Creative Component Research

Concept

Whats batch effect?
Batch effects occur because measurements are affected by laboratory conditions, reagent lots and personnel differences. In principal component analysis (PCA), samples often cluster by laboratory, processing day or experimental plan rather than by their biological characteristics.

General Process

Dataset: Heterogeneous datasets with low replicate numbers in TCGA, GTEx and IHEC.

Target: determine if batch adjustment leads to an improvement in data quality, especially in cases of low replicate numbers. Or how the performance of BEA methods could be assessed and compared objectively in a heterogeneous scenario with few replicates and many diverse sample types[5].

Mian Idea: The method borrows information from the Cell Ontology to establish if batch adjustment leads to a better agreement between observed pairwise similarity and similarity of cell types inferred from the ontology.

Result: Batch effects in heterogeneous datasets with low replicate numbers cannot be adequately adjusted.

Method/Main idea in Detail

Start: consider only the similarity of samples belonging to the same sample type. eg: a liver cell may be expected to be more similar in expression to a kidney cell , or the liver cells from different part.

Overview of the method:
(a) Given a gene expression matrix as input, we compute similarities for all pairs of samples, giving rise to a matrix D of observed similarities.
(b) Using an ontology as input, we compute a matrix O of expected similarities based on the path lengths between the terms corresponding to each sample.
(c)we recompute the similarity in matrix O after BEA.
(d)Finally, we correlate for each sample two vectors, namely the observed sample similarities from matrix D to the expected similarities in matrix O that correspond to their sample type before and after BEA. Then we get ontology scores for each sample in gene expression matrix. we can compare ontology scores before and after BEA to check the null hypothesis.

d)H_0: BEA does not lead to significantly higher correlation of expected and observed similarity, i.e. the ontology score will not improve.

Correcting for Batch Effects:
(a)Combat
(b)SVA
(c)RUV

Results

(a) Use randomization experiments to assess the benefit of using the Cell Ontology for computing ontology scores. Results indicate that the ontology score indeed informs about the sample similarities we can expect from gene expression data. Besides, the better the ontology and the assigned sample labels reflect the underlying biology, the higher the ontology score will be.
(b) When add Gaussian noise N(μ=10, 1) to all genes in GTEX, the score decreases when the fraction of samples exposed to noise increases. This means score is sensitive to noise.
(c) Original GTEx data clusters per tissue, after adding noise, GTEx data become two large clusters. After BEA with Combat shows that samples no longer cluster by batches and instead cluster by tissue as in the original data except liver samples. And we can use the ontology score to circumvent this subjective assessment directly and effectively. (the score decreases when noise is added and is nearly restored to the original level after adjustment via Combat with the notable exception of liver samples.)
(d) In heterogeneous datasets, the RUV method does not perform well according to the ontology score, while SVA seems to be able to successfully adjust GTEx data, while exhibiting poor performance on TCGA data. Combat shows only marginal improvement in the ontology scores. Besides, in datasets with only few samples are available per tissue/ cell-type and the overlap of tissues/cell-types being present, RUV performs favorably for DEEP and Roadmap data, while Combat seems to obtain good results on Blueprint samples. For ENCODE data, no adjustment method is able to improve the score.

Correcting for Batch Effects—Combat

Concept: comBat allows users to adjust for batch effects in datasets where the batch covariate is known. It uses either parametric or non-parametric empirical Bayes frameworks for adjusting data for batch effects. Users are returned an expression matrix that has been corrected for batch effects.

Conditions: The input data are assumed to be cleaned and normalized before batch effect removal.

Main Idea: Combat was originally designed to adjust batch effects in microarray datasets with small batch sizes in mind. It can be used with or without adding group variables whereas the batch variables have to be provided.
let Y_{ijg} represent the expression value for gene g for sample j from batch i. Assumes that

model

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\epsilon_{ijg}$$

where α_g is the overall gene expression, X is a design matrix for sample conditions, and β_g is the vector of regression coefcients corresponding to X. The error terms, ϵ_{ijg}, can be assumed to follow a Normal distribution with expected value of zero and variance σ_g^2 . The γ_{ig} and δ_{ig} represent the additive and multiplicative batch effects of batch i for gene g, respectively.

Combat in R: source("http://bioconductor.org/biocLite.R")
biocLite("sva")
library(sva)
combatdata = ComBat(dat=expdata, batch=batch, mod=label, par.prior=T, prior.plots=F)

Evaluation: Notably, many methods are not appropriate when batch sizes are small (less than 10), which is often the case. In order to account for this situation, we have to use EB framework for adjusting for additive, multiplicative, and exponential (when data have been log transformed) batch effects. After using EB method, Downstream analyses are appropriate for the combined data without having to worry about batch effects. From the plot[3], the amount of batch parameter(their variance and mean) shrinkage that occurred for the adjustments for 200 genes from one of the batches.

Robustness: from plot[3], data with outliers in batch is barely adjusted, the batches without outliers are adjusted correctly in the EB data, which means EB method can robustly dealing with outliers.

Disadvantages: Combat is critical for use with small batches. However, for cases with large batches or substantial batch effects, combat method should be similar to the two-way ANOVA approach. While two-way ANOVA approach is sensitive to unbalanced data.
when study groups are not evenly distributed across batches, batch adjustment using ComBat on unbalanced data sets may be not only harmful for small sample size and batch sizes, but also harmful for large one[4].
When the group batch distribution is unbalanced, i.e.where batches do not have the same composition of groups, this will lead to deflated estimates of the estimation errors, and over-confidence in the results. The problem do not have matter with sample size.
The size and impact of the problem will depend greatly on how unbalanced the group-batch distribution is: if it is only moderately unbalanced,it need not be a concern, whereas in heavily unbalanced cases it may have a huge influence.

Advantages: (a) EB adjustments allow for the combination of multiple data sets and are robust to small sample sizes. This method uses an empirical Bayes approach to avoid over-correcting which is critical for use with small batches.
(b) It will moderate the side-effects of batch adjustments.
(c) Combat also allow covariates to be included in the batch adjustment.

RUV–Removing Unwanted Variation from High Dimensional Data with Negative Controls

Conditions: RUV can be used in the case where there is no predefined factor of interest.
It can also be the case that one needs to normalize a dataset without knowing which factors of interest will be studied. our aim is to remove the unwanted variation without losing the variation of interest.

Main Idea: model of RUV is a linear model with a term representing the variation of interest and another term representing the unwanted variation:

$$Y = X\beta + W\alpha + \epsilon$$

with Y $\in R^{m\times n}$, X $\in R^{m\times p}$, $\beta \in R^{p\times n}$, W $\in R^{m\times k}$, $\alpha \in R^{k\times n}$, and $\epsilon \in R^{m\times n}$.
Y is the observed matrix of expression of n genes for m samples, X represents the p factors of interest, W represents the k unwanted factors and ϵ represent some noise, typically $\epsilon_j \sim N(0, \sigma_\epsilon^2 I_m)$, j =1,...,n. Both α and β are modeled as fixed,i.e., non-random.

result: In the paper that written by Jacob in 2015[2], he use two methods to correct the batch effect with RUV, One method uses the negative control gene-based estimator of unwanted factors, and estimates the effect of these factors on gene expression using a random effect model. The second method relies on replicate samples and estimates the unwanted variation using the variation observed in differences of replicates. Both estimators can be improved by joint modeling of the variation of interest and the unwanted variation.
All the methods we introduce are available in the bioconductor package RUV normalize. The replicate-based method performed less well than the control gene based one(unless a really large number of replicates was available)but was unaffected by poor quality control genes and to large confounding level.
Jacob were able to verify that both proposed methods provide a better correction even in the case where the factor of interest and the unwanted factors are totally confounded.

comBAT and RUV comparison

ComBat performs well to remove an observed batch if it is largely independent from the signal of interest, and also have a better adjustment with small sample size.
Naive RUV-2 performs well to estimate and remove unobserved unwanted variation from gene expression data when the factor of interest is also unobserved. it also performs well to remove a batch if the batch is not too associated with the factor of interest.
RUV-4 may be used when the goal of the analysis is to determine which of the features are truly associated with a given factor of interest. One nice property of RUV-4 is that it is not necessary to estimate the number of unwanted factors in the model[1], and it is suitable to use in High dimensional data.

Estimation Procedure for ComBAT

Step 1: Standardization Procedure
The aim of the standardization procedure is to reduce gene-to-gene variation in the data. from the model we know that $\alpha_g, \beta_g, \gamma_g$, and σ_g^2 to differ across genes. To more clearly extract the common batch biases from the data, the standardization procedure standardizes all genes to have the similar overall mean and variance. Without standardization, these differences will bias the EB estimates of the prior distribution of batch effect, the gene-specific variation increases the noise in the data and inflates the prior variance, decreasing the amount of shrinkage that occurs.
So fristly, we can use ordinary least-squares approach to estimate the model parameters $\alpha_g, \beta_g, \gamma_{ig}$ as $\hat{\alpha_g}, \hat{\beta}_g, \hat{\gamma_{ig}}$ for i = 1,...,m and g = 1,...,G.
Then we estimate $\hat{\sigma_g^2} = \frac{1}{N}\sum_{ij}(Y_{ijg} - \hat{\alpha_g} - X\hat{\beta}_g - \hat{\gamma_{ig}})^2$ (N is the total number of samples).
The standardized data, are now calculated by

$$Z_{ijg} = \frac{Y_{ijg} - \hat{\alpha_g} - X\hat{\beta}_g}{\hat{\sigma_g^2}}$$

Step 2: EB batch effect parameter estimates using parametric empirical priors
we assume that the standardized data, Z_{ijg}, satisfy the distributional form, $Z_{ijg} \sim N(\gamma_{ig}, \delta_{ig}^2)$. That means, our data model is

$$f(Z_{ijg}|\gamma_{ig}, \delta_{ig}^2)$$

we assumed that the parametric forms for prior distributions on the batch effect parameters to be $\pi(\theta)$, that is

$$\gamma_{ig} \sim N(\gamma_i, \tau_i^2)$$

and

$$\delta_{ig}^2 \sim InverseGamma(\lambda_i, \theta_i)$$

When mean and variance are not independent for each other in Z_ijg distribution, we can find posterior by $\pi(\theta_1, \theta_2) = \pi(\theta_1|\theta_2)\pi(\theta_2)$, that is $p(\gamma_{ig}, \delta_{ig}^2|Z_{ijg}) = p(\gamma_{ig}|\delta_{ig}^2, Z_{ijg})p(\delta_{ig}^2|Z_{ijg})$.
If we want to find posterior distributions of γ_{ig} and δ_{ig}^2 , we need to estimate the hyperparameters $\gamma_i, \tau_i^2, \lambda_i, \theta_i$, and we can use Method of Moments (MM) to achieve this:
Letting $\hat{\gamma_{ig}} = \frac{\sum_j Z_{ijg}}{n_i}$ (batch i sample mean for gene g), we have

$$\bar{\gamma_i} = \frac{\sum_g \hat{\gamma_{ig}}}{G}$$

$$\bar{\tau_i^2} = \frac{\sum_g (\hat{\gamma_{ig}} - \bar{\gamma_i})^2}{G-1}$$

Letting $\hat{\delta_{ig}^2} = \frac{\sum_j (Z_{ijg} - \hat{\gamma_{ig}})^2}{n_i - 1}$, we have

$$\bar{V_i} = \frac{\sum_g \hat{\delta_{ig}^2}}{G}$$

$$\bar{S_i^2} = \frac{\sum_g (\hat{\delta_{ig}} - \bar{V_i})^2}{G-1}$$

We also have

$$\bar{V_i} = \frac{\beta}{\alpha - 1} = \frac{\theta_i}{\lambda_i - 1}$$

$$\bar{S_i^2} = \frac{\theta_i^2}{(\lambda_i - 1)^2(\lambda_i - 2)}$$

Solve this equation, we can get

$$\bar{\lambda_i} = \frac{\bar{V_i} + 2\bar{S_i^2}}{\bar{S_i^2}}$$

$$\bar{\theta_i} = \frac{\bar{V_i^3} + \bar{V_i}\bar{S_i^2}}{\bar{S_i^2}}$$

After estimation parameter, we can caculate the conditional posterior distribution for γ_{ig} and δ_{ig}^2.
the γ_{ig} conditional posterior distribution is:

$$p(\gamma_{ig}|Z_{ig}, \delta_{ig}^2) \propto L(Z_{ig}|\gamma_{ig}, \delta_{ig}^2)\pi(\gamma_{ig})$$

$$p(\gamma_{ig}|Z_{ig}, \delta_{ig}^2) \propto \left\{ -\frac{1}{2}\left(\frac{n_i\tau_i^2 + \delta_i^2 g}{\delta_{ig}^2\tau_i^2}\right)\left[\gamma_{ig}^2 - 2\left(\frac{\tau_i^2\sum_j Z_{ijg} + \delta_{ig}^2\gamma_i}{n_i\tau_i^2 + \delta_{ig}^2}\right)\gamma_{ig}\right]\right\}$$

The distribution above can be determined to be the kernel of a normal distribution with expected value that within the function exp() above. Given $\hat{\gamma_{ig}}, \hat{\delta_{ig}}, \bar{\gamma_i}, \bar{\tau_i^2}$ as input, we have:

$$\gamma_{ig}^\star = \hat{E}[\gamma_{ig}|Z_{ig}, \delta_{ig}^{2^\star}] = \frac{n_i\bar{\tau_i^2}\hat{\gamma_{ig}} + \delta_{ig}^{2^\star}\bar{\gamma_i}}{n_i\bar{\tau_i^2} + \delta_{ig}^{2^\star}}$$

the δ_{ig}^2 conditional posterior distribution is:

$$p(\delta_{ig}^2|Z_{ig}, \gamma_{ig}) \propto L(Z_{ig}|\gamma_{ig}, \delta_{ig}^2)\pi(\delta_{ig}^2)$$

$$p(\delta_{ig}^2|Z_{ig}, \gamma_{ig}) \propto \left(\delta_{ig}^2\right)^{-(\frac{n_i}{2}+\lambda_i)-1} exp\left\{-\frac{\theta_i + \frac{1}{2}\sum_j (Z_{ijg} - \gamma_{ig})^2}{\delta_{ig}^2}\right\}$$

Which can be identified as an Inverse Gamma distribution with expected value within function exp() above. Given $\hat{\gamma_{ig}}, \hat{\sigma_{ig}}, \bar{\gamma_i}, \bar{\tau_i^2}$ as input, we have:

$$\hat{E}[\delta_{ig}^2|Z_{ig}, \gamma_{ig}^\star] = \frac{\bar{\theta_i} + \frac{1}{2}\sum_j (Z_{ijg} - \gamma_{ig}^\star)^2}{\frac{n_i}{2} + \bar{\lambda_i} - 1}$$

Finally, however, we should notice that there are no closed form solutions for $\gamma_{ig}^{\star}$ and $\delta_{ig}^{2\star}$ parameters, and therefore they must be found iteratively. We can use $\hat{\delta_{ig}^2}$ as starting value, calculate and estimate of $\gamma_{ig}^{\star}$. Then use the newly found $\gamma_{ig}^{\star}$ to estimate $\delta_{ig}^{2\star}$. Iterate the previous steps until convergence. This can be shown to be a simple case of the EM Algorithm (Dempster et al., 1977), and typically only a few iterations (less than 30) are necessary to achieve very accurate estimates for the EB batch adjustments.

EB Batch Effect Parameter Estimates using Nonparametric Empirical Priors

The parametric forms for the prior estimates were not satisfactory for data set 2, leading to the need for more flexible options for the prior distributions, so we use a nonparametric empirical prior to accommodate these data. We assume that the data has been standardized as in Step 1, and that the standardized data, Z_{ijg}, satisfies the distributional form, $Z_{ijg} \sim N(\gamma_{ig}, \delta_{ig}^2)$. Also let $\hat{\gamma_{ig}} = \frac{\sum_j Z_{ijg}}{n_i}$ and $\hat{\delta_{ig}^2} = \frac{\sum_j (Z_{ijg} - \hat{\gamma_{ig}})^2}{n_i - 1}$ as in the previous section. Let Z_{ig} be a vector containing Z_{ijg} for $j = 1, , n_i$.

From the data model Z_{ig} and unkown prior, we assume the posterior distribution to be $p(Z_{ig}, \gamma_{ig}, \delta_{ig}^2)$, so the posterior expectation of Z_{ig} is given by

$$E[\gamma_{ig}] = \int \gamma_{ig} p(Z_{ig}, \gamma_{ig}, \delta_{ig}^2) d(\gamma_{ig}, \delta_{ig}^2)$$

Let $\pi(\gamma_{ig}, \delta_{ig}^2)$ be the (unspecified) density function for the prior, where $p(Z_{ig}, \gamma_{ig}, \delta_{ig}^2)$ in the equation above is

$$p(Z_{ig}, \gamma_{ig}, \delta_{ig}^2) = \frac{L(Z_{ig}|\gamma_{ig}, \delta_{ig}^2)\pi(\gamma_{ig}, \delta_{ig}^2}{\int L(Z_{ig}|\gamma_{ig}, \delta_{ig}^2)\pi(\gamma_{ig}, \delta_{ig}^2) d(\gamma_{ig}, \delta_{ig}^2)}$$

Letting $\int L(Z_{ig}|\gamma_{ig}, \delta_{ig}^2)\pi(\gamma_{ig}, \delta_{ig}^2) d(\gamma_{ig}, \delta_{ig}^2) = C(Z_{ig})$, we estimate both Z_{ig} and the integral of $E(\gamma_{ig})$ using Monte Carlo integration over the empirically estimated pairs $(hat\gamma_{ig}, \hat{\delta_{ig}^2})$, which are considered random draws from $\pi(\gamma_{ig}, \delta_{ig}^2)$.

Let $W_{ig} = L(Z_{ig}|\hat{\gamma_{ig}}, \hat{\delta_{ig}^2})$ for $g = 1, ..., G$. So $\hat{C}(Z_{ig}) = \frac{\sum_g W_{ig}}{n}$. So $E[\gamma_{ig}]$ can be estimated by

$$\gamma_{ig}^{\star} = E[\gamma_{ig}] = \frac{\sum_g W_{ig}\hat{\gamma_{ig}}}{n\hat{C}(Z_{ig})} = \frac{\sum_g W_{ig}\hat{\gamma_{ig}}}{\sum_g W_{ig}}$$

The same method is used to find the posterior expectation of δ_{ig}^2, which is

$$\delta_{ig}^{2\star} = \frac{\sum_g W_{ig}\hat{\delta_{ig}^2}}{\sum_g W_{ig}}$$

Step 3: Adjust the data for batch effects
After calculating the adjusted batch effect estimators, $\gamma_{ig}^{\star}$ and $\delta_{ig}^{2^{\star}}$, we now adjust the data. The EB batch adjusted data $Y_{ijg}^{\star}$ can be calculated as

$$Y_{ijg}^{\star} = \frac{\hat{\sigma_g}}{\hat{\delta_{ig}^{\star}}}(Z_{ijg} - \hat{\gamma_{ig}^{\star}}) + \hat{\alpha_g} + X\hat{\beta}_g$$

Estimation Procedure for RUV

Conditions
we are interested in the case where X is not observed. Our objective in general will be to estimate Wα and remove it from Y.
Our model is

$$Y = X\beta + W\alpha + \epsilon$$

with Y $\in R^{m\times n}$, X $\in R^{m\times p}$, $\beta \in R^{p\times n}$, W $\in R^{m\times k}$, $\alpha \in R^{k\times n}$, and $\epsilon \in R^{m\times n}$.
α estimation improvment
when X and W are not expected to be orthogonal, and $W = \hat{W}_2$, $X\beta = 0$: the naive RUV-2 estimator of α is formally given by

$$min||Y - \hat{W}_2\alpha||_F^2$$

Where $\alpha \in R^{k\times n}$, and this is the maximum likelihood estimator of α for our model.
If we keep the same model and endow α with a distribution $\alpha_j \overset{iid}{\sim} N(0, \sigma_\alpha^2 I_k)$, j =1,...,n, the maximum a posteriori estimator of α becomes:

$$min||Y - \hat{W}_2\alpha||_F^2 + \nu||\alpha||_F^2$$

where $\nu = \frac{\sigma_\epsilon^2}{\sigma_\alpha^2}$. Here again like with σ_ϵ, we limit ourselves to a model where σ_α is common to all genes.

the only difference between the naive RUV-2 estimator and the newly introduced random α RUV-2 is the penalty term: the latter is a ridge regression against $\hat{W}_2$ maximum a posteriori for a random α model whereas the former is an ordinary regression maximum likelihood for a fixed α model. In this context where X is unobserved and $X\beta$ is set to 0 to estimate α, this difference can be important if X and W are correlated.
Assuming some structure is known on the unobserved $X\beta$ term, it is possible to write a joint estimator of (Xβ,α) given W rather than fixing $X\beta$=0:

$$min\left\{||Y - W_2\alpha - X\beta||_F^2 + \nu||\alpha||_F^2\right\}$$

where $X\beta \in M$, M is a subset of $R^{m\times n}$, $\beta \in R^{k\times n}$ and $\alpha \in R^{k\times n}$ Such a joint scheme can be used to build a different estimator of W: once an estimate of $X\beta|W,\alpha$ becomes available, W can be re-estimated using and SVD on the residuals Y-$X\beta$ rather than the control genes.

Using negative control samples" to remove unwanted variation
Symmetrically to the negative control genes used to estimate W, we now consider negative control samples for which the factor of interest X is 0.
one way of obtaining such control samples is to use replicate samples,i.e.,samples that come from the same tissue but which were hybridized in two different settings, say across time or platform. The profile formed by the difference of two such replicates should only be influenced by unwanted variation– those whose levels differ between the two replicates. In particular, the X of this difference should be 0. By construction, this approach is only able to deal with unwanted variation with respect to which replicates are available, which is often the case for technical unwanted variation but rarely the case for biological unwanted variation.
More generally when there are more than two replicates, one may take all pairwise differences or the differences between each replicate and the average of the other replicates.

algorithm
(1) We denote by d the indices of artificial control samples formed by differences of replicates, and we therefore have X^d =0 where X^d are the rows of X indexed by d. We assume that k = d (k is the cols of α) and therefore $\hat{\alpha} = Y^d$.
(2) Use the rows of Y corresponding to control samples $Y^d = W^d\alpha + \epsilon^d$ to estimate α. Assuming i.i.d. noise

$$\epsilon_j \sim N(0, \sigma_\epsilon^2 I_m)$$

, j =1,...,n, the $W^d\alpha$ matrix maximizing the likelihood of this model is argmin

$$min||Y^d - W^d\alpha||_F^2$$

This argmin is reached for $\hat{W^d\alpha} = PE_kQ^\top$, where $Y_d = PEQ^\top$ is the SVD of Y_d, and E_k is the diagonal matrix with the k largest singular values as its k first entries and 0 on the rest of the diagonal. We can use $\hat{\alpha} = E_kQ^\top$
(3) Caculated the $\hat{W^d\alpha} = Y\hat{\alpha}^\top(\hat{\alpha}\hat{\alpha}^\top)^{-1}\hat{\alpha}$. Since $\hat{\alpha} = Y^d$, so we have

$$\hat{W^d\alpha} = Y_c(Y_c^d)^\top(Y_c^d(Y_c^d)^\top)^{-1}Y^d$$

where Y_c is the control genes.
(4)

References

[1] Johann A Gagnon-Bartsch, Laurent Jacob, and Terence P Speed. Removing unwanted variation from high dimensional data with negative controls. *Berkeley: Tech Reports from Dep Stat Univ California*, pages 1–112, 2013.

[2] Laurent Jacob, Johann A Gagnon-Bartsch, and Terence P Speed. Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics*, 17(1):16–28, 2015.

[3] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[4] Vegard Nygaard, Einar Andreas Rødland, and Eivind Hovig. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, 17(1):29–39, 2016.

[5] Florian Schmidt, Markus List, Engin Cukuroglu, Sebastian Köhler, Jonathan Göke, and Marcel H Schulz. An ontology-based method for assessing batch effect adjustment approaches in heterogeneous datasets. *Bioinformatics*, 34(17):i908–i916, 2018.