CovBat using Multilevel PCA

Andrew Chen

Site Effect in Between-Site Principal Component Scores

Original CovBat harmonizes the principal component scores across all sites, where the PCs and scores are both determined using the full dataset. In practice, this method works quite well, but there are some potential drawbacks. One problem is that variance in the largest site may dominate the PCA, leaving important site effects in smaller sites relegated to lower eigenvalue PCs. Another is that correcting on PCs determined using the whole sample may remove some wanted variation within subjects of a single site, which may negatively impact downstream analyses.

In this document, we propose an alternative method that leverages multilevel PCA to identify important directions of variation separately for both between-site and between-subjects variation. We base our multilevel PCA off multilevel FPCA (Di et al. 2009), which has been used in the analogous case of repeated measurements within a single subject. Our proposed method would isolate between-site variation and then correct on the associated PC scores, leaving within-site variation untouched. This method may potentially enable more targeted harmonization and mitigate the loss of wanted heterogeneity.

Let y_{ij}, $i = 1, \dots, M$, $j = 1, \dots, n_i$ be the $p \times 1$ observation vectors. Before performing PCA, we want to remove the site effect in the mean and variance of each variable. To do this, we turn to ComBat. In this context, ComBat assumes that the observations follow

$$y_{ij} = \alpha_{ij} + x'_{ij}\beta + \gamma_i + \delta'_i e_{ij}$$

where α_{ij} is the intercept vector, x'_{ij} are the covariates, β is the vector of regression coefficients, γ_i is the $p \times 1$ mean site effect vector, δ_i is the $p \times 1$ variance site effect vector, and e_{ij} is the error vector where $e_{ij} \sim N(\mathbf{0}, \sigma' I)$. σ is the vector of variances, consistent across site. ComBat then estimates γ_i and δ_i then residualizes with respect to those parameters, yielding ComBat-adjusted observations y_{ij}^{ComBat}. We additionally residualize on the intercept and covariates, so now $y_{ij} \sim N(0, \sigma' I)$.

To remove potential covariance site effects, we now assume that y_{ij}^{ComBat} have mean 0 and covariances Σ_i. We hope to recover observations without the covariance site effect such that the true observations should have mean 0 and covariance Σ.

Multilevel PCA splits the covariance of Y into between-site and within-site covariances then performs separate PCA analyses on each. For a D-dimensional dataset this is expressed via

$$\Sigma = \sum_{k=1}^{D} Var(\Lambda_k^{(1)})\phi_k^{(1)}{\phi_k^{(1)}}^T + \sum_{l=1}^{D} Var(\Lambda_l^{(2)})\phi_l^{(2)}{\phi_l^{(2)}}^T$$

We can approximate the observations using multilevel PCA as $y_{ij}^{ComBat} \approx \sum_{k=1}^{K} \lambda_{ik}^{(1)}\phi_k^{(1)} + \sum_{l=1}^{L} \lambda_{ijl}^{(2)}\phi_l^{(2)}$ where K and L are chosen to capture some portion of the variation in the observations. We assume that the site effect is captured by the site-level principal component scores $\lambda_{ik}^{(1)}$ for each site indexed by i. For maximal correction, we could simply harmonize across sites by setting the score for each site equal to the mean score that is

$${\lambda_{ik}^{(1)}}^{mCovBat} = \frac{1}{M}\sum_{j}^{M} \lambda_{jk}^{(1)}$$

for all k. Alternatively, we could harmonize some subset $\mathscr{K}$ of the between-site scores.

Based on multilevel PC assumptions (more details in later section), this would drive the variance of the between-site scores included in $\mathscr{K}$ to be zero, leaving

$$Cov(Y^{mCovBat}) = \sum_{k \notin \mathscr{K}} \widehat{Var}(\Lambda_k^{(1)})\phi_k^{(1)}{\phi_k^{(1)}}^T + \sum_{l=1}^{D} \widehat{Var}(\Lambda_l^{(2)})\phi_l^{(2)}{\phi_l^{(2)}}^T$$

where each $\widehat{Var}(\Lambda_k^{(1)})$ are estimated from the $\lambda_{ik}^{(1)}$ across i and $\widehat{Var}(\Lambda_l^{(2)})$ are estimated from the $\lambda_{ijl}^{(2)}$ across i and j. This harmonized covariance arises from the fact that mCovBat drives $\widehat{Var}(\Lambda_k^{(1)}) = 0$ for $k \in \mathscr{K}$.

Intriguingly, this harmonization would not affect site-specific covariance matrices. There is some question then of whether or not this method would remove site covariance effects. Furthermore, whether or not this aggressive harmonization technique would remove wanted site heterogeneity warrants further discussion.

To implement this method, we would need to separate the within-site and between-site covariances in a manner similar to how Di et al. 2009 estimates "total", "between" and "within" covariances. We use method of moments along with the fact that the total covariance $K_T = K_W + K_B$ where K_W is the within-site covariance and K_B is the between-site covariance. Adapting the paper from the functional data setting to the vector setting, we write our estimators in terms of the whole sample covariance matrix $\hat{\Sigma}$, the sample site covariance matrices $\hat{\Sigma}_i$, and the total number of subjects where $N = \sum_{i=1}^{M} n_i$ as $\hat{K}_W = \frac{1}{N}\sum_{i=1}^{M} n_i \hat{\Sigma}_i$ and $\hat{K}_T = \hat{\Sigma}$ then define the estimator $\hat{K}_B = \hat{K}_T - \hat{K}_W$

Multilevel PC Method Properties

From multilevel PCA assumptions, we have that for datasets Y_i comprised of draws from site-specific distributions with mean zero and covariances Σ_i that $Y_{ij} \approx \sum_{k=1}^{P} \lambda_{ik}^{(1)} \phi_k^{(1)} + \sum_{l=1}^{S} \Lambda_l^{(2)} \phi_l^{(2)}$ where P and S are less than the dimension of the original dataset, $\lambda_{ik}^{(1)}$ is a draw from the between-site PC score distribution $\Lambda_k^{(1)}$, and $\Lambda_l^{(2)}$ are uncorrelated random variables with mean 0. $\Lambda_k^{(1)}$ and $\Lambda_l^{(2)}$ may be correlated. Then we can derive the covariance of Y_i as

$$\begin{aligned} Cov(Y_i) &\approx Cov\left(\sum_{k=1}^{P} \lambda_{ik}^{(1)} \phi_k^{(1)} + \sum_{l=1}^{S} \Lambda_l^{(2)} \phi_l^{(2)}\right) \\ &= \sum_{l=1}^{S} Var(\Lambda_{il}^{(2)}) \phi_l^{(2)} {\phi_l^{(2)}}^T \end{aligned}$$

We also know from the multilevel PCA assumptions that the full data covariance is

$$Cov(Y) = \sum_{k=1}^{D} Var(\Lambda_k^{(1)}) \phi_k^{(1)} {\phi_k^{(1)}}^T + \sum_{l=1}^{D} Var(\Lambda_l^{(2)}) \phi_l^{(2)} {\phi_l^{(2)}}^T$$

The variances of within-site and between-site score variances can be estimated via method of moments where there are M within-site scores and $N = \sum_{i=1}^{M} n_i$ between-site scores.