbol{v}_i^T\rangle\langle \boldsymbol{\mu}_c\rangle^T)\boldsymbol{Q}^{-T}] \\
&\quad + \langle \log \frac{p(\check{\boldsymbol{W}}|\check{\alpha})p(\check{\alpha})}{q(\check{\boldsymbol{W}})q(\check{\alpha})} \rangle + \mathrm{const.}
\end{aligned}$$

Updating variational distributions

We use a mean field approximation to learn the variational distributions $q(\boldsymbol{\Theta})$ in (1). Below, we summarize the variational updates for $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$, and clustering parameters $\boldsymbol{\mu}$ and $\boldsymbol{z}$.

Updating α_{mk}: $q(\alpha_{mk}) = \text{Ga}(\tilde{a}_\alpha, \tilde{b}_\alpha)$ such that $\tilde{a}_\alpha = a_\alpha + P_m/2$ and $\tilde{b}_\alpha = b_\alpha + \langle \boldsymbol{w}_k^{(m)} \boldsymbol{w}_k^{(m)^\top} \rangle/2$. Moreover, $\langle \alpha_{mk} \rangle = (\tilde{a}_\alpha)(\tilde{b}_\alpha)$ and $\langle \log \alpha_{mk} \rangle = \psi(\tilde{a}_\alpha) - \log(\tilde{b}_\alpha)$, for $k = 1, \ldots, K$ and $m = 1, \ldots, M$.

Updating γ_m: $q(\gamma_m) = \text{Ga}(\tilde{a}_\gamma, \tilde{b}_\gamma)$ where $\tilde{a}_\gamma = a_\gamma + (NP_m)/2$. Also:
For $m = 1, \cdots, M_1$,

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

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

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

Updating τ_m: Update the distribution of τ_m which is proportional to

$$\tau_m^{\frac{P_m}{2} + a_\tau - 1} (\tau_m + 1)^{-\frac{P_m}{2}} \exp \left[-\tau_m \left(b_\tau + \frac{1}{2} (\tau_m + 1)^{-1} \alpha_{mk} \tilde{\boldsymbol{w}}_k^{(m)} \tilde{\boldsymbol{w}}_k^{(m)\top} \right) \right]$$

and calculate $\langle \tau_m \rangle$ using importance sampling (Rubin, 1987).

Updating $\boldsymbol{\mu}_c$ and z_i for the clustering task:

$$\begin{aligned}
q(\boldsymbol{\mu}_c) &\sim \mathcal{N}((\eta + \sum_i^N q(z_i = c))^{-1} \sum_i^N q(z_i = c) \boldsymbol{v}_i^\top, (\eta + \sum_i^N q(z_i = c))^{-1}) \\
\langle \boldsymbol{\mu}_c \rangle &= (\eta + \sum_i^N q(z_i = c))^{-1} \sum_i^N q(z_i = c) \boldsymbol{v}_i^\top \\
\langle \boldsymbol{\mu}_c \boldsymbol{\mu}_c^\top \rangle &= \langle \boldsymbol{\mu}_c \rangle \langle \boldsymbol{\mu}_c \rangle^\top + (\eta + \sum_i^N q(z_i = c))^{-1} \boldsymbol{I} \\
q(z_i = c) &\propto \exp[-\frac{1}{2} \sum_{k=1}^K \langle \boldsymbol{\mu}_{ck}^2 \rangle + \langle \boldsymbol{v}_{ik}^2 \rangle - 2 \langle \boldsymbol{v}_{ik} \rangle \langle \boldsymbol{\mu}_{ck} \rangle].
\end{aligned}$$

Out-of-sample prediction

For out-of-sample data points $\boldsymbol{Y}_*$, we would like to infer $q(\boldsymbol{V}_*) \approx p(\boldsymbol{V}_*|\boldsymbol{Y}_*^{(1)}, \cdots, \boldsymbol{Y}_*^{(M)})$. Based on the chain rule, we have

$$\begin{aligned} q(\boldsymbol{V}_*) &\propto p(\boldsymbol{V}_*) \prod_m \int p(\boldsymbol{Y}_*^{(m)}|\boldsymbol{X}_*^{(m)}) p(\boldsymbol{X}_*^{(m)}|\boldsymbol{V}_*) d\boldsymbol{X}_*^{(m)} = p(\boldsymbol{V}_*) \prod_m p(\boldsymbol{Y}_*^{(m)}|\boldsymbol{V}_*) \qquad (6) \\ &\propto \int p(\boldsymbol{V}_*|\boldsymbol{X}_*^{(m)}, \ldots, \boldsymbol{X}_*^{(m)}) \prod_m p(\boldsymbol{X}_*^{(m)}|\boldsymbol{Y}_*^{(m)}) d\boldsymbol{X}_*^{(m)} \end{aligned}$$

For ordinal-based views, (6) is used, where the likelihood term $p(\boldsymbol{Y}_*^{(m)}|\boldsymbol{V}_*)$ 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 real-based views, given that $p(\boldsymbol{X}_*^{(m)}|\boldsymbol{V}_*), m = M_1 + 1, \ldots, M_1 + M_2$ is a Gaussian distribution, we can directly apply (6) and obtain the Gaussian posterior $q(\boldsymbol{V}_*) \approx p(\boldsymbol{V}_*|\boldsymbol{X}_*^{(M_1+1)}, \cdots, \boldsymbol{X}_*^{(M_1+M_2)})$.

Finally, combining the ordinal- and real-based views in a sequential manner, we get the overall out-of-sample prediction for $\boldsymbol{V}_*$:

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

1.2 MCMC derivation

For ordinal-based views, we use the algorithm proposed in Albert and Chib (1997) to learn the real matrices $\boldsymbol{X}^{(m)}$, $m = 1, \ldots, M_1$, and the cut-points $\boldsymbol{g}^m = (g_1^m, \ldots, g_{L_m-1}^m)$. Specifically,

Sampling $X_{ij}^{(m)}$: $(\boldsymbol{X}_{ij}^{(m)}|y_{ij}^{(m)} = h) \sim \mathcal{N}_{(g_{h-1}^m, g_h^m)}(\boldsymbol{v}_i \tilde{\boldsymbol{W}}_{:j}^{(m)}, \gamma_m^{-1})$, a truncated normal distribution on the interval (g_{h-1}^m, g_h^m).

Sampling the cut-points: As considered in Albert and Chib (2001), we transform $\boldsymbol{g}^m$ into a real-valued vector $\boldsymbol{\lambda}^m = (\lambda_1^m, \ldots, \lambda_{L_m-1}^m)$ such that $\lambda_1^m = \log g_1^m$ and $\lambda_h^m = \log(g_h^m - g_{h-1}^m)$ for $2 \leq h \leq L_m - 1$ with inverse map given by $g_h^m = \sum_{i=1}^h \exp(\lambda_i^m)$ for $1 \leq h \leq L_m - 1$. The transformed cut-point vector $\boldsymbol{\lambda}^m$ follows a multivariate normal prior distribution $\pi(\boldsymbol{\lambda}^m) = \mathcal{N}(\boldsymbol{0}, \delta \boldsymbol{I})$ and the full conditional posterior does not have a close form, therefore a Metropolis-Hastings step is employed. Specifically, sample the cut-point vector $\boldsymbol{g}^m$ by drawing $\boldsymbol{\lambda}^m$ and then using the inverse map as follows: (1) Draw a proposal value $\tilde{\boldsymbol{\lambda}}^m \sim \mathcal{N}(\boldsymbol{\lambda}^m, \delta \boldsymbol{I})$; (2) Accept the new values in $\tilde{\boldsymbol{\lambda}}^m$ with probability $\min\left\{1, \frac{f(\boldsymbol{Y}^{(m)}|\boldsymbol{X}^{(m)}, \tilde{\boldsymbol{\lambda}}^m)\pi(\tilde{\boldsymbol{\lambda}}^m)}{f(\boldsymbol{Y}^{(m)}|\boldsymbol{X}^{(m)}, \boldsymbol{\lambda}^m)\pi(\boldsymbol{\lambda}^m)}\right\}$ such that $f(\boldsymbol{Y}^{(m)}|\boldsymbol{X}^{(m)}, \boldsymbol{\lambda}^m) = \prod_{i=1}^{N}\prod_{j=1}^{P_m}\left[\Phi(g_{y_{ij}^{(m)}} - x_{ij}^{(m)}) - \Phi(g_{y_{ij}^{(m)}-1} - x_{ij}^{(m)})\right]$ and $g_h^m = \sum_{i=1}^{h}\exp(\lambda_i^m)$ for $1 \leq h \leq L_m - 1$.

The MCMC sampling scheme for the proposed multi-view factor model is derived considering a spike-and-slab prior for the view-specific factor loading matrix $\boldsymbol{W}^{(m)}$; that is, the ARD prior is replaced by a group-wise spike-and-slab prior as proposed in Klami et al. (2013). For inference with this prior we use the Gibbs sampling strategy described in Knowles and Ghahramani (2011) with small modifications for the multi-view setting. More details about this implementation could be found in Klami et al. (2013). We note that both the VB scheme and the Gibbs sampler work well in practice and produce similar results. In Section 3, in this supplementary material, we show comparisons between both methods.

2 Computational time

For model fitting via MCMC we ran 50,000 iterations, with a burn-in of 10,000, and then every 5th sample was collected to yield 8,000 posterior samples. We ran two parallel chains starting from different initial values to assess convergence. The convergence for some of the parameters was checked using the Gelman-Rubin diagnostic (Brooks and Gelman, 1998). Given that good mixing was observed, we use these samples to compute posterior estimates. On the hand, the proposed VB algorithm converged (in terms of the variational lower bound) in about 25 iterations; then, in all our experiments we consider a total of 50 iterations for the VB method.

The MCMC and VB algorithms were implemented in Matlab 8.3 and the experiments were performed on a computer with 2.53 GHz Core 2 Duo processor and 4GB memory. To give a sense of computation times, approximately 2 seconds were required per MCMC iteration; that is, approximately 27 hours were required to run 50,000 iterations. For the VB algorithm, approximately 117 seconds were required per iteration; that is, approximately 2 hours were required to run 50 VB iterations.

3 Comparison with MCMC results

We present some comparisons between VB- and MCMC-based results for predicted psychiatric scores and ordinal responses. We also compare the marginal distribution obtained from both inference strategies for some parameters.

3.1 Predicting psychiatric scores

For comparison purposes, we again consider the task of predicting the three psychiatric scores (thought, internalizing and externalizing), but now using the MCMC algorithm to fit the model. These experiments are performed as discussed in the main paper. Table 1 shows the average of the coefficient of determination (R^2) and root mean square error (RMSE) calculated over 10 runs and based on the VB and MCMC algorithms. As we can see, the MCMC- and VB-based results are pretty similar indicating that the predictive performance based on both algorithms are almost the same.

Table 1: Average and standard deviation of the coefficient of determination (R^2) and root mean square error (RMSE) for psychiatric score predictions calculated over 10 runs.

	R^2		RMSE	
	VB	MCMC	VB	MCMC
Thought	0.958 ± 0.022	0.957 ± 0.020	3.121 ± 1.013	3.105 ± 0.950
Internalizing	0.954 ± 0.023	0.955 ± 0.022	3.301 ± 1.123	3.297 ± 1.120
Externalizing	0.956 ± 0.022	0.957 ± 0.022	3.014 ± 0.901	3.011 ± 0.897

3.2 Predicting ordinal responses

In addition to the matrix completion task for fMRI data presented in Section 5.2 (in the main paper), we also consider ordinal matrix completion for questionnaires. For this task, we perform analysis considering the same model hyperparameters as in Section 5. We hide $20\%, 30\%, \ldots, 90\%$ data in the NEO-PI-R ordinal-based view and predict the missing data. Figure 1 shows the average mean absolute error (MAE) for different percentage of missingness over 10 runs, based on predictions from both the proposed VB algorithm and MCMC. As we can see, the VB results are competitive with the MCMC-based predictions.

3.3 Marginal posterior distributions

Figure 2 shows marginal distributions for three parameters in the model, obtained with the VB and the MCMC algorithms. Specifically, we examine the posterior distribution of the view-specific precision parameters γ_m for $m = 1, 5, 15$, associated with the views NEO-PI-R, anger left amygdala, and psychiatric scores, respectively. We note that both inference strategies provide basically the same results. We also checked the posterior distribution from other parameters and the results are also similar, indicating that the VB-based results are comparable with the MCMC sampling scheme that require a large number of iterations to guarantee convergence.

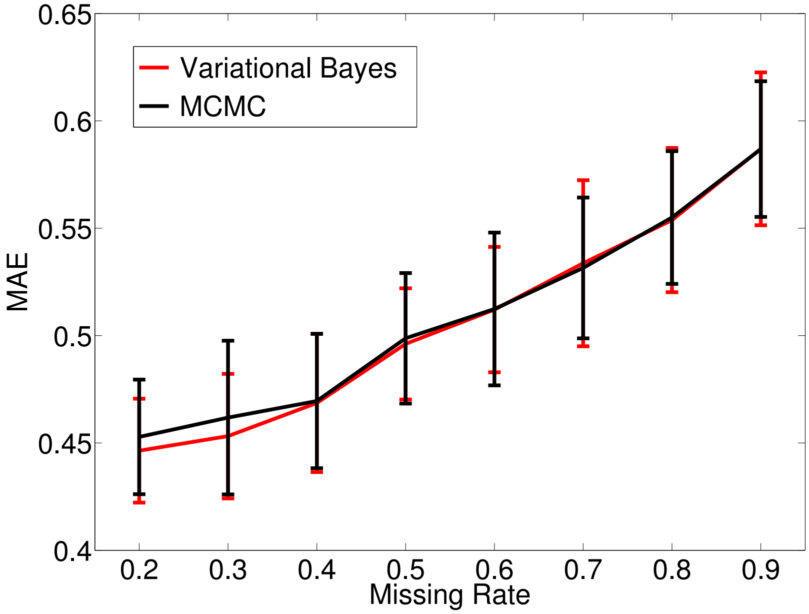
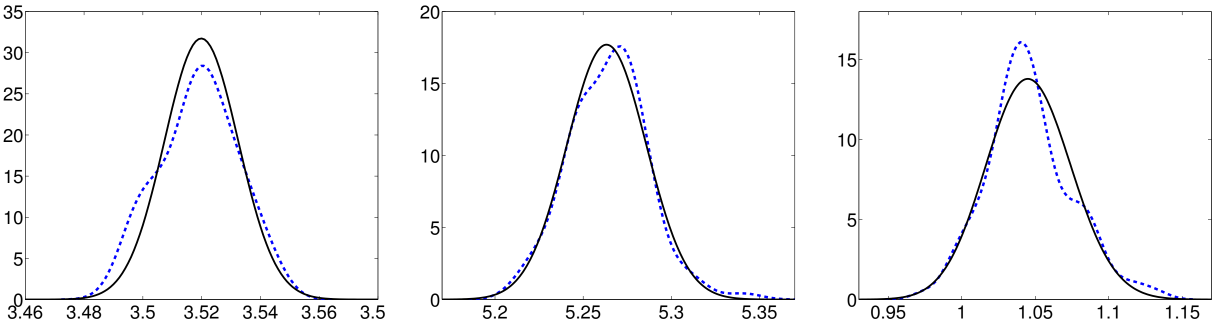


Figure 1: Average mean absolute error (MAE) for the ordinal responses over 10 runs as a functions of the fraction of missing data. Error bars indicate the standard deviation around the mean.



γ_1 - NEO

γ_5 - Anger Left Amyg.

γ_{15} - Psych. Scores

Figure 2: Marginal posterior distributions for three view-specific parameters - γ_m for $m = 1, 5, 15$ - associated with views from NEO, anger left amygdala, and psychiatric scores, respectively. Black lines represent the approximate posterior distributions obtained from the VB algorithm. Dashed blue lines represent the posterior distributions obtained with the MCMC algorithm.

4 Additional results on neuroscience data

In addition to the results presented in Section 5 (in the main paper), here we include results considering additional views to fit the model. Specifically, we now include additional 18 self-report questionnaires (described in Table 2) as part of the ordinal-based views, to fit the model (in our previous analysis, we only considered the NEO-PI-R questionnaire as one of the ordinal-based views). From this new analysis, we could identify relationships between spectrum of behaviors (assessed from some questionnaires), amygdala/VS reactivity and SNP biomarkers; therefore the inclusion of those questionnaires could be of interest to assess the relationships between views.

4.1 Additional self-report questionnaires

In addition to the self-report NEO-PI-R questionnaire, we also have access to additional 18 self-report standard psychological surveys. All participants completed a battery of questions (618 in total) assessing a spectrum of behaviors and experiences related to amygdala function such as negative affect, impulsivity, antisocial personality and anxiety. Table 2 shows a complete list of the behavioral battery considered in new analysis as well as the numbers of questions per questionnaire. Most of them following a 4-, 5- or 6-point Likert scale format (Likert, 1932).

Table 2: Description of the 19 self-report questionnaires used in the neuroscience data analysis.

Abbreviation	Questionnaire Name	Items
NEO PI-R	NEO Personality Inventory Revised	240
STAXI	State Trait Anger Expression Inventory	10
BPAQ	Buss Perry Agression Questionnaire	29
MCSDS	Marlowe-Crowne Social Desirability Scale	33
ERQ	Emotion Regulation Questionnaire	10
BCOPE	Brief COPE Questionnaire	20
LESS	Life Event Scale for Students	46
PSS	Perceived Stress Scale	10
CTQ	Childhood Trauma Questionnaire	28
RQ	Reaction Questionnaire	22
IRI	Interpersonal Relatedness Inventory	7
BIS	Barratt Impulsiveness Scale	30
EDI	Eating Disorders Inventory	25
PSQI	Pittsburgh Sleep Quality Index	13
DEMO	Social Demographic and Family History	49
ISEL	Interpersonal Support Evaluation List	12
ISI	Insomnia Severity Index	7
SPQ	Antisocial Process	22
CSQ	Coping Skills Questionnaire	5

4.2 Results

The new model fitting consider now $M_1 = 22$ ordinal-based views (19 from questionnaires, 3 from SNP data) and $M_2 = 11$ real-based views (from fMRI responses and psychiatric scores). As before, we perform analysis considering an initial value of $K = 50$ latent factors and prior

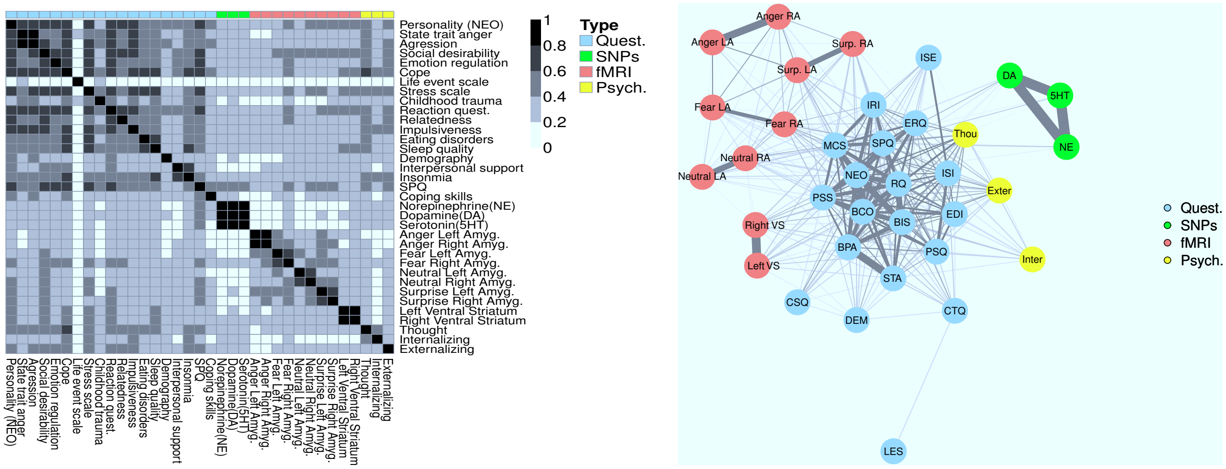


Figure 3: **Left**: Inferred view-correlation matrix considering all the 19 questionnaires. **Right**: Graph representation of the view-correlation matrix (correlations greater than or equal to 0.3 are highlighted) where every node represents a view. Nodes are colored according to the type of data. Nodes representing questionnaires are named using the first three letters of the questionnaire abbreviation (see Table 2).

hyperparameters $a_\alpha = b_\alpha = a_\tau = b_\tau = 0.01$, and use the proposed VB algorithm to fit the model. Figure 3 shows the inferred view-correlation matrix (computed as described in Section 2.3, in the paper) considering a total of $M = 33$ views. A graphical representation of the correlation matrix is shown in the right panel on Figure 3. Note that the view associated with the psychiatric scores was split into thought, internalizing and externalizing disorders. As the figures show, the model discovers high pairwise correlations between different questionnaires as well as between other types of data. For instance, the NEO-PI-R questionnaire is highly correlated with the majority of self-report questionnaires, except for the one related with *Life Event Scale* (LES) evaluation. Also, we note that the LES questionnaire is poorly correlated with the other views (correlations lower than 0.3), except for the one associated with *Childhood Trauma Questionnaire* (CTQ).

Finally, we compare the predictive performance, in terms of psychiatric scores and fMRI predictions, for two scenarios (i) considering all the 19 questionnaires as ordinal-based views and additional available views, and (ii) considering only the NEO-PI-R questionnaire and additional available views; with "additional available views" representing SNP and fMRI data when predicting psychiatric scores, and SNP data when predicting fMRI responses, respectively. Table 3 shows the average and standard deviation of the coefficient of determination (R^2) and root mean square error (RMSE) averaged over 10 runs, for psychiatric scores and fMRI predictions, for each scenario. The R^2 and RMSE values are similar for both scenarios, indicating that the inclusion of more questionnaires does not improve the predictive performance significantly.

Table 3: Average and standard deviation of the coefficient of determination (R^2) and root mean square error (RMSE) for psychiatric score and fMRI predictions calculated over 10 runs.

	R^2		RMSE	
	NEO	All Quest.	NEO	All Quest.
Thought	0.958 ± 0.022	0.959 ± 0.024	3.121 ± 1.013	3.111 ± 1.051
Internalizing	0.954 ± 0.023	0.956 ± 0.024	3.301 ± 1.123	3.295 ± 1.141
Externalizing	0.956 ± 0.022	0.960 ± 0.024	3.014 ± 0.901	3.011 ± 0.920
Anger Left Amygdala	0.940 ± 0.021	0.942 ± 0.022	9.970 ± 1.210	9.950 ± 1.220
Anger Right Amygdala	0.939 ± 0.022	0.940 ± 0.024	9.990 ± 1.210	10.010 ± 1.220
Fear Left Amygdala	0.945 ± 0.023	0.948 ± 0.025	13.070 ± 2.010	12.950 ± 2.070
Fear Right Amygdala	0.954 ± 0.022	0.956 ± 0.022	13.690 ± 2.040	13.650 ± 2.040
Neutral Left Amygdala	0.955 ± 0.019	0.960 ± 0.020	9.390 ± 1.150	9.410 ± 1.160
Neutral Right Amygdala	0.957 ± 0.020	0.961 ± 0.022	9.560 ± 1.300	9.550 ± 1.320
Surprise Left Amygdala	0.941 ± 0.022	0.942 ± 0.025	14.510 ± 3.210	14.500± 3.230
Surprise Right Amygdala	0.945 ± 0.023	0.944 ± 0.023	11.520 ± 3.010	11.540 ± 3.011
Left VS	0.932 ± 0.026	0.936 ± 0.026	5.500 ± 1.010	5.470 ± 1.030
Right VS	0.935 ± 0.025	0.940 ± 0.026	4.870 ± 1.020	4.85 ± 1.025

References

Albert, J. H. and Chib, S. (1997). Bayesian methods for cumulative, sequential and two-step ordinal data regression models. Technical report, Department of Mathematics and Statistics, Bowling Green State University, Ohio.

Albert, J. H. and Chib, S. (2001). Sequential ordinal modeling with applications to survival data. *Biometrics*, 57:829–836.

Brooks, S. P. and Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455.

Klami, A., Virtanen, S., and Kaski, S. (2013). Bayesian Canonical Correlation Analysis. *Journal of Machine Learning Research*, 14:965–1003.

Knowles, D. and Ghahramani, Z. (2011). Nonparametric Bayesian sparse factor models with application to gene expression modeling. *Annals of Applied Statistics*, 5:1534–1552.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55.

Rubin, D. (1987). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm. *Journal of the American Statistical Association*, 82:543–546.