Supplementary material:

Maximum likelihood estimation

In this supplement we derive the maximum likelihood estimation algorithm for the proposed Gaussian factor analysis model. We note that the log-likelihood associated with the proposed model is:

$$\mathcal{L} = \sum_{i=1}^{N} p \log 2\pi + \log \det \Sigma^{(i)} + \operatorname{tr}\left(\Sigma^{(i)^{-1}} K^{(i)}\right). \tag{1}$$

In the case of the loading matrix, the gradient update is defined as:

$$\frac{\partial \mathcal{L}}{\partial W} = \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial \Sigma^{(i)}} \frac{\partial \Sigma^{(i)}}{\partial W} = \sum_{i=1}^{N} \left(\underbrace{-\Sigma^{(i)^{-1}} + \Sigma^{(i)^{-1}} K^{(i)} \Sigma^{(i)^{-1}}}_{M^{(i)}} \right) W G^{(i)} \tag{2}$$

We note that the main computational burden is associated with computing $M^{(i)}$. Using the Sherman-Woodbury identity, we may write $M^{(i)}$ as:

$$\begin{aligned} M^{(i)} = &- v^{(i)^{-1}} I + v^{(i)^{-1}} W A^{(i)} W^T + v^{(i)^{-2}} K^{(i)} \\ &- 2 v^{(i)^{-2}} W A^{(i)} W^T K^{(i)} \\ &+ v^{(i)^{-2}} W A^{(i)} W^T K^{(i)} W A^{(i)} W^T, \end{aligned}$$

from which it follows that computing the gradient of the log-likelihood with respect to the loading matrix, W, incurs a computational cost of $\mathcal{O}(p^3)$.

In the case of the latent connectivity matrix, $G^{(i)}$, the update is equivalent to the score matching algorithm. This follows from:

$$\frac{\partial \mathcal{L}}{\partial G^{(i)}} = \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial \Sigma^{(i)}} \frac{\partial \Sigma^{(i)}}{\partial G^{(i)}} \tag{3}$$

$$= \sum_{i=1}^{N} \left(-\Sigma^{(i)^{-1}} + \Sigma^{(i)^{-1}} K^{(i)} \Sigma^{(i)^{-1}} \right) W^T W. \tag{4}$$

Setting equation (4) to equal zero implies that $I = K^{(i)} \Sigma^{(i)^{-1}}$, which after re-arranging yields:

$$G^{(i)} = W^T K^{(i)} W - v^{(i)} I. \tag{5}$$

Score matching estimation

In this supplement we provide a detailed derivation for the score matching algorithm presented in Section 4. We begin by explicitly writing the score matching objective in terms of parameters W, $\{G^{(i)}\}$ and $\{v^{(i)}\}$. This is specified as:

$$J = \sum_{i=1}^{N} -\text{tr}\left(\Omega^{(i)}\right) + \frac{1}{2}\text{tr}\left(\Omega^{(i)}\Omega^{(s)}K^{(i)}\right) \tag{6}$$

$$= \sum_{i=1}^{N}\Bigg[-{v^{(i)}}^{-1}\text{tr}\left(I\right) + {v^{(i)}}^{-1}\text{tr}(A^{(i)}) + \frac{1}{2}{v^{(i)}}^{-2}\text{tr}(K^{(i)}) \tag{7}$$

$$-{v^{(i)}}^{-2}\text{tr}(W^TK^{(i)}WA^{(i)}) + \frac{1}{2}{v^{(i)}}^{-2}\text{tr}(W^TK^{(i)}WA^{(i)}A^{(i)})\Bigg] \tag{8}$$

where $A^{(i)} = G^{(i)}(G^{(i)}+v^{(i)}I)^{-1}$ as in the original text. We may now directly compute the derivatives with respect to each of the parameters in the proposed latent variable model. The derivative for the loading matrix is:

$$\frac{\partial J}{\partial W} = \sum_{i=1}^{N} {v^{(i)}}^{-2}K^{(i)}W\left(\frac{1}{2}A^{(i)}A^{(i)} - A^{(i)}\right). \tag{9}$$

The derivative with respect to latent connectivities can be obtained via the chain rule as:

$$\frac{\partial J}{\partial G^{(i)}} = \frac{\partial J}{\partial A^{(i)}}\frac{\partial A^{(i)}}{\partial G^{(i)}} \tag{10}$$

$$= {v^{(i)}}^{-2}\left[v^{(i)}I - W^TK^{(i)}W\left(I - A^{(i)}\right)\right]\frac{\partial A^{(i)}}{\partial G^{(i)}} \tag{11}$$

$$= {v^{(i)}}^{-2}\left[v^{(i)}I - W^TK^{(i)}W\left(I - A^{(i)}\right)\right]\left[(G^{(i)} + v^{(i)}I)^{-1}\left(I - A^{(i)}\right)\right]. \tag{12}$$

We note that setting equation (12) to zero implies that the middle term must be zero, as both $(G^{(i)} + v^{(i)}I)^{-1}$ and $(I - A^{(i)})$ cannot be zero. By equating the middle term with zero, we obtain:

$$v^{(i)}\left(W^TK^{(i)}W\right)^{-1} = I - A^{(i)}, \tag{13}$$

which after re-arranging yields:

$$A^{(i)} = G^{(i)}(G^{(i)} + v^{(i)}I)^{-1} = I - v^{(i)}\left(W^TK^{(i)}W\right)^{-1}. \tag{14}$$

Finally, we may re-arranging for $G^{(i)}$ to obtain:

$$G^{(i)} = W^TK^{(i)}W - v^{(i)}I \tag{15}$$

which is the same update as obtained in equation (5) above. Before deriving the derivative of the score matching objective with respect to $v^{(i)}$, we state the following identities which will of use later on:

- We may eigendecompose the latent connectivity $G^{(i)}$ as follows:

$$G^{(i)} = V_iD_iV_i^T, \tag{16}$$

where V_i is a matrix of eigenvectors and D_i is a diagonal matrix of eigenvalues $d_1, \ldots, d_k$. Therefore we may write $A^{(i)}$ as follows:

$$\begin{aligned} A^{(i)} &= G^{(i)}(G^{(i)} + v^{(i)}I)^{-1} \\ &= V_i \; D_i \; V_i^T \; V_i \; \text{diag}\left(\frac{1}{d_j + v^{(i)}}\right) \; V_i^T \\ &= V_i \; \text{diag}\left(\frac{d_j}{d_j + v^{(i)}}\right) \; V_i^T \end{aligned}$$

As a result, we can compute the derivative of $A^{(i)}$ with respect to $v^{(i)}$ as follows:

$$\frac{\partial A^{(i)}}{\partial v^{(i)}} = V_i \; \text{diag}\left(\frac{-d_j}{(d_j + v^{(i)})^2}\right) \; V_i^T = \tilde{D}_i. \tag{17}$$

Furthermore, because V_i are eigenvectors, we have that $\frac{\partial \text{tr}(A^{(i)})}{\partial v^{(i)}} = \text{tr}(\tilde{D}_i)$.

- Using the same arguments, we may write the derivative of $A^{(i)}A^{(i)}$ with respect to $v^{(i)}$ as follows:

$$\frac{\partial A^{(i)}A^{(i)}}{\partial v^{(i)}} = V_i^T \; \text{diag}\left(\frac{-2d_j^2}{(d_j + v^{(i)})^3}\right) \; V_i^T = \tilde{\tilde{D}}_i \tag{18}$$

Using equations (17) and (18) we may therefore write the derivative of the score matching objective with respect to $v^{(i)}$ as follows:

$$\begin{aligned} \frac{\partial J}{\partial v^{(i)}} = \; & {v^{(i)}}^{-3} \text{tr}\left(v^{(i)}I - K^{(i)}\right) && (19) \\ & + {v^{(i)}}^{-3} \text{tr}\left(W^T K^{(i)} W \left[2A^{(i)} - v^{(i)}\tilde{D}_i - A^{(i)}A^{(i)} + v^{(i)}\tilde{\tilde{D}}_i\right]\right) && (20) \\ & - {v^{(i)}}^{-2} \text{tr}\left(A^{(i)} - v^{(i)}\tilde{D}_i\right) && (21) \end{aligned}$$

As such, we define:

$$\begin{aligned} H_1^{(i)} &= 2A^{(i)} - v^{(i)}\tilde{D}_i - A^{(i)}A^{(i)} + v^{(i)}\tilde{\tilde{D}}_i \\ H_2^{(i)} &= -{v^{(i)}}^{-2} \text{tr}\left(A^{(i)} - v^{(i)}\tilde{D}_i\right) \end{aligned}$$

Additional experiments: cluster recovery

Results in Section 6 demonstrated that the proposed method is able to reliably recover the loading matrix, W, in terms of mean squared error. In this section we provide additional results demonstrating that the clusters inferred from the estimated loading matrix accurately reflect the true clustering of variables. The adjusted Rand Index was employed in order to quantify the similarity between the estimated and true clusterings. Results are shown in Figure 1 for Gaussian factor analysis models and latent Bayesian networks along the top and bottom row respectively. In each case, we note that the proposed model is able to accurately cluster variables, inclusively when compared with traditional clustering algorithms such as k-means and hierarchical clustering. Furthermore, we note that as the number of classes increases from $N = 1$ to $N = 10$, the accuracy of the proposed method improves as it is able to combine observations across classes.

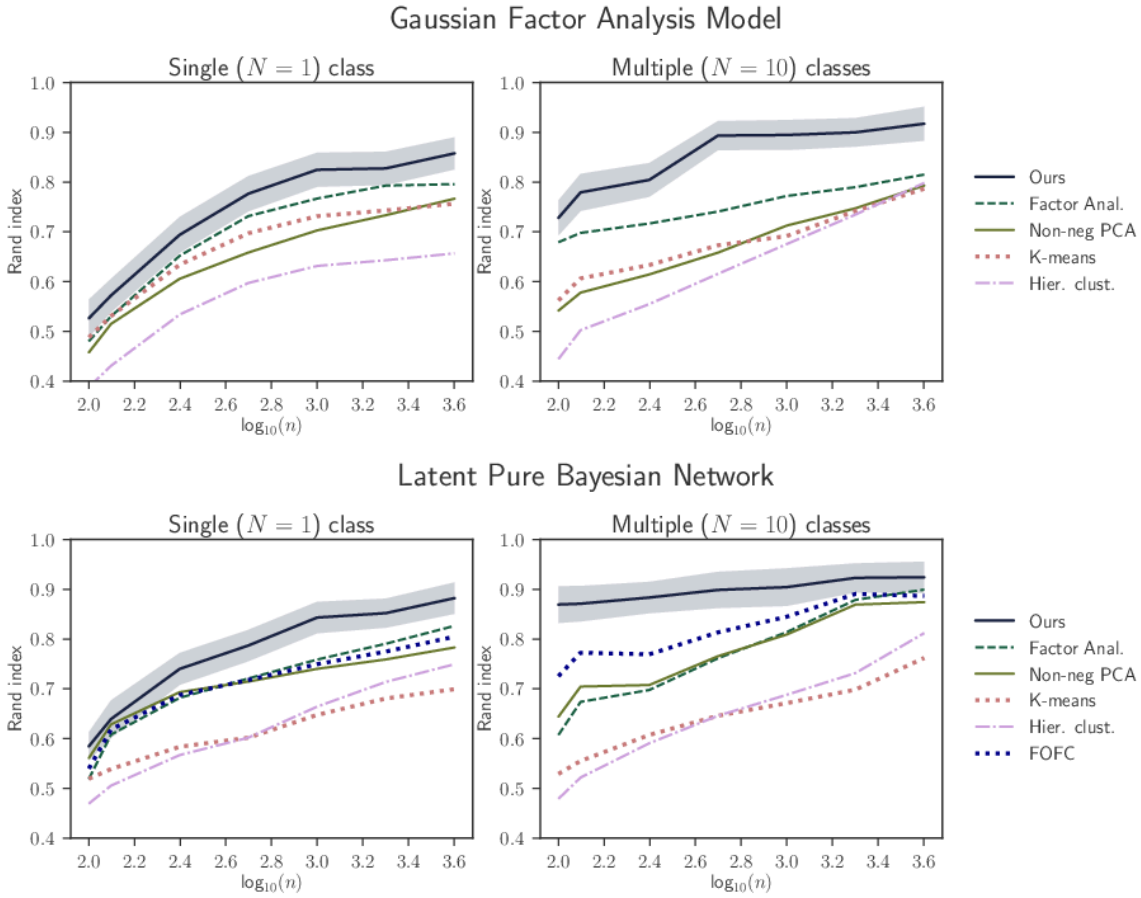

Figure 1: Adjusted Rand index scores for variable clustering as inferred by the estimated loading matrices. Results are shown for Gaussian factor analysis models (top) and latent Bayesian networks (bottom) as well as for $N = 1$ and $N = 10$ classes in the left and right columns respectively. Shaded regions correspond to 95% error bars.