Recall the full data model is

$$Y_{p\times n} = B_{p\times d}X_{d\times n} + \underbrace{L_{p\times K}C_{K\times n}}_{\text{Cell Type Effect}} + \underbrace{\Gamma_{p\times r}H_{r\times n}}_{\text{Additional Confounding}} + \underbrace{E_{p\times n}}_{MN_{p\times n}\left(0,\Sigma=\text{diag}\left(\sigma_1^2,\ldots,\sigma_p^2\right),I_n\right)}$$

$$C_{K\times n} = \Omega_{K\times d}X_{d\times n} + D\Xi_{K\times n}, \text{ where } \text{Var}\left(D\Xi_{K\times n}\right) = \left(DD^T\right)\otimes I_n = \Lambda\otimes I_n$$

$$H_{r\times n} = \alpha_{r\times d}X_{d\times n} + W_{r\times n}$$

The α term is included for completeness. If we understand our problem, we would expect this to be 0. We need to account for ΓH, since this term is almost always present in experimental data. In our case, we have individuals with and without cell type training data:

$$Y_1 = BX_1 + LC + \Gamma\alpha X_1 + \Gamma W_1 + E_1$$
$$Y_2 = BX_2 + LD\Xi + L\Omega X_2 + \Gamma\alpha X_2 + \Gamma W_2 + E_2$$

Estimating L, Γ, Σ

We use REML by rotating out X_1 and X_2 from the above two equations to estimate L, Γ, Σ. Given r, I have a working method that estimates L, Γ, Σ. In practice, however, we need to estimate r, which is challenging. For this reason, n_1 cannot be too small.

Correcting For Cell Type

We first rotate out C from the first equation:

$$Y_1Q_C = \tilde{Y}_1 = B\underbrace{\tilde{X}_1}_{X_1Q_C} + \Gamma\alpha\tilde{X}_1 + \Gamma\underbrace{\tilde{W}_1}_{W_1Q_C} + \underbrace{\tilde{E}_1}_{E_1Q_C}$$

and then compute the QR decomposition of $\tilde{X}_1 = \tilde{R}_1^T\tilde{Q}_1^T$:

$$\tilde{Y}_1\tilde{Q}_1\left[,1:d\right] = \tilde{Z}_1 = B\tilde{R}_1^T + \Gamma\left(\alpha\tilde{R}_1^T + \tilde{W}_1^{(1)}\right) + \tilde{E}_1^{(1)}.$$

Note that for $\tilde{R}_1$ small (i.e. cell type explains a large portion of the variability in X_1), we need to account for spurious correlations due to $\tilde{W}_1^{(1)}$ (even if $\alpha = 0$). For the next set of individuals, we have

$$Y_2Q_2\left[,1:d\right] = \tilde{Z}_2 = BR_2^T + \Gamma\left(\alpha R_2^T + W_2^{(1)}\right) + L\left(\Omega R_2^T + D\Xi^{(1)}\right) + E_2^{(1)}.$$

For $\Lambda\otimes\left(R_2^{-1}R_2^{-T}\right) = \Lambda\otimes\left(X_2X_2^T\right)^{-1}$ small in comparison to Ω, we can ignore the spurious correlation caused by $\Xi^{(1)}$. This basically means that if the cellular variability Λ is LESS than the correlation between C and X (like it is in the cases we are considering), then we can ignore spurious correlations, even for small sample sizes in X_2. We can check this by comparing the estimates for D and ΩR_2^T. If we know Ω and D, then we can approximate the distribution for $\Omega + D\Xi^{(1)}R_2^{-T}$. If we were to then estimate $\Omega + D\Xi^{(1)}R_2^{-T}$ directly from the data and found that the estimate strayed too far from Ω, we would know that the data are not informative enough to estimate cell type.

This idea is important because in some cases we are betting off estimating the confounding directly from the data. I am still working on a method to determine when we should use training data, and the variance of $D\Xi^{(1)}R_2^{-T}$ in relation to Ω will be an important factor.

As of now, I have a method that seems to control the false discovery rate at a nominal level, even for large confounding and noisy estimates of L. There are two scenarios we should consider

1. $\Lambda\otimes\left(X_2X_2^T\right)^{-1}$ is small. In this case, we can estimate $\Omega + D\Xi^{(1)}R_2^{-T}$ from the data and see how far the estimate is from Ω. If it is close, we know that we can probably estimate the confounding from the data. If not, then we need to use training data. This may be useful in experimental design. If one was involved in a large study (i.e. COPSAC), we could get cell type data on only a fraction of the individuals, estimate L, Ω and $\Lambda = DD^T$. If we find that our estimate for $\Omega + D\Xi^{(1)}R^{-T}$ is close to Ω, then we know the data are informative enough to estimate cell type. If not, then we should collect enough training data to apply my method.

2. $\Lambda\otimes\left(X_2X_2^T\right)^{-1}$ is large. In this case, you should use the first set of samples with cell type information to estimate L and see if we can reliably estimate cell type composition from the data. If we can, then we can ignore cell type and estimate it from the data.

Alternative Assumptions on L

In the above section, I assumed that L was full rank, i.e. methylation differed across all K cell types. It may be the case that only a subset of the cell types show different methylation patters, or we can explain the variability in methylation caused by cell type with a low dimensional projection. If we only consider the first set of individuals and ignore BX_1, we can estimate L over the space of rank $s \leq K$ matrices using GLS:

$$\min_{L \text{ rank } s} \left\|\Psi^{-1/2} Y_1 - \Psi^{-1/2} L C_1\right\|_F^2 \underbrace{=}_{\Psi^{-1/2} L \Lambda^{1/2} = \tilde{L}, \tilde{C}_1 = \Lambda^{-1/2} C_1, \tilde{Y}_1 = \Psi^{-1/2} Y_1} \min_{\tilde{L}} \left\|\tilde{Y}_1 - \tilde{L}\tilde{C}_1\right\|_F^2 = \min_{\tilde{\ell} \in \mathbb{R}^{p \times s}, \tilde{u} \in \mathbb{R}^{K \times s}, A} \left\|\tilde{Y}_1 - \tilde{\ell} A \tilde{u}^T \tilde{C}_1\right\|_F^2$$

where $\Psi = \Sigma + \Gamma\Gamma^T$, $\Lambda = \frac{1}{n_1} C_1 C_1^T$ (i.e. $\tilde{C}_1 \tilde{C}_1^T = I_K$), A is a diagonal matrix with s entries all greater than 0 and $\tilde{\ell}^T \tilde{\ell} = \tilde{u}^T \tilde{u} = I_s$ (i.e. $\tilde{\ell} A \tilde{u}^T$ is the SVD of $\tilde{L}$). Expanding the objective function, we see that

$$\arg\min_{\tilde{\ell}, \tilde{u}, A} \left\|\tilde{Y}_1 - \tilde{\ell} A \tilde{u}^T \tilde{C}_1\right\|_F^2 = \operatorname{Tr}\left(\tilde{C}_1^T \tilde{u} A \tilde{\ell}^T \tilde{\ell} A \tilde{u}^T \tilde{C}_1\right) - 2 \operatorname{Tr}\left(\tilde{Y}_1^T \tilde{\ell} A \tilde{u}^T \tilde{C}_1\right) = \operatorname{Tr}\left(A^2\right) - 2 \operatorname{Tr}\left(\tilde{Y}_1^T \tilde{\ell} A \tilde{u}^T \tilde{C}_1\right)$$

We see the above objective function has a minimum when $\tilde{u}^T \tilde{C}_1 \tilde{Y}_1 \tilde{\ell} = A$, meaning $\tilde{\ell}$ and $\tilde{u}$ are the first s right and left singular vectors of $\tilde{C}_1 \tilde{Y}_1$, respectively. This is also the solution to the usual CCA problem! We need not worry about the high dimensional nature of the problem, since it is assumed that Γ is low rank.

Note that if Γ were known, $\operatorname{Var}\left(\hat{\tilde{u}}\right) = O\left(\frac{1}{pn_1}\right)$. If we estimate $\hat{\Gamma}$ where each row has variance $O\left(\frac{1}{n_1}\right)$, then the singular vectors of $\tilde{C}_1 Y_1^T \hat{\Psi}^{-1} Y_1 \tilde{C}_1^T \in \mathbb{R}^{K \times K}$ should have variance roughly $O\left(\frac{1}{n_1}\right)$. Therefore, with enough training data, we can get a good starting point for $\hat{\tilde{u}}$ and use the rest of the data to refine that estimate (a good starting point is essential, since the likelihood that includes the second set of individuals is non-convex). This seems to work well in simulations. This idea may be important if we believe some cell types have very similar methylation patterns.