Suppose we have methylation data for p CpG sites on the logit scale (i.e. logit(beta value)) for two sets of individuals, Y_1, Y_2, where $Y_i \in \mathbb{R}^{p\times n_i}$ and n_i the number of individuals in each group. Let $X_i \in \mathbb{R}^{d\times n_i}$ be the covariate matrix for d covariates of interest (i.e. sex, asthma status, etc.) and $C_i \in [0,1]^{K\times n_i}$ be the cell type matrix for $K+1$ cell types ($K+1$ since X_i has an intercept). If $c_s^{(1)}$ is the s^{th} column of C_1, then $\mathbf{1}^T c_s^{(1)} < 1$. The data are then modeled linearly as

$$Y_i = B_{p\times d} X_i + L_{p\times K} C_i + E_i,\ E_i \sim MN_{P\times n_i}\left(0, \Sigma_{p\times p}, I_{n_i}\right)$$

In our set up, we observed X_1, X_2 and C_1 but NOT C_2. For now, we assume that the columns of C_1 follow a dirichlet distribution with

$$C_1 \mid X_1, \Omega \sim \text{Dir}\left(\alpha \Omega X_1^T\right)$$

where $\Omega \in \mathbb{R}^{K\times d}$ and $\alpha > 0$ is a concentration parameter. Note that

$$E\left(C_1 \mid X_1, \Omega\right) = \Omega X_1^T$$

$$\text{Var}\left(c_s^{(1)} \mid X_1, \Omega\right) = \frac{1}{\alpha+1}\left(\text{diag}\left(\Omega x_{1_s}\right) - \Omega x_{1_s}\left(\Omega x_{1_s}\right)^T\right)$$

and

$$E\left(Y_1 \mid X_1, X_2, C_1\right) = BX_1 + LC_1$$

$$E\left(Y_2 \mid X_1, X_2, C_1\right) = BX_2 + L\Omega X_2$$

0.1 Estimating Ω

Let $l\left(\Omega, \alpha \mid C_1\right) = \log p\left(C_1 \mid \Omega, \alpha\right)$ and let $\hat{\Omega}, \hat{\alpha} = \hat{\Omega}\left(C_1\right), \hat{\alpha}\left(C_1\right)$ be the MLE. In order to estimate B and L I condition on C_1, which meaning there is no randomness in $\hat{\Omega}, \hat{\alpha}$ in the usual frequentist setting. In order to introduce uncertainty back into the estimates $\hat{\Omega}, \hat{\alpha}$, I assume there is a prior $p\left(\Omega, \alpha\right)$ with compact support on Ω, α. If we define $h = \left(\text{vec}\left(\Omega\right), \alpha\right)$, we have

$$\log p\left(\Omega, \alpha \mid C_1\right) = Const + \underbrace{l\left(\Omega, \alpha \mid C_1\right)}_{O_P\left(n_1^{1/2}\right)} + \underbrace{\log p\left(h\right)}_{O_p(1)} = Const + l\left(\hat{h} \mid C_1\right) - \underbrace{\frac{1}{2}\left(h-\hat{h}\right)^T H\left(\hat{h}\right)\left(h-\hat{h}\right)}_{\text{Kernel of a } N\left(\hat{h}, H\left(\hat{h}\right)^{-1}\right)} + o_P\left(\left\|h-\hat{h}\right\|^2\right) + \log p\left(h\right)$$

where $H = -\nabla_{hh}^2 l \mid_{\hat{h}} \approx \mathcal{I}\left(\hat{h}\right)$, the Fisher information at $\hat{h}$. Since $\frac{1}{2}\left(h-\hat{h}\right)^T H\left(\hat{h}\right)\left(h-\hat{h}\right)$ dominates the above expression for large n_1, we may approximate $p\left(h \mid C_1\right)$ with a normal distribution:

$$p\left(h \mid C_1\right) \approx N\left(\hat{h}, \mathcal{I}\left(\hat{h}\right)^{-1}\right).$$

This is exactly the Bayesian central limit theorem. So long as $n_1 = O\left(n_2\right)$, the uncertainty in h is asymptotically negligible. However, the above result will be important in proving we need $n_1 = O\left(n_2\right)$ to justify using $\hat{h} = \hat{h}\left(C_1\right)$ as a plugin estimator for h.

0.2 Quasi-Likelihood Estimator

Since the Normal + Dirichlet is a difficult likelihood to work with, I use a quasi-likelihood estimator. For each CpG site $g = 1, \ldots, p$, let y_{i_g} be the g^{th} row of Y_1, ℓ_g the g^{th} column of L and β_g the g^{th} column of B. Let $\tilde{\beta}_g = \begin{pmatrix}\beta_g \\ \ell_g\end{pmatrix}$. The model for these individual sites is

$$y_{1_g} = X_1^T \beta_g + C_1^T \ell_g + \epsilon_{1_g},\ \epsilon_{1_g} \sim N\left(0, \sigma_g^2\right)$$

$$y_{2_g} = X_2^T \beta_g + C_2^T \ell_g + \epsilon_{2_g},\ \epsilon_{2_g} \sim N\left(0, \sigma_g^2\right)$$

where ϵ_{i_g} is independent of C_2. The **means** of the random variables are

$$\mu_{1_g} = \mu_{1_g}\left(\beta_g, \ell_g\right) = E\left(y_{1_g} \mid X_1, X_2, C_1\right) = X_1^T \beta_g + C_1^T \ell_g$$

$$\mu_{2_g} = \mu_{2_g}\left(\beta_g, \ell_g\right) = E\left(y_{2_g} \mid X_1, X_2, C_1\right) = X_2^T\beta_g + \underbrace{E\left(C_2 \mid X_2, C_1\right)^T}_{\approx X_2^T\hat{\Omega}^T} \ell_g.$$

The **variances** are

$$\mathrm{Var}\left(y_{1_g} \mid X_1, X_2, C_1\right) = \sigma_g^2 I_{n_1}$$

$$\mathrm{Var}\left(y_{2_g} \mid X_1, X_2, C_1\right) = \mathrm{Var}\left(C_2^T\ell_g \mid X_2, X_1, C_1\right) + \sigma_g^2 I_{n_2} = G\left(\ell_g\right).$$

Let c_s and c_t be the s and t^{th} columns of C_2, respectively. Define

$$R_s = \frac{1}{\alpha+1}\left(\mathrm{diag}\left(\Omega x_{2_s}\right) - \Omega x_{2_s}\left(\Omega x_{2_s}\right)^T\right) \text{ and } V_s = E_{\Omega,\alpha|C_1,X_1,X_2} R_s.$$

$$S_{ss} = \underbrace{\mathrm{Var}_{\Omega|C_1,X_1,X_2}\left(E\left(c_s \mid C_1, X_1, X_2, \Omega\right)\right)}_{\text{Variance taken over } p\left(h \mid C_1, X_1, X_2\right)}$$

$$S_{st} = \mathrm{Var}_{\Omega|C_1,X_1,X_2}\left(E\left(c_s \mid C_1, X_1, X_2, \Omega\right), E\left(c_t \mid C_1, X_1, X_2, \Omega\right)\right)$$

0.3 Review of Current Methods

To my knowledge there are three current methods to correct for cell type in methylation experiments, 2 by Houseman and 1 by Zhou.

1. **Houseman, 2012**: Houseman uses a training set consisting of methylation measured in cell types sorted by flow cytometry. The user can then feed in their methylation data matrix (on the beta scale) and get back the predicted cell types for each individuals. The problems with this method are:

   (a) Predicted cell type depends on population. Michelle compared her measured cell proportions with Houseman's predicted values and found that the flow cytometry data was very different that Houseman's predictions.

   (b) Houseman does not provide any sort of uncertainty in the prediction, only a point estimate. This is important in assessing the uncertainty in our parameter estimates.

   (c) Their model does not use covariates as information in deconvolving cell type. If two people have similar covariates, then they should also have similar cell types.

2. **Houseman, 2014**: Houseman uses an unsupervised method to estimating regression coefficients in the presence of unmeasured cell type confounders. The shortcomings of this method are:

   (a) One is forced to make restrictive assumptions about the sparsity of B and $L\Omega$. In an independent paper, Hastie 2016 shows that in order to get consistent estimators of B in the unsupervised regime, one needs to make the assumption that the sparsity of B INCREASES as the sparsity of L INCREASES. That is, if the unobserved cell type does not explain a large portion of the variability in methylation (which we have seen happen, e.g. Michelles Hutterite data), then one has no hope of estimating B. In fact, whenever cell type DOES NOT depend on your covariates you drastically reduce your power to do inference on B when you do not observe cell type.

   (b) Even when $L\Omega$ is dense and B is sparse, Houseman uses a poor estimator for $\hat{B}$. The asymptotic properties of his estimator are not known and it is unrealistic to get p-values for individual sites. The best we can do is get site-specific confidence intervals. We can actually improve upon this method by using a different estimator.

3. **Zhou, 2014**: Zhou uses a linear regression treating methylation as the covariate and disease status as the response. He first forms that individual relatedness matrix Φ using the centered methylation matrix as a proxy for individual relatedness. He then iteratively adds principle components of Φ until only a small fraction of sites are significantly correlated with the phenotype. The problems here are:

   (a) The method is very adhoc and the statistical properties are not very well understood.

   (b) You can only look at one covariate at a time. You cannot test the hypothesis that methylation is uncorrelated with ≥ 2 covariates at once.

   (c) You lose power when performing OLS with binary response (in comparison to a GLM).