Bayesian information sharing enhances detection of regulatory associations in rare cell types
(Supplementary Materials)

1 Supplementary Figures

Supplementary Figure 1: Noise of network inference algorithms are approximately Gaussian. For a single cell type in each of the BEELINE, Tabula Muris, and mouse blood lineage datasets, we generated 50 sets of bootstrapped cell populations, each of which contain the same number of cells as the original population. We applied Pearson correlation, GENIE3, and PIDC on each of the bootstrapped cell populations and evaluated the distributions of 10,000 randomly selected edges across the 50 bootstrapped samples using the Shapiro-Wilk test of normality. The $-\log_{10}(q)$ values for the selected edges, calculated by applying Bonferonni correction on the test’s $p$ values, are depicted in the histograms. Percentages labeled on top of each histogram correspond to the proportion of edges that are associated with $q$ values greater than 0.05.
Supplementary Figure 2: Standard deviation estimates increase with decreasing sample size. For each cell type considered in the Tabula muris and mouse blood lineage datasets, we identified the top 10% of edges by estimated standard deviation based on five bootstrapped samples. Cell types are ordered according to the number of cells that represent them in the dataset. Cell type population sizes range from 1,939 cells to 103 cells in the Tabula Muris dataset, while the cell type population sizes for the mouse blood lineage dataset range from 1,860 cells to 30 cells.
Supplementary Figure 3: Network accuracy results for ShareNet are robust with respect to the number of mixture components. We compared the accuracy of dendritic cell networks inferred with Pearson correlation, GENIE3, and PIDC, with and without ShareNet. The dendritic cell population is downsampled to 50, 100, and 200 cells. We tested a range of mixture component numbers for ShareNet across the set of downsampling conditions. The inferred dendritic cell networks were compared with the corresponding cell type-specific ChIP-seq networks. While increasing the number of mixture components appears to improve the accuracy of networks inferred with Pearson correlation for population sizes of 50, the effect of ShareNet on network accuracy is generally robust to the number of mixture components used, as long as multiple mixture components are used.
Supplementary Figure 4: Number of bootstrapped samples until convergence of edge score standard deviation estimates. For a single cell type in each of the BEELINE, Tabula Muris, and mouse blood lineage datasets, we generated 50 sets of bootstrapped cell populations, each of which contain only 200 cells. We applied Pearson correlation, GENIE3, and PIDC on each of the bootstrapped cell populations and re-calculated the standard deviation of each edge's score upon the addition of each bootstrapped sample. We evaluated the number of bootstrapped samples required for the standard deviation estimate for each edge score to change by less than 20% and considered the distribution of the number of bootstrapped samples required to reach this convergence threshold. Around 5 bootstrapped samples were required to reach this convergence threshold for 80% of the edges across the various datasets and network inference algorithms.

Supplementary Figure 5: GENIE3 edge scores are highly variable for small populations of cells in the BEELINE dataset. (a) The dispersion (standard deviation divided by mean) of the top 10% of edges ranked by edge score across networks inferred using GENIE3, Pearson correlation, and PIDC for cell population sizes of 50, 100, and 200 cells. (b) The fraction of edges associated with high dispersion (> 0.5) across the various network inference algorithms and cell population sizes.
Supplementary Figure 6: ShareNet improves the accuracy of networks inferred in the presence of small populations of dendritic cells. We calculated AUPRC values for the four cell types that were included alongside the dendritic cells in the downsampling analyses. Networks inferred using ShareNet frequently outperformed the corresponding networks inferred without ShareNet, while also avoiding notable deterioration of accuracy despite sharing information with a downsampled cell type.

Supplementary Figure 7: ShareNet improves the accuracy of networks inferred using the Bayesian Variable Selection (BVS) model. Posterior probabilities of inclusion $p(\gamma_{i,j}^{(c)}|e)$ learned from BVS were used to construct networks for each cell type. $p$-values indicate the statistical significance of ShareNet’s improvement over the original methods calculated using the one-sided Wilcoxon signed-rank test. ShareNet obtains significant improvement over the original methods in all but one comparison (BEELINE).
Supplementary Figure 8: Gene Ontology (GO) enrichment analysis on genes implicated in edges uniquely identified in ShareNet-Pearson networks relative to Pearson networks for MSCs in fat. Genes were ranked by frequency of occurrence in the uniquely identified edges when retaining the top 1% of edges in each network. The TreeMap plot was generated using REVIGO on the significant GO terms (FDR q < 0.01) identified in the enrichment analysis.
2 Derivation of ShareNet

In this section, we derive the parameter updates in the coordinate ascent variational inference procedure used in ShareNet.

2.1 Model Description

We first present the generative process for the Bayesian information sharing framework in full.

\[
(\mu_k, \Sigma_k^{-1}) \sim NW(\mu_0, \beta_0, \Psi, \nu)
\]

\[
u \sim \text{Categorical}(1/K, \ldots, 1/K)
\]

\[
z_{i,j|u_{i,j}} \sim \mathcal{N}(\mu_{u_{i,j}}, \Sigma_{u_{i,j}})
\]

\[
e_{i,j}^{(c)} \sim \mathcal{N}(z_{i,j}^{(c)}, \sigma_{i,j}^{(c)2})
\]

Here, \(e_{i,j}^{(c)}\) represents the observed score for the edge connecting gene \(i\) to gene \(j\) in cell type \(c\), and we model it as a sample from a univariate Gaussian \(\mathcal{N}(e_{i,j}^{(c)}|z_{i,j}^{(c)}, \sigma_{i,j}^{(c)2})\). The mean parameter \(z_{i,j}^{(c)}\) of the Gaussian is a latent variable that represents the true score of the edge if calculated in an ideal scenario absent of noise. The variance parameter \(\sigma_{i,j}^{(c)2}\) captures the degree of variation in the observed edge score. We treat this variance parameter as a hyperparameter, estimating it by bootstrap sampling each cell type in the dataset of interest, applying a chosen network inference algorithm on the bootstrapped samples, and then calculating the standard error, which we use as the standard deviation in this part of the model.

Working up the hierarchical model, the \(C\)-dimensional vector of latent variables \(z_{i,j} = (z_{i,j}^{(1)}, \ldots, z_{i,j}^{(C)})\) containing the true edge scores for edge \((i, j)\) across the \(C\) cell types is a sample drawn from a mixture of multivariate Gaussian distributions with \(K\) mixture components. Each of the \(K\) Gaussians is parameterized by a mean \(\mu_k\) and an inverse covariance \(\Sigma_k^{-1}\) parameter that are drawn independently from the same Normal-Wishart prior. The set of \(K\) covariance matrices \(\Sigma_1, \ldots, \Sigma_K\) represents the set of cell type-to-cell type sharing patterns.

Taking all of these distributions into consideration, the joint density for the full hierarchical model can be written as follows. For simplicity of representation, we denote each (regulator gene \(i\)-target gene \(j\)) combination with a unique subscript \(n\).

\[
p(\mu, \Sigma^{-1}, u, z, e) = \prod_{k=1}^{K} p(\mu_k, \Sigma_k^{-1}) \prod_{n=1}^{N} \prod_{k=1}^{K} p(z_n|\mu_k, \Sigma_k^{-1})u_{n,k} \prod_{n=1}^{N} p(u_n)p(e_n|z_n, D_n^{-1})
\]

Here, \(D_n = \text{diag}(1/\hat{\sigma}_n^{(1)2}, \ldots, 1/\hat{\sigma}_n^{(C)2})\) is the diagonal matrix containing the reciprocal of the squared standard error terms for edge \(n\) across cell types 1 to \(C\).

2.2 Variational Inference

As the latent variable \(z\) represents the true edge scores, we are interested in calculating the posterior distribution of this latent variable given the observed edge scores.

\[
p(z|e, \mu, \Sigma^{-1}, u) = \frac{p(\mu, \Sigma^{-1}, u, z, e)}{p(e)}
\]

We are similarly interested in calculating the posterior distributions of the remaining latent variables \((\mu, \Sigma^{-1}, u)\) given the observed edge scores. However, exact computation of these posterior distributions is intractable, as computing each of these distributions requires computing the evidence term \(p(e)\) in the denominator.

\[
p(e) = \int \int \int \int \prod_{k=1}^{K} p(\mu_k, \Sigma_k^{-1}) \prod_{n=1}^{N} \prod_{k=1}^{K} p(z_n|\mu_k, \Sigma_k^{-1})u_{n,k} \prod_{n=1}^{N} p(u_n)p(e_n|z_n, D_n^{-1})d\mu d\Sigma^{-1}du dz
\]

We can see that computing this evidence term requires integrating over \(u\), which is intractable to compute for even moderate sizes of \(N\) and \(K\). As a result, we resort to variational inference as a tool for approximating these distributions.
2.2.1 Setup

In variational inference, we propose a family of approximating (variational) distributions and then seek to identify the member of this family of distributions that minimizes the KL divergence to the posterior distribution that we aim to approximate \( p(U|x) \).

\[
q^*(U) = \arg \min_{q(U)} \text{KL}(q(U)||p(U|x))
\]

Here, we define \( U = (\mu, \Sigma^{-1}, u, z) \) to be the set of latent variables associated with our model and \( x = e \) are the observed edge scores. Unfortunately, computation of the KL divergence objective function is also intractable because it requires computing \( \log p(e) \).

\[
\text{KL}(q(\mu, \Sigma^{-1}, u, z)||p(\mu, \Sigma^{-1}, u, z, e)) = E_q[\log q(\mu, \Sigma^{-1}, u, z)] - E_q[\log p(\mu, \Sigma^{-1}, u, z, e)]
\]

We can, however, calculate an alternative objective, the ELBO (evidence lower bound), which is equivalent to the negative KL divergence up to an added constant.

\[
\text{ELBO}(q) = E_q[\log p(\mu, \Sigma^{-1}, u, z, e)] - E_q[\log q(\mu, \Sigma^{-1}, u, z)] = E_q[\log p(e|\mu, \Sigma^{-1}, u, z)] - KL(q(\mu, \Sigma^{-1}, u, z)||p(\mu, \Sigma^{-1}, u, z, e))
\]

We, therefore, seek to maximize the ELBO as our objective.

2.2.2 Variational family of distributions

In our setup, we assume that the variational distribution factorizes as

\[
q(U) = q(\mu, \Sigma^{-1}, u, z) = \prod_{k=1}^{K} q(\mu_k, \Sigma_k^{-1}) \prod_{n=1}^{N} q(z_n)q(u_n).
\]

Given this family of distributions, our variational inference goal is to identify the member of this family that maximizes the ELBO.

2.2.3 Coordinate ascent variational inference: general

Let us first consider a general solution for an optimal factor in variational inference. For each factor in the variational distribution, its optimal distribution is proportional to the exponential of the expected log of the joint, where the expectation is taken with respect to the latent variables described by the remaining variational factors (18).

\[
q^*_j(\theta_j) \propto \exp \left( E_{i \neq j}[\log p(\theta_i, x)] \right)
\]

Here, each \( \theta_j \) represents the set of latent variables that are being described by a variational factor \( j \). We can apply this solution to determine the form of each of the factors in our variational distribution.

2.2.4 Coordinate ascent variational inference: \( q(u_n) \)

Starting with the \( q(u_n) \) factor in our variational distribution, we first state the general form of the optimal solution.

\[
q(u_n) \propto \exp \left( E_q(\mu, \Sigma^{-1}, u_{-n}, z, e)[\log p(\mu, \Sigma^{-1}, u, z, e)] \right)
\]
Because our prior over \((\mu, \Sigma^{-1})\) in our hierarchical model is Gaussian, our optimal solution is a Normal-Wishart distribution parameterized as
\[ q \text{.} \]

2.2.5 Coordinate ascent variational inference: \(q(\mu_k, \Sigma_k^{-1})\)

We apply a similar procedure for determining the optimal solution for the \(q(\mu_k, \Sigma_k^{-1})\) factor. The general solution for this factor is as follows.

\[
\log q^*(u_n) = E_q(\mu, \Sigma^{-1}, u_{-n}, z, e) [\log p(\mu, \Sigma^{-1}, u, z, e)] + \text{constant}
\]
\[
= \log p(u_n) + E_q(\mu, \Sigma^{-1}, z) \left[ \sum_{k=1}^{K} u_{n,k} \log p(z_n | \mu_k, \Sigma_k^{-1}) \right] + \text{constant}
\]
\[
= \sum_{n=1}^{N} \sum_{k=1}^{K} u_{n,k} \log \rho_{n,k} + \text{constant},
\]

where we define
\[
\log \rho_{n,k} = \frac{1}{K} + \frac{1}{2} E_q(\Sigma^{-1})[\log |\Sigma_k^{-1}|] - \frac{C}{2} \log(2\pi) - \frac{1}{2} E_q(\mu, \Sigma^{-1}, z) \left[ (z_n - \mu_k)^T \Sigma_k^{-1} (z_n - \mu_k) \right].
\]

If we exponentiate both sides, our optimal solution takes on the form \(q(u_n) \propto \prod_{k=1}^{K} \phi_{n,k}^{u_{n,k}}\), which when normalized is a categorical distribution.

\[
q(u) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \phi_{n,k}^{u_{n,k}}, \quad \text{where} \quad \phi_{n,k} = \frac{\rho_{n,k}}{\sum_{j=1}^{K} \rho_{n,j}}.
\]

From this result, we also have \(E_q(u_n | u_{-n,k}) = \tilde{\phi}_{n,k}\), which will be used in our derivations of the other variational factors below. We will also return to the expectation terms within Equation 1 in the derivations below.

Because our prior over \((\mu_k, \Sigma_k^{-1})\) is a Normal-Wishart distribution and the distribution over each \(z_n\) in our hierarchical model is Gaussian, our optimal solution is a Normal-Wishart distribution parameterized as follows.

\[
q(\mu_k, \Sigma_k^{-1}) = NW(\mu_k, \Sigma_k^{-1} | \tilde{\alpha}_k, \tilde{\beta}_0, \tilde{B}_k, \tilde{\nu}_k)
\]
\[
\tilde{\alpha}_k = \frac{\beta_0 \mu_0 + \sum_{n=1}^{N} \tilde{\phi}_{n,k} E_q(z_n)}{\beta_0 + N_k}
\]
\[
\tilde{B}_k^{-1} = \frac{(\mu_0 - \sum_{n=1}^{N} \tilde{\phi}_{n,k} E_q(z_n)) (\mu_0 - \sum_{n=1}^{N} \tilde{\phi}_{n,k} E_q(z_n))^{T}}{1/\beta_0 + 1/N_k}
\]
\[
+ \tilde{\Psi}^{-1} + \sum_{n=1}^{N} \tilde{\phi}_{n,k} E_q(z_n) \left[ (z_n - \mu_k)(z_n - \mu_k)^T \right]
\]
\[
N_k = \sum_{n=1}^{N} \tilde{\phi}_{n,k}
\]

9
\[ \tilde{v}_k = \nu + \sum_{n=1}^{N} \tilde{\phi}_{n,k} \]

Similar to the above, we will return to the expectation terms expressed in these equations in a subsequent section.

### 2.2.6 Coordinate ascent variational inference: \( q(z_n) \)

\[
\log q^*(z_n) = \sum_{k=1}^{K} \mathbb{E}_{\mu, \Sigma^{-1}, u} [u_{n,k} \log p(z_n | \mu_k, \Sigma_k^{-1})] + \log p(e_n | z_n, D_n^{-1}) + \text{constant}
\]

\[
= \sum_{k=1}^{K} \mathbb{E}_{\mu, \Sigma^{-1}, u} [u_{n,k} \log \mathcal{N}(z_n | \mu_k, \Sigma_k^{-1})] + \log \mathcal{N}(e_n | z_n, D_n^{-1}) + \text{constant}
\]

Because the prior over \( z_n \) in the hierarchical model is Gaussian and the distribution over \( z_n \) is also Gaussian, \( q^*(z_n) \) is taken on the following Gaussian distribution.

\[
q^*(z_n) = \mathcal{N}(m_n, \tilde{S}_n)
\]

\[
\tilde{S}_n = \left( \sum_{k=1}^{K} \tilde{\phi}_{n,k} (\tilde{v}_k B_k + D_n) \right)^{-1}
\]

\[
\tilde{m}_n = \tilde{S}_n \left( \sum_{k=1}^{K} \tilde{\phi}_{n,k} \mu_k (\tilde{v}_k B_k + e_n D_n) \right)
\]

### 2.2.7 Full set of coordinate ascent updates

Using the above expressions for the various factors in our variational distribution, we can write out the complete set of corresponding variational parameter update equations.

For the variational factor \( q(\mu_k, \Sigma_k^{-1}) \), the update equations for \( \tilde{a}_k \) and \( \tilde{B}_k^{-1} \) are as follows.

\[
\tilde{a}_k = \frac{\beta_0 \mu_0}{\beta_0 + N_k} + \sum_{n=1}^{N} \tilde{\phi}_{n,k} \mathbb{E}_{q(z)}[z_n] = \frac{\beta_0 \mu_0}{\beta_0 + N_k} + \sum_{n=1}^{N} \tilde{\phi}_{n,k} m_n
\]

\[
\tilde{B}_k^{-1} = \frac{(\mu_0 - \sum_{n=1}^{N} \tilde{\phi}_{n,k} \mathbb{E}_{q(z)}[z_n]/N_k)(\mu_0 - \sum_{n=1}^{N} \tilde{\phi}_{n,k} \mathbb{E}_{q(z)}[z_n]/N_k)^T}{1/\beta_0 + 1/N_k}
\]

\[
+ \Psi^{-1} + \sum_{n=1}^{N} \sum_{n=1}^{N} \tilde{\phi}_{n,k} \mathbb{E}_{q(z)}[(z_n - \mu_k)(z_n - \mu_k)^T]
\]

\[
= \frac{(\mu_0 - \sum_{n=1}^{N} \tilde{\phi}_{n,k} m_n/N_k)(\mu_0 - \sum_{n=1}^{N} \tilde{\phi}_{n,k} m_n/N_k)^T}{1/\beta_0 + 1/N_k}
\]

\[
+ \Psi^{-1} + \sum_{n=1}^{N} \sum_{n=1}^{N} \tilde{\phi}_{n,k} ((m_n - \tilde{\mu}_k)(m_n - \tilde{\mu}_k)^T + \tilde{S}_n)
\]

For the \( q(u_n) \) factor, we had Equation 1, which contained the following expectation terms that we can now write out.

\[
\mathbb{E}_{q(\Sigma^{-1})}[\log |\Sigma_k^{-1}|] = \sum_{c=1}^{C} \psi \left( \frac{\tilde{v}_c + 1 - c}{2} \right) + C \log 2 + \log |\tilde{B}_k|
\]

\[
\mathbb{E}_{q(\mu, \Sigma^{-1}, z)}[(z_n - \mu_k)^T \Sigma_k^{-1}(z_n - \mu_k)] = (m_n - \tilde{\mu}_k)^T (\tilde{v}_k \tilde{B}_k)(m_n - \tilde{\mu}_k) + \text{Tr}((\tilde{v}_k \tilde{B}_k)\tilde{S}_n) + \mathbb{E}_{q(\Sigma^{-1})}[\text{Tr}(\tilde{\Sigma}^{-1})]
\]
\[(\tilde{m}_n - \tilde{\mu}_k)^T (\tilde{\nu}_k \tilde{B}_k) (\tilde{m}_n - \tilde{\mu}_k) + \text{Tr}((\tilde{\nu}_k \tilde{B}_k) \tilde{S}_n) + C \tilde{\beta}^{-1}_k\]

Using the above two equations, our revised expression for Equation 1 then becomes

\[
\log \rho_{n,k} = \log \frac{1}{K} + \frac{1}{2} \left( \sum_{c=1}^{C} \psi \left( \frac{\tilde{\nu}_k + 1 - c}{2} \right) + C \log 2 + \log |\tilde{B}_k| \right) - \frac{C}{2} \log(2\pi) \\
- \frac{1}{2} \left( (\tilde{m}_n - \tilde{\mu}_k)^T (\tilde{\nu}_k \tilde{B}_k) (\tilde{m}_n - \tilde{\mu}_k) + \text{Tr}((\tilde{\nu}_k \tilde{B}_k) \tilde{S}_n) + C \tilde{\beta}^{-1}_k \right)
\]

We now have a complete set of coordinate ascent update equations for our variational parameters.

\[
\tilde{a}_k = \beta_0 \mu_0 + \sum_{n=1}^{N} \tilde{\phi}_{n,k} \tilde{m}_n
\]

\[
\tilde{B}_k^{-1} = (\mu_0 - \sum_{n=1}^{N} \tilde{\phi}_{n,k} \tilde{m}_n / N_k)(\mu_0 - \sum_{n=1}^{N} \tilde{\phi}_{n,k} \tilde{m}_n / N_k)^T \\
+ \Psi^{-1} + \sum_{n=1}^{N} \tilde{\phi}_{n,k} \left( (\tilde{m}_n - \tilde{\mu}_k)(\tilde{m}_n - \tilde{\mu}_k)^T + \tilde{S}_n \right)
\]

\[
N_k = \sum_{n=1}^{N} \tilde{\phi}_{n,k}
\]

\[
\tilde{\nu}_k = \nu + \sum_{n=1}^{N} \tilde{\phi}_{n,k}
\]

\[
\tilde{\phi}_{n,k} = \frac{\rho_{n,k}}{\sum_{j=1}^{K} \rho_{n,j}}
\]

\[
\log \rho_{n,k} = \log \frac{1}{K} + \frac{1}{2} \left( \sum_{c=1}^{C} \psi \left( \frac{\tilde{\nu}_k + 1 - c}{2} \right) + C \log 2 + \log |\tilde{B}_k| \right) - \frac{C}{2} \log(2\pi) \\
- \frac{1}{2} \left( (\tilde{m}_n - \tilde{\mu}_k)^T (\tilde{\nu}_k \tilde{B}_k) (\tilde{m}_n - \tilde{\mu}_k) + \text{Tr}((\tilde{\nu}_k \tilde{B}_k) \tilde{S}_n) + C \tilde{\beta}^{-1}_k \right)
\]

\[
\tilde{S}_n = \left( \sum_{k=1}^{K} \tilde{\phi}_{n,k} (\tilde{\nu}_k \tilde{B}_k) + D_n \right)^{-1}
\]

\[
\tilde{m}_n = \tilde{S}_n \left( \sum_{k=1}^{K} \tilde{\phi}_{n,k} \tilde{\mu}_k (\tilde{\nu}_k \tilde{B}_k) + e_n D_n \right)
\]
3 Derivation of Bayesian Variable Selection (BVS)

3.1 Model Description

Our derivation of the Bayesian variable selection with ShareNet as prior contains many of the same terms as above, as we seek to leverage a mixture of Gaussians to share information across cell types. As a result, the distributions that are used in this model overlap to some degree with the ones described earlier. In the Bayesian variable selection with ShareNet model, the generative process is as follows.

\[(\mu_k, \Sigma_k^{-1}) \sim \text{NW}(\mu_0, \beta_0, \Psi, \nu)\]
\[u_{i,j} \sim \text{Categorical}(1/K, \ldots, 1/K)\]
\[z_{i,j} | u_{i,j} \sim \mathcal{N}(\mu_{u_{i,j}}, \Sigma_{u_{i,j}})\]
\[\gamma_{i,j}^{(c)} \sim \text{Bernoulli}(\beta_{i,j})\]
\[\beta_{i,j}^{(c)} | \gamma_{i,j}^{(c)} \sim \begin{cases} N(0, \sigma_\beta^2) & \text{if } \gamma_{i,j}^{(c)} = 1, \\ \delta_0(\beta_{i,j}^{(c)}) & \text{otherwise,} \end{cases}\]
\[y_{j,n}^{(c)} \sim \mathcal{N}\left(\sum_i X_{i,n}^{(c)} \beta_{i,j}^{(c)} \gamma_{i,j}^{(c)}, \sigma_\epsilon^2\right)\]

Here, \(\beta_{i,j}^{(c)}\) represents the coefficient of the linear model describing effect of regulator gene \(i\) on target gene \(j\), and \(\gamma_{i,j}^{(c)}\) is a corresponding binary indicator variable that indicates whether or not regulator gene \(i\) is included in the network model for cell type \(c\). We also include “spike and slab” priors (17) on \(\beta_{i,j}^{(c)}\) and \(\gamma_{i,j}^{(c)}\).

Based on our model, we have the following joint density for our Bayesian variable selection problem.

\[
\log p(\mu, \Sigma^{-1}, u, y, \gamma, \beta, z | X) = \sum_{c=1}^{C} \sum_{j=1}^{J} \left\{ -\frac{N(c)}{2} \log(2\pi\sigma_\epsilon) - \frac{y_j^{(c)} \mu_j^{(c)} - y_j^{(c)} \beta_j^{(c)} X_j^{(c)} T X_j^{(c)}}{2\sigma_\epsilon} \right. \\
- \frac{1}{2\sigma_\epsilon^2} \sum_{i \in R(j)} \left( \frac{y_j^{(c)} \mu_j^{(c)} - y_j^{(c)} \beta_j^{(c)} X_j^{(c)} T X_j^{(c)}}{2\sigma_\epsilon^2} \right) X_j^{(c)} T X_j^{(c)} \\
+ \sum_{c=1}^{C} \sum_{j=1}^{J} \sum_{i \in R(j)} (-\log(2\pi\sigma_\beta^2) - \frac{\beta_j^{(c)} \gamma_{i,j}^{(c)} \log(\gamma_{i,j}^{(c)}) + (1 - \gamma_{i,j}^{(c)}) \log(1 - \gamma_{i,j}^{(c)})}{2}) \\
+ \sum_{c=1}^{C} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{n=1}^{N} u_{j,i,k} \left( -\frac{C}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma_k^{-1}| - \frac{1}{2} (z_{j,i} - \mu_k)^T \Sigma_k^{-1} (z_{j,i} - \mu_k) \right) \\
+ \sum_{k=1}^{K} \left( -\frac{C}{2} \log(2\pi) - \frac{1}{2} \log |(\beta_0 \Psi)^{-1}| - \frac{1}{2} (\mu - \mu_0)^T (\beta_0 \Psi)(\mu - m_0) \right) \\
+ \sum_{k=1}^{K} \nu - \frac{C}{2} \log |\Sigma_k^{-1}| - \frac{1}{2} \text{Tr}(\Psi^{-1} \Sigma_k^{-1}) - \log \left( \frac{\nu}{2\pi} \right)^{C/2} \left( \Psi \right)^{\nu/2 \Gamma_C(\nu/2) - 1}}
3.2 Variational Inference

Similar to the previous model, we seek to define a family of variational distributions from which we identify an optimal approximating distribution. In this scenario, the variational distribution factorizes as follows.

\[
q(\gamma_j, \beta_j, z_j, S^{-1}) = \left( \prod_{c=1}^{C} \prod_{i \in R(j)} q(\beta_{j,i}^{(c)}, \gamma_{j,i}^{(c)}) \right) \left( \prod_{i \in R(j)} q(z_{j,i}) \right) \left( q(S^{-1}) \right)
\]

\[
= \left( \prod_{c=1}^{C} \prod_{i \in R(j)} q(\beta_{j,i}^{(c)}, \gamma_{j,i}^{(c)})q(\gamma_{j,i}^{(c)}) \right) \left( \prod_{i \in R(j)} q(z_{j,i}) \right) \left( q(S^{-1}) \right)
\]

We also define the distributions for the following set of variational factors.

\[
q(\beta_{j,i}^{(c)}, \gamma_{j,i}^{(c)}) = \begin{cases} 
N(\beta_{j,i}^{(c)}, \gamma_{j,i}^{(c)}, \mu_{j,i}^{(c)}, \pi_{j,i}^{(c)}) & \text{if } \gamma_{j,i}^{(c)} = 1 \\
\delta_0(\beta_{j,i}^{(c)}) & \text{otherwise}
\end{cases}
\]

\[
q(\gamma_{j,i}^{(c)}) = (\alpha_{j,i}^{(c)})^{\gamma_{j,i}^{(c)}} (1 - \alpha_{j,i}^{(c)})^{1 - \gamma_{j,i}^{(c)}}
\]

\[
q(z_{j,i}) = N(z_{j,i}|\tilde{m}_{j,i}, \tilde{S}_{j,i})
\]

3.2.1 ELBO

Given our joint density and the general form of our variational distributions, we can write out the ELBO.

\[
\text{ELBO}(q) = \mathcal{L}(q) = \mathbb{E}_q(\gamma_j, \beta_j, z_j, S^{-1})[\log \mathbb{P}(X_j, \gamma_j, \beta_j | X_{-j}, z_j, S^{-1}, \theta) - \log q(\gamma_j, \beta_j, z_j, S^{-1})]
\]

\[
= \mathbb{E}_q(\gamma_j, \beta_j, z_j, S^{-1})[\log \mathbb{P}(X_j, \gamma_j, \beta_j, z_j, S^{-1} | X_{-j}, \theta)] - \mathbb{E}_q(\gamma_j, \beta_j, z_j, S^{-1})[\log q(\gamma_j, \beta_j, z_j, S^{-1})]
\]

We split up the first and second terms in the ELBO and consider them separately at first for simplicity.

\[
\mathbb{E}_q \left[ \log p(y_j, \gamma_j, \beta_j, z_j, S^{-1} | X_j) \right] = \sum_{c=1}^{C} \left\{ \sum_{i \in R(j)} \left( -\frac{N(c)}{2} \log(2\pi \sigma_i) - \frac{y_j^{(c)T} y_j^{(c)}}{2\sigma_i^{(c)}} + \frac{1}{2\sigma_i^{(c)}} \sum_{i \in R(j) \neq i} \mathbb{E}_q \left[ (\gamma_{j,i}^{(c)} \beta_{j,i}^{(c)}) (\gamma_{j,i}^{(c)} \beta_{j,i}^{(c)}) \right] x_i^{(c)T} x_i^{(c)} \right) - \frac{1}{2\sigma_i^{(c)}} \sum_{i \in R(j) \neq i} \mathbb{E}_q \left[ (\gamma_{j,i}^{(c)} \beta_{j,i}^{(c)})^2 \right] x_i^{(c)T} x_i^{(c)} \right) \right.
\]

\[
+ \sum_{c=1}^{C} \left( \sum_{i \in R(j)} \left( -\frac{\log(2\pi \sigma_i^2)}{2} - \mathbb{E}_q \left[ \frac{\beta_{j,i}^{(c)^2} \sigma_i^2}{\sigma_i^2} \right] \right) \right)
\]

\[
+ \sum_{c=1}^{C} \mathbb{E}_q \left[ \gamma_{j,i}^{(c)} \log(g(z_{j,i}^{(c)})) + (1 - \gamma_{j,i}^{(c)}) \log(1 - g(z_{j,i}^{(c)})) \right]
\]

\[
+ \sum_{i \in R(j)} \left( -\frac{1}{2} \mathbb{E}_q \left[ \log(2\pi |S|) \right] - \frac{1}{2} \mathbb{E}_q \left[ (z_{j,i} - m)^T S^{-1} (z_{j,i} - m) \right] \right) - \log(2^{np/2}V^{n/2} \Gamma_p(n/2)) + \frac{n - p - 1}{2} \mathbb{E}_q \left[ \log |S^{-1}| \right] - \frac{1}{2} \mathbb{E}_q \left[ \text{Tr}(V^{-1} S^{-1}) \right]
\]

\[
\mathbb{E}_q[\log q(\gamma_j, \beta_j, z_j, S^{-1})] = \left( \sum_{c=1}^{C} \sum_{i \in R(j)} \mathbb{E}_q \left[ \log q(\beta_{j,i}^{(c)} | \gamma_{j,i}^{(c)}) \right] + \mathbb{E}_q \left[ \log q(\gamma_{j,i}^{(c)}) \right] \right) + \left( \sum_{i \in R(j)} \mathbb{E}_q \left[ \log q(z_{j,i}) \right] \right)
\]
3.2.2 Expectation Terms

From the above expression of the ELBO, we observe that there are a number of expectation terms which we will need to calculate in order to calculate the ELBO. We write out this set of expectation equations, which will be useful for explicitly writing out the form of the ELBO.

\[
\mathbb{E}_q[\gamma_{j,i}^{(c)}] = \alpha_{j,i}^{(c)}
\]

\[
\mathbb{E}_q[\gamma_{j,i}^{(c)} \beta_{j,i}^{(c)}] = \sum_{\gamma_{j,i}^{(c)}} \int_{\beta_{j,i}^{(c)}} \gamma_{j,i}^{(c)} \beta_{j,i}^{(c)} q(\beta_{j,i}^{(c)} | \gamma_{j,i}^{(c)}) q(\gamma_{j,i}^{(c)}) d\beta_{j,i}^{(c)}
\]

\[
= \int_{\beta_{j,i}^{(c)}} \beta_{j,i}^{(c)} \alpha_{j,i}^{(c)} \mathcal{N}(\beta_{j,i}^{(c)} | \mu_{j,i}^{(c)}, \sigma_{j,i}^{(c)}^2) d\beta_{j,i}^{(c)} = \alpha_{j,i}^{(c)} \mu_{j,i}^{(c)}
\]

\[
\mathbb{E}_q[\gamma_{j,i}^{(c)} \beta_{j,i}^{(c)} \gamma_{j,i}^{(c)}] = \sum_{\gamma_{j,i}^{(c)}} \int_{\beta_{j,i}^{(c)}} \gamma_{j,i}^{(c)} \beta_{j,i}^{(c)} \gamma_{j,i}^{(c)} q(\beta_{j,i}^{(c)} | \gamma_{j,i}^{(c)}) q(\gamma_{j,i}^{(c)}) d\beta_{j,i}^{(c)}
\]

\[
= \int_{\beta_{j,i}^{(c)}} \beta_{j,i}^{(c)} \alpha_{j,i}^{(c)} \mathcal{N}(\beta_{j,i}^{(c)} | \mu_{j,i}^{(c)}, \sigma_{j,i}^{(c)}^2) d\beta_{j,i}^{(c)} = \alpha_{j,i}^{(c)} (\sigma_{j,i}^{(c)}^2 + \mu_{j,i}^{(c)}^2)
\]

\[
\mathbb{E}_q[\log q(\gamma_{j,i}^{(c)}, \beta_{j,i}^{(c)})] = \sum_{\gamma_{j,i}^{(c)}} \int_{\beta_{j,i}^{(c)}} \log q(\gamma_{j,i}^{(c)}, \beta_{j,i}^{(c)}) q(\beta_{j,i}^{(c)} | \gamma_{j,i}^{(c)}) q(\gamma_{j,i}^{(c)}) d\beta_{j,i}^{(c)}
\]

\[
= \alpha_{j,i}^{(c)} \log(\alpha_{j,i}^{(c)}) + (1 - \alpha_{j,i}^{(c)}) \log(1 - \alpha_{j,i}^{(c)}) - \frac{1}{2} \left( 1 + \log(2\pi \sigma_{j,i}^{(c)}^2) \right)
\]

\[
\mathbb{E}_q[\log q(z_{j,i})] = \int_{z_{j,i}} \left( \log q(z_{j,i}) \right) q(z_{j,i}) dz_{j,i} = -\frac{1}{2} \log |\bar{S}| - \frac{D}{2} (1 + \log(2\pi))
\]

\[
\mathbb{E}_q[\log p(z_{j,i} | m, S)] = \mathbb{E}_q \left[ -\frac{1}{2} \log |S| - \frac{1}{2} (z_{j,i} - m)^T S^{-1} (z_{j,i} - m) \right]
\]

\[
= -\frac{1}{2} \log |S| - \frac{1}{2} \mathbb{E}_q \left[ (z_{j,i} - m)^T S^{-1} (z_{j,i} - m) \right]
\]

\[
= -\frac{1}{2} \log |S| - \frac{1}{2} \text{Tr} \left( S^{-1} \mathbb{E}_q \left[ (z_{j,i} - m)(z_{j,i} - m)^T \right] \right)
\]

\[
= -\frac{1}{2} \log |S| - \frac{1}{2} \text{Tr} \left( S^{-1} \mathbb{E}_q \left[ z_{j,i} z_{j,i}^T - z_{j,i} m^T - m z_{j,i}^T + m m^T \right] \right)
\]

\[
= -\frac{1}{2} \log |S| - \frac{1}{2} \text{Tr} \left( S^{-1} \mathbb{E}_q \left[ (m_{j,i} - \bar{m}_{j,i}) + (\tilde{S}_{j,i} - \bar{S}_{j,i}) \bar{m}_{j,i} - \tilde{m}_{j,i} m^T - m \bar{m}_{j,i}^T + m m^T \right] \right)
\]

\[
= -\frac{1}{2} \log |S| - \frac{1}{2} \text{Tr} \left( S^{-1} (m_{j,i} - m) + \text{Tr}(S^{-1} \tilde{S}_{j,i}) \right)
\]

\[
\mathbb{E}_q[(z_{j,i} - m)^T S^{-1} (z_{j,i} - m)] = (\bar{m}_{j,i} - m)^T S^{-1} (\tilde{m}_{j,i} - m) + \text{Tr}(S^{-1} \tilde{S}_{j,i})
\]

\[
\mathbb{E}_q[\log p(S^{-1})] = \mathbb{E}_q \left[ -\log(2^{np/2} |V|^{n/2} \Gamma_p(n/2)) + \frac{n - p - 1}{2} \log |S^{-1}| - \frac{1}{2} \text{Tr}(V^{-1} S^{-1}) \right]
\]

\[
= -\log(2^{np/2} |V|^{n/2} \Gamma_p(n/2)) + \frac{n - p - 1}{2} \log |S^{-1}| - \frac{1}{2} \text{Tr}(V^{-1} S^{-1})
\]
3.2.3 Approximating the Expectation of Log Sigmoid

Here, we seek to approximate $E_q[\log g(z_j^{(c)})]$ by first defining a grid of samples from the factor in our variational distribution that approximates the distribution for $z_{j,i}$. In our current setup, this factor corresponds to a multivariate Gaussian centered at $\tilde{m}_{j,i}$ with covariance $\tilde{S}_{j,i}$. Because we are focusing on approximating a single scalar value $E_q[\log g(z_j^{(c)})]$ here, we need only to consider the components of the multivariate Gaussian that directly correspond to element $(c)$, namely $\tilde{m}^{(c)}_{j,i}$ in the mean and $\tilde{S}^{c,c}_{j,i}$ in the covariance. We refer to these values as $\tilde{m}$ and $\tilde{s}$, respectively, below for simplicity.

We choose a grid of samples $\tilde{m}, \tilde{m} \pm \eta \sqrt{s}, \tilde{m} \pm 2\eta \sqrt{s}, \ldots, \tilde{m} \pm K\eta \sqrt{s}$, where $\eta$ and $K$ are hyperparameters. Using this grid of samples, we approximate our expectation term as

$$E_q[\log g(z_j^{(c)})] \approx \sum_{t=-K}^{K} w_t \log(g(\tilde{m} + t\eta \sqrt{s})),$$

where $w_t = \Phi(t + \frac{1}{2}) - \Phi(t - \frac{1}{2})$.

Similarly, we approximate $E_q[\log(1 - g(z_j^{(c)}))]$ as $E_q[\log(1 - g(z_j^{(c)}))] \approx \sum_{t=-K}^{K} w_t \log(1 - g(\tilde{m} + t\eta \sqrt{s}))$.

With these approximations, let us now write out the ELBO.

$$L(q) = \sum_{c=1}^{C} \left\{ - \frac{N^{(c)}}{2} \log(2\pi \sigma_\epsilon) - \frac{y_j^{(c)T} y_j^{(c)}}{2\sigma_\epsilon^2} + \frac{1}{\sigma_\epsilon^2} \sum_{i \in R(j)} \alpha_{j,i}^{(c)} \mu_{j,i}^{(c)} x_i^{(c)T} x_j^{(c)} \right\}$$

$$- \frac{1}{2\sigma_\epsilon^2} \sum_{i \in R(j)} \sum_{i' \neq i} (\alpha_{j,i}'^{(c)} \mu_{j,i}'^{(c)})(\alpha_{j,i}^{(c)} \mu_{j,i}^{(c)}) x_i^{(c)T} x_j^{(c)}$$

$$- \frac{1}{2\sigma_\epsilon^2} \sum_{i \in R(j)} \alpha_{j,i}^{(c)} \mu_{j,i}^{(c)} x_i^{(c)T} x_i^{(c)}$$

$$+ \sum_{c=1}^{C} \sum_{i \in R(j)} \alpha_{j,i}^{(c)} \left( - \frac{\log(2\pi \sigma_\beta^2)}{2} - \frac{s_{j,i}^{(c)^2} + \mu_{j,i}^{(c)^2}}{2\sigma_\beta^2} \right) - \frac{1}{2} \left( - \psi_p(n/2) - p \log(2) - \log |S^{-1}| \right) - \frac{D}{2} \log(2\pi) - \frac{1}{2} \left( \tilde{m}_{j,i} - m \right)^T S^{-1} (\tilde{m}_{j,i} - m) + \text{Tr}(S^{-1} \tilde{S}_{j,i})$$

$$+ \sum_{c=1}^{C} \sum_{i \in R(j)} - \alpha_{j,i}^{(c)} \log(\alpha_{j,i}^{(c)}) - (1 - \alpha_{j,i}^{(c)}) \log(1 - \alpha_{j,i}^{(c)}) + \frac{1}{2} \left( 1 + \log(2\pi s_{j,i}^{(c)^2}) \right)$$

$$+ \sum_{i \in R(j)} \frac{1}{2} \log |\tilde{S}_{j,i}| + \frac{D}{2} (1 + \log(2\pi))$$

$$- \log(2^{np/2} |V|^{n/2} \Gamma_p(n/2)) + \frac{n - p - 1}{2} \log |S^{-1}| - \frac{1}{2} \text{Tr}(V^{-1} S^{-1})$$
3.2.4 Coordinate ascent variational inference: $\tilde{S}_{j,i}$

Let us first write out the terms in the ELBO that contain $\tilde{S}_{j,i}$ or any of its elements.

\[
\mathcal{L}(q) = k + \sum_{c=1}^{C} \sum_{i \in R(j)} \alpha_{j,i}^{(c)} \left( \sum_{t=-K}^{K} w_{i} \log \left( g \left( \tilde{m}_{j,i}^{(c)} + t\eta \sqrt{\tilde{S}_{j,i}^{c,c}} \right) \right) \right) + (1 - \alpha_{j,i}^{(c)}) \left( \sum_{t=-K}^{K} w_{i} \log \left( 1 - g \left( \tilde{m}_{j,i}^{(c)} + t\eta \sqrt{\tilde{S}_{j,i}^{c,c}} \right) \right) \right) \\
+ \sum_{i \in R(j)} \frac{1}{2} \log |\tilde{S}_{j,i}| - \frac{D}{2} \log(2\pi) - \frac{1}{2} \left( (\tilde{m}_{j,i} - \mathbf{m})^T \mathbf{S}^{-1}(\tilde{m}_{j,i} - \mathbf{m}) + \text{Tr}(\mathbf{S}^{-1}\tilde{S}_{j,i}) \right)
\]

Now, we take the partial derivative with respect to $\tilde{S}_{j,i}$.

\[
\frac{\partial \mathcal{L}}{\partial \tilde{S}_{j,i}} = \frac{\partial}{\partial \tilde{S}_{j,i}} \left( \alpha_{j,i}^{(c)} \left( \sum_{t=-K}^{K} w_{i} \log \left( g \left( \tilde{m}_{j,i}^{(c)} + t\eta \sqrt{\tilde{S}_{j,i}^{c,c}} \right) \right) \right) + (1 - \alpha_{j,i}^{(c)}) \left( \sum_{t=-K}^{K} w_{i} \log \left( 1 - g \left( \tilde{m}_{j,i}^{(c)} + t\eta \sqrt{\tilde{S}_{j,i}^{c,c}} \right) \right) \right) \right) \\
- \frac{1}{2} \mathbf{S}^{-1} + \frac{1}{2} \tilde{S}_{j,i}^{-1}
\]

We simplify each of the terms in the first line to clean up the calculation of the partial derivative, replacing $\tilde{S}_{j,i}^{c,c}$ with $s$ and making the following substitutions: $a = e^{-\tilde{m}_{j,i}^{(c)}}$ and $k_{t} = \eta t$. We also apply the sigmoid function $g(x) = \frac{1}{1+e^{-x}}$.

\[
\frac{\partial}{\partial s} \left( \log \left( 1 + e^{-a(s) + t\eta \sqrt{s}} \right) \right) = \frac{\partial}{\partial s} \left( \log \left( 1 + \frac{e^{-a(s) + t\eta \sqrt{s}}}{e^{-a(s)}} \right) \right) = \frac{\partial}{\partial s} \left( \log \left( 1 + ae^{-ks} \right) \right)
\]

\[
= k \frac{2\sqrt{s}}{1 + ae^{-ks}}
\]

\[
= \frac{k}{2\sqrt{s}} \left( 1 - \frac{1}{1 + ae^{-ks}} \right) = \frac{k}{2\sqrt{s}} \left( 1 - \frac{1}{1 + e^{-(\tilde{m}_{j,i}^{(c)} + \eta \sqrt{s})}} \right) = \frac{\eta t}{2\sqrt{s}} \left( 1 - g(\tilde{m}_{j,i}^{(c)} + \eta \sqrt{s}) \right)
\]

\[
\frac{\partial}{\partial s} \left( \log \left( \frac{1}{1 + e^{-a(s) + t\eta \sqrt{s}}} \right) \right) = \frac{\partial}{\partial s} \left( \log \left( 1 - \frac{1}{1 + e^{-a(s) + t\eta \sqrt{s}}} \right) \right) = \frac{\partial}{\partial s} \left( \log(1 - \frac{1}{1 + ae^{-ks}}) \right)
\]

\[
= -\frac{k}{2\sqrt{s}} \frac{1}{1 + ae^{-ks}} = -\frac{\eta t}{2\sqrt{s}} \frac{1}{1 + e^{-(\tilde{m}_{j,i}^{(c)} + \eta \sqrt{s})}} = -\frac{\eta t}{2\sqrt{s}} g(\tilde{m}_{j,i}^{(c)} + \eta \sqrt{s})
\]

Substituting these terms back into our expression for $\frac{\partial \mathcal{L}}{\partial \tilde{S}_{j,i}}$ yields the following.

\[
\frac{\partial \mathcal{L}}{\partial \tilde{S}_{j,i}} = -\frac{1}{2} \mathbf{S}^{-1} + \frac{1}{2} \tilde{S}_{j,i}^{-1} + \mathbf{D}
\]

Here, $\mathbf{D} = \text{diag}(d^{(1)}, ..., d^{(C)})$, where

\[
d^{(c)} = \sum_{t=-K}^{K} \left( \alpha_{j,i}^{(c)} w_{i} \frac{\eta t}{2\sqrt{s}} \left( 1 - \frac{1}{1 + e^{-(\tilde{m}_{j,i}^{(c)} + \eta s)}} \right) + (1 - \alpha_{j,i}^{(c)}) w_{i} \frac{\eta t}{2\sqrt{s}} \left( \frac{1}{1 + e^{-(\tilde{m}_{j,i}^{(c)} + \eta s)}} \right) \right)
\]

\[
= \sum_{t=-K}^{K} \left( \alpha_{j,i}^{(c)} w_{i} \frac{\eta t}{2\sqrt{s}} \left( 1 - g(\tilde{m}_{j,i}^{(c)} + \eta \sqrt{s}) \right) - (1 - \alpha_{j,i}^{(c)}) w_{i} \frac{\eta t}{2\sqrt{s}} g(\tilde{m}_{j,i}^{(c)} + \eta \sqrt{s}) \right)
\]

\[
= \sum_{t=-K}^{K} w_{i} \frac{\eta t}{2\sqrt{s}} \left( \alpha_{j,i}^{(c)} \left( 1 - g(\tilde{m}_{j,i}^{(c)} + \eta \sqrt{s}) \right) - (1 - \alpha_{j,i}^{(c)}) \left( g(\tilde{m}_{j,i}^{(c)} + \eta \sqrt{s}) \right) \right)
\]

\[
= \sum_{t=-K}^{K} w_{i} \frac{\eta t}{2\sqrt{s}} \left( \alpha_{j,i}^{(c)} - (1 - \alpha_{j,i}^{(c)}) \left( g(\tilde{m}_{j,i}^{(c)} + \eta s) - g(\tilde{m}_{j,i}^{(c)} + \eta \sqrt{s}) + \alpha_{j,i}^{(c)} g(\tilde{m}_{j,i}^{(c)} + \eta \sqrt{s}) \right) \right)
\]

16
We then calculate the remaining partial derivative terms in the equation above. Similar to before, we start out by isolating the terms in the ELBO that contain \( \tilde{L} \).

Taking the partial derivative with respect to \( \tilde{L} \), we get the following.

Taking the partial derivative with respect to \( \tilde{m}_{j,i} \), we get the following.

Our expression for the partial derivative can now be written as follows.
3.2.6 Coordinate ascent variational inference: $s_{j,i}^{[c]}$

\[
\frac{\partial \mathcal{L}}{\partial s_{j,i}^{[c]}} = -\frac{1}{2\sigma^2} \alpha_{j,i}^{(c)} (X_i^{(c)\top} X_i^{(c)}) - \frac{1}{2\sigma^2} + \frac{1}{2s_{j,i}} = 0
\]

\[
\frac{1}{s_{j,i}^{[c]^2}} = \frac{1}{\sigma^2} (X_i^{(c)\top} X_i^{(c)}) + \frac{1}{\sigma^2} \alpha_{j,i}^{(c)} (X_i^{(c)\top} X_i^{(c)}) + \frac{\sigma^2}{\sigma^2}
\]

3.2.7 Coordinate ascent variational inference: $\mu_{j,i}^{(c)}$

\[
\frac{\partial \mathcal{L}}{\partial \mu_{j,i}^{(c)}} = \frac{1}{\sigma^2} \alpha_{j,i}^{(c)} (X_i^{(c)\top} y^{(c)}) - \frac{1}{\sigma^2} (X_i^{(c)\top} X_i^{(c)}) \alpha_{j,i}^{(c)} \mu_{j,i}^{(c)} - \frac{1}{\sigma^2} \sum_{i' \neq i} (X_i^{(c)\top} X_{i'}^{(c)}) \alpha_{j,i'}^{(c)} \mu_{j,i'}^{(c)} - \frac{1}{\sigma^2} \mu_{j,i}^{(c)} - \frac{1}{s_{j,i}^{[c]^2}} = \mu_{j,i}^{(c)} \left( \frac{1}{\sigma^2} + \frac{1}{s_{j,i}^{[c]^2}} \right)
\]

\[
\mu_{j,i}^{(c)} = \frac{\alpha_{j,i}^{(c)} s_{j,i}^{[c]^2}}{\sigma^2} (X_i^{(c)\top} y^{(c)}) - \sum_{i' \neq i} \alpha_{j,i'}^{(c)} \mu_{j,i'}^{(c)} (X_i^{(c)\top} X_{i'}^{(c)})
\]

3.2.8 Coordinate ascent variational inference: $\alpha_{j,i}^{(c)}$

\[
\frac{\partial \mathcal{L}}{\partial \alpha_{j,i}^{(c)}} = \frac{1}{\sigma^2} \mu_{j,i}^{(c)} (X_i^{(c)\top} y^{(c)}) - \frac{1}{\sigma^2} \sum_{i' \neq i} \alpha_{j,i'}^{(c)} \mu_{j,i'}^{(c)} (X_i^{(c)\top} X_{i'}^{(c)})
\]

\[
- \frac{1}{\sigma^2} (s_{j,i}^{[c]^2} + \mu_{j,i}^{(c)^2}) X_i^{(c)\top} X_i^{(c)}
\]

\[
+ \left( \sum_{t=-K}^K w_t \log \left( g \left( \tilde{m}_{j,i}^{(c)} + \eta \sqrt{S_{j,i}^{c,c}} \right) \right) \right) - \left( \sum_{t=-K}^K w_t \log \left( 1 - g \left( \tilde{m}_{j,i}^{(c)} + \eta \sqrt{S_{j,i}^{c,c}} \right) \right) \right)
\]

\[
- (1 + \log(\alpha_{j,i}^{(c)})) + (log(1 - \alpha_{j,i}^{(c)})) + 1
\]

\[
u_{j,i} = (1 + \log(\alpha_{j,i}^{(c)})) + (log(1 - \alpha_{j,i}^{(c)})) + 1
\]

\[
u_{j,i} = \frac{1 - \alpha_{j,i}^{(c)}}{\alpha_{j,i}^{(c)}} = 0
\]

\[
\Rightarrow \alpha_{j,i}^{(c)} = \frac{1}{1 + e^{-u_{j,i}^{(c)}}}
\]

\[
u_{j,i} = \frac{1}{\sigma^2} \mu_{j,i}^{(c)} (X_i^{(c)\top} y^{(c)}) - \frac{1}{\sigma^2} \sum_{i' \neq i} \alpha_{j,i'}^{(c)} \mu_{j,i'}^{(c)} (X_i^{(c)\top} X_{i'}^{(c)})
\]
\[ -\frac{1}{2\sigma^2} (s_j^{(c)} + \mu_{j,i}^{(c)}) X_i^{(c)^T} X_i^{(c)} + b + 2 \sum_{c' \in N(c)} W_{c,c'} (2\alpha_{j,i}^{(c')} - 1) \]

\[ = \frac{\mu_{j,i}^{(c)}}{\sigma^2} \left( X_i^{(c)} y^{(c)} - \sum_{i' \neq i} \alpha_{j,i'}^{(c)} \mu_{j,i'}^{(c)} X_i^{(c)^T} X_i^{(c)} \right) - \frac{1}{2\sigma^2} (s_j^{(c)} + \mu_{j,i}^{(c)}) X_i^{(c)^T} X_i^{(c)} \]

\[ + \left( \sum_{t=-K}^{K} w_t \log \left( g \left( \tilde{m}_{j,i}^{(c)} + t\eta \sqrt{\tilde{S}_{j,i}^{(c)}} \right) \right) \right) - \left( \sum_{t=-K}^{K} w_t \log \left( 1 - g \left( \tilde{m}_{j,i}^{(c)} + t\eta \sqrt{\tilde{S}_{j,i}^{(c)}} \right) \right) \right) \]
4 Runtime and Memory Usage

ShareNet’s runtime and memory usage are dependent on a variety of factors, most notably the number of cell types and the number of putative edges in a network. The main bottleneck for both of these computational resources is related to the $\tilde{S}$ variational parameter, an $N \times C \times C$ matrix, where $N$ is the number of putative edges in a network and $C$ is the number of cell types. The memory associated with this parameter, therefore, scales with complexity $O(NC^2)$. Meanwhile, the matrix inversion step in the coordinate ascent update for $\tilde{S}$ (Eq. 8) scales with complexity $O(NC^3)$. As a result, the number of network edges and cell types considered in the use of ShareNet can dramatically influence both the runtime and memory consumption, a phenomenon we observe across the three datasets, which spanned a broad range of these factors.

For the BEELINE analyses, we considered 267,750 putative edges for each of the five cell types in the dataset. In the Tabula muris analyses, we inferred networks for 52 cell types, each of which contained 2,640,456 putative edges. Lastly, the mouse blood lineage dataset contained 24 cell types, and each network was associated with 1,207,152 putative edges. The average total runtime for ShareNet was 2 minutes, 116 minutes, and 411 minutes for the BEELINE, mouse blood lineage, and Tabula muris datasets, respectively. Peak memory usage was approximately 664 MB, 11.9 GB, and 111 GB for the same three datasets. These runtime and memory usage statistics do not include runs of the network inference algorithms required to obtain initial network estimates and standard deviation estimates, which ShareNet uses as input. All methods were run on a 2.40GHz Intel Xeon E5-2695v2 central processing unit.

While the runtime and memory usage statistics may appear sizeable for larger datasets, we note that both of these computational resources can be substantially reduced by leveraging prior biological knowledge. For example, the number of putative edges considered in a network can be drastically reduced to consider only edges supported by transcription factor binding motifs or chromatin accessibility instead of all possible transcription factor-target gene pairs as was done in this study. In addition, we observed that cell types that are dissimilar tend to engage in less information sharing with each other in ShareNet (Fig. 5a, 6b). Retaining only cell types that are more similar to each other for use in ShareNet rather than using an entire collection of varied cell types could, therefore, also reduce the computational burden for such large dataset scenarios without diminishing performance.