ComBat with Covariance Effect Estimation Notes

Andrew Chen

Site Effect in Principal Component Scores

A heuristic approach to this problem proceeds as follows. Perform PCA on each site's observations separately. Ideally, the PCs identified would be shared between sites. In this case, the difference in covariance between sites may be captured in differences in the distribution of PC scores. By regressing on site-specific center and scaling parameters, perhaps we could remove the site effect in variance and covariance.

This approach corresponds to a model that essentially works on the spectral decomposition of each site's covariance matrix. 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 $y_{ij} \sim N(0, \sigma' I)$. An argument could be made for not including the scaling parameter in the model since subsequent PCA steps may be able to remove the scaling effect. However, should a variable be scaled overly large due to site effect, it may dominate the PCA analysis for that site. So to be safe, we remove the scaling effect via ComBat.

To remove potential covariance site effects, we now assume that y_{ij}^{ComBat} have mean 0 and covariances Σ_i. We now look at the spectral decompositions for each covariance matrix $\Sigma_i = \Phi_i \Lambda_i \Phi'_i$ where Φ_i are the $p \times p$ matrices with the PCs as the columns and Λ_i are the diagonal eigenvalue matrices. However, we hope to recover observations without the covariance site effect such that the true observations should have mean 0 and covariance $\Sigma = \Phi \Lambda \Phi'$.

We assume that $\Sigma_i \approx \Phi \Lambda_i \Phi'$. To estimate Φ there are several methods that could be employed. The simplest would be to just take the full data matrix and perform PCA to obtain the eigenvectors in Φ. However, there is no guarantee that those eigenvectors best capture the variation in the data independent of site effect. MetaPCA provides an alternative means to determine eigenvectors that may more optimally describe the data across sites. Another estimation procedure can be taken from the PVD paper. This procedure would take some number of eigenvectors from each site, arrange them into a separate matrix, then perform PCA on that matrix to obtain the major directions of variation among the eigenvectors. Hope to discuss these options further.

Once Φ is estimated, we can approximate the observations $y_{ij}^{ComBat} \approx \sum_{k=1}^{K} \lambda_{ijk}\phi_k$ where K is chosen to capture some portion of the variation in the observations and $\Phi = \begin{bmatrix}\phi_1 & \cdots & \phi_p\end{bmatrix}$. We assume that the site effect lies within the principal component scores λ_{ijk} for each participant indexed by j so we impose a model

$$\lambda_{ijk} = \mu_{ik} + \rho_{ik}\epsilon_{ijk}$$

where $\epsilon_{ijk} \sim N(0, \tau_k)$ and μ_{ik}, ρ_{ik} are the center and scale parameters corresponding to each principal component indexed by k. Note that this is almost exactly the ComBat model, except for each of the k principal component scores. After imposing a prior on each parameter, we can then estimate each of the k pairs of center and scale parameters via empirical Bayes using the n_i observations per site. Then after lazily dubbing this method CovBat we can remove the site effect by residualizing to obtain

$$\lambda_{ijk}^{CovBat} = \frac{\lambda_{ijk} - \hat{\mu}_{ik}}{\hat{\rho}_{ik}} \qquad y_{ij}^{CovBat} = \sum_{k=1}^{K} \lambda_{ijk}^{CovBat}\phi_k$$

PC Method Properties

From standard PCA results, we have that for a random vector Y with mean zero and covariance Σ that $Y = \sum_{k=1}^{P} \Lambda_k \phi_k$ where P is the dimension of the original dataset and Λ_k are uncorrelated random variables with mean 0. Then we can derive the covariance of Y as

$$Cov(Y) = Cov(\sum_{k=1}^{P} \Lambda_k \phi_k) = \sum_{k=1}^{P} Var(\Lambda_k) \phi_k \phi_k^T$$

Given a random sample, y_i such that $i = 1, \ldots, n$ and $y_i \approx \sum_{k=1}^{P} \lambda_{ik} \phi_k$ we can approximate $Var(\Lambda_k) \approx \frac{1}{n-1} \sum_{i=1}^{n} (\lambda_{ik} - \bar{\lambda_k})^2$. This property indicates that removal of site effect in the covariance of Y can be accomplished by harmonizing the PC score variances $Var(\Lambda_k)$ and that we can investigate this harmonization by looking at the relative amount of $\phi_k \phi_k^T$ within each site.

Functional Connectivity Harmonization

To extent this method to functional connectivity, we turn to Population Value Decomposition (PVD; Crainiceanu et al. 2011). In the functional connectivity setting, we are instead dealing with sample functional connectivity matrices Σ_{ij} for each of j subjects in i sites. Under the PVD model, these correlation matrices can be approximated via

$$\Sigma_{ij} = \mathbf{P}\mathbf{V}_{ij}\mathbf{P}^T + \mathbf{E}_{ij}$$

where estimation of $\mathbf{P}$ proceeds according to the original paper. To obtain an analogous principal components decomposition of these functional connectivity matrices, we turn to the reduced-dimension $A \times A$ matrices $\mathbf{V}_{ij}$. Let $\mathbf{v}_{ij}$ be the vectorized versions of the $\mathbf{V}_{ij}$ and perform PCA on these vectors to obtain

$$\mathbf{V_{ij}} = \sum_{k=1}^{K} \Lambda_{ijk} \boldsymbol{\phi}_k + \boldsymbol{\eta}_{ij}$$

where Λ_{ijk} are uncorrelated random coefficients, $\boldsymbol{\phi}_k$ are the eigenvectors obtained from PCA analysis arranged into $A \times A$ matrices, and η_{ij} is a noise process. By multiplying on the left by $\mathbf{P}$ and right by $\mathbf{P}^T$ we obtain

$$\Sigma_{ij} = \sum_{k=1}^{K} \Lambda_{ijk} \mathbf{P} \boldsymbol{\phi}_k \mathbf{P}^T + \mathbf{P} \boldsymbol{\eta}_i \mathbf{P}^T + \mathbf{E}_{ij} = \sum_{k=1}^{K} \Lambda_{ijk} \mathbf{\Phi}_k + \mathbf{e}_{ij}$$

after some substitutions. From here, we can proceed to correct on the scores across sites via a ComBat-like procedure as detailed in the first section. In this case, we are instead correcting on eigenmatrices $\mathbf{\Phi}_k$ rather than eigenvectors, and performing the correction on scores obtained from the lower-dimensional representations $\mathbf{V}_{ij}$. This approach should harmonize the functional connectivity matrices across sites, but further exploration is required to determine exactly what harmonization on these scores would do to the sample properties.

Unconstrained Covariance Scaling Parameter

Previously, we discussed estimation of the covariance parameters through a two-stage approach involving application of original ComBat followed by SVD. This method would work as follows. First, assume that the $p \times n_i$ dimensional data matrices Y_i $i = 1, \ldots, M$ follow

$$Y_i = A + X\beta + \Gamma_i + \Delta_i E$$

where A is the intercept matrix, X is the design matrix, β is the coefficient vector, Γ_i is the mean site effect matrix, Δ_i is the $p \times p$ covariance site effect parameter matrix, and E is the error matrix where each column e_j, $j = 1, \ldots, n_i$ follows $e_j \sim N(0, \Sigma)$. Original ComBat should be able to remove the intercept, covariates effect, and mean site effects so that subsequently,

$$Y_i^{ComBat} \approx \Delta_i^* E \qquad Cov(Y_i) = \Delta_i^* E E' \Delta_i^{*T} = \Delta_i^* B \Delta_i^{*T}$$

where Δ_i^* is the covariance site effect after ComBat, which should remove site effects from the diagonal elements. Optimally, Δ_i^* would just be Δ_i with variances on the diagonal uninfluenced by site. However, this model involves $p(p-1)/2$ parameters for Δ_i^* and B which are multiplied together. After further consideration, it is likely that this model is intractable or at least extremely difficult to estimate.