Covariance Adjustment for Batch Effects in Gene Expression Data

Jeongyoun Ahn

Department of Statistics
University of Georgia

June 24, 2013

joint work with Jungae Lee and Kevin K. Dobbin
Data from multiple batches

- Several data sets (batches) are obtained for the same research goal.
- Individual studies have relatively small sample with a lot of genes.
- Merging data or inter-laboratory study to discover reliable bio-marker and robust prognostic models.
- Significant biases among batches trade off increased sample size.
- Ignoring it can lead erroneous conclusions.
Batch effect

- Systematic bias due to different sites or different times.
- From subtle inconsistencies in sample maintenance, RNA extraction techniques, hybridization protocols, or many other experimental condition.
- Making laboratory protocols uniform could reduce, but not eliminate, batch effects.
- Been found in mass spectrometry data, copy number abnormality data, methylation array data, and DNA sequencing data.
Batch effect example

Breast cancer data (Wang et al., 2005; Desmedt et al., 2007) collected at two different laboratories.
Batch effect example from Luo et al. (2010)
Batch effect example

- **MD Anderson Dataset** (Cross Hybridization Time)
  - PC1: 63.1%
  - PC2: 27.7%

- **Hammer Dataset** (Cross Hybridization Time)
  - PC1: 12.5%
  - PC2: 4.6%

- **Iconix Dataset**
  - PC1: 11.8%
  - PC2: 8.4%

- **UAMS Dataset** (Cross Generation)
  - PC1: 43.0%
  - PC2: 9.0%

- **NB Dataset** (Cross Channel)
  - PC1: 80%
  - PC2: 60%
Existing Methods for Batch Adjustment

- Mean Centering/Standardization
- Distance Weighted Discrimination
- Empirical Bayes Method
- Cross Platform Normalization
Mean Centering & Standardization

$Y_{ijg}$: gene expression value of patient $j$ for gene $g$ in batch $i$.

- Mean centering: The adjusted value $Y_{ijg}^*$ is the mean-centered value, i.e., all genes have mean zero within a batch.

$$Y_{ijg}^* = Y_{ijg} - \bar{Y}_{i.g}$$

- Standardization: Makes the standard deviation 1.

$$Y_{ijg}^* = \frac{Y_{ijg} - \bar{Y}_{i.g}}{SD_{i.g}}$$

- (Unsaid) Assumption: the batch bias only exists in the location shift and scale change for each gene.
Distance Weighted Discrimination (DWD)

2. Find the direction to maximize separation of batch biases.
3. Each subpopulation is shifted on the given direction until its mean reaches the separating hyperplane.
DWD Adjustment

Benito et al, Bioinf 2004, fig 1. The batch effect isn't clearly dominant. LDA is ok here.

Copyright 2004-2010, KR Coombes, KA Baggerly, BM Broom
Distance Weighted Discrimination (DWD)

1. Any linear discrimination can be used.
2. Essentially “incomplete” mean centering.
3. Assumes the batch effect exists in the shift of the means.
Empirical Bayes Method (Johnson et al., 2007)

- Uses a two way ANOVA for each gene:

\[ Y_{ijgc} = \alpha_g + \beta_{gc} + \gamma_{ig} + \delta_{ig} \epsilon_{ijgc}, \]

- batch \( i \), sample \( j \), bio-label \( c \), and gene \( g \).
- \( \epsilon_{ijgc} \sim N(0, \sigma^2_g) \), \( \beta_{gc} \): biological effect, \( (\gamma_{ig}, \delta_{ig}) \): batch effects.

- Adjust multiplicative effect - not only location, but also scatter.
- Use of biological information is controversial.
Proposed approach

We believe

- Batch effect can alter gene-to-gene relationships as well.
- Not all batches are equal. Some are better.
- Direct use of biological signal can induce a different bias.
Model

Assuming no batch effect, the unobserved data

$$X^* = (X^*_1, \ldots, X^*_p)^T.$$  

$$E(X^*) = \mu^* \text{ and } \text{Var}(X^*) = \Sigma^* \ (p:\text{the number of genes}).$$  

$$Z = \Sigma^{*-1/2}(X^* - \mu^*) \text{ has } E(Z) = 0 \text{ and } \text{Var}(Z) = I.$$  

Data with batch effect

$$X_{ij} = (X_{ij1}, \ldots, X_{ijp})^T.$$  

batch $i \ (i = 1, \ldots, k)$ and $j$th array $(j = 1, \ldots, n_i).$  

$$E(X_{ij}) = \mu_i \text{ and } \text{Var}(X_{ij}) = \Sigma_i \text{ for batch } i.$$
Batch effect function

The $i$th batch effect can be expressed as a function $f_i$,

$$f_i : X^*_j \longrightarrow X_{ij}.$$

Here $X^*_j$ is the $j$th array (sample) that would have been observed without batch effect. Then,

$$X_{ij} = \Sigma_i^{\frac{1}{2}} Z_j + \mu_i$$

$$= \Sigma_i^{\frac{1}{2}} (\Sigma^*^{-\frac{1}{2}} (X^*_j - \mu^*)) + \mu_i$$

$$= \Sigma_i^{\frac{1}{2}} \Sigma^*^{-\frac{1}{2}} X^*_j - \Sigma_i^{\frac{1}{2}} \Sigma^*^{-\frac{1}{2}} \mu^* + \mu_i$$

$$= f_i(X^*_j)$$
Batch adjustment function

The function $f_i$ ($i$th batch effect) is an affine transformation:

$$f_i(X^*) = A_i X^* + b_i,$$

where $A_i = \Sigma_i^{1/2} \Sigma^*^{-1/2}$ and $b_i = -\Sigma_i^{1/2} \Sigma^*^{-1/2} \mu^* + \mu_i$.

The “optimal” batch adjustment is to apply the inverse function $f^{-1}_i$

$$f^{-1}_i(Y) = A_i^{-1} (Y - b_i).$$

$$\hat{A}^{-1}_i = \hat{\Sigma}^{*1/2} \hat{\Sigma}_i^{-1/2}, \quad \hat{b}_i = \hat{\Sigma}_i^{1/2} \hat{\Sigma}^*^{-1/2} \hat{\mu} + \hat{\mu}_i$$
Factor model (Fan et al., 2008) for high dimensional data

\[ \mathbf{y} = \mathbf{Bf} + \mathbf{e}, \]

where \( \mathbf{y} = (Y_1, \ldots, Y_p)^T \), \( \mathbf{f} = (F_1, \ldots, F_q)^T \) (observable factors), \( \mathbf{B} \) is \( p \times q \) regression coefficient matrix, and \( \mathbf{e} = (\epsilon_1, \ldots, \epsilon_p)^T \). \( \mathbb{E}(\mathbf{e}|\mathbf{f}) = \mathbf{0} \) and \( \text{cov}(\mathbf{e}|\mathbf{f}) = \Sigma_0 \) is diagonal.
Covariance estimation based on the factor model

Let \((f_1, y_1), \ldots, (f_n, y_n)\) be \(n\) iid samples of \((f, y)\).

\[
S = \Sigma = \hat{\text{cov}}(y) = \hat{B}\hat{\text{cov}}(f)\hat{B}^T + \hat{\Sigma}_0,
\]

where \(\hat{B} = YX^T(XX^T)^{-1}\), \(X = (f_1, \ldots, f_n)\), \(\hat{\text{cov}}(f)\) is the sample covariance matrix of \(f\), and \(\hat{\Sigma}_0 = \text{diag}(\frac{1}{n}\hat{E}\hat{E}^T)\) with \(\hat{E} = Y - \hat{B}X\).

\(S\) is always invertible even if \(p > n\).
Choice of Factors

- Functional similarity (Pathway): Gene Ontology or Gene Set
- Evolutionary similarity (DNA Sequence): Gene Family
- Empirical approach: K-means gene clustering
Batch Adjustment

With $\hat{A}^{-1}_i = S^* \frac{1}{2} S_i^{-1} (i = 1, \ldots, k)$,

- $S_i$ and $S_i^{-1}$ are estimated by the factor model for batch $i$.
- $S^* = S_{\text{pooled}} = \frac{(n_1-1)S_1 + (n_2-1)S_2 + \cdots + (n_k-1)S_k}{(n_1+n_2+\cdots+n_k-k)}$.

Modified samples

$X^*_{ij} = \hat{f}^{-1}_i(X_{ij}) = \hat{A}^{-1}_i(X_{ij} - \hat{b}_i)$
Sparse Estimation

- The original estimator can be unstable and noisy.
- Consider sparse estimator in order to achieve better homogeneity.
- For off-diagonal element,

\[ \hat{A}^{-1}(\delta) = \{A_{j\ell}(\delta)\} = A_{j\ell} \cdot I(|A_{j\ell}| > \delta), j \neq \ell. \]

- \( \delta \) is determined so that the sample covariances of the batches after adjustment are most “similar”.

Covariance Adjustment for Batch Effects in Gene Expression Data

- Proposed Method
- Sparse Adjustment
Ideal batch

- When one of the batches is believed to be superior quality.
- Facility has the most experience with the technology, better reproducibility on technical replicates.
- Say $m$-th batch, then one can make other batches mimic it.

$$\hat{A}_i^{-1} = S_m^{\frac{1}{2}} S_i^{-\frac{1}{2}}, \quad i = 1, \ldots, k.$$
Asymptotic Property

Under some regularity conditions,

\[ \left\| S_i^{\frac{1}{2}} S_i^{-\frac{1}{2}} - \sum_i^{\frac{1}{2}} \sum_i^{-\frac{1}{2}} \right\|_F = o_p \left( (p^4 q^5 \log(n)/n) \right) \]
Simulated two batches

- Normal distribution with $p = 800$, $n = 50$ each
- $\mu_j$: Uniform$(0, 1.4)$ for Batch 1, and Uniform$(-1.4, 0)$ for Batch 2
- $\Sigma = UDU'$, where $U$ is a orthonormal matrix and $D$ is a diagonal matrix that contains eigenvalues.
- The eigenvalues are set to $\lambda_k = pk^{-1}/3$ for Batch 1 and $\lambda = p \exp(-0.042k)/6$ for Batch 2.
- 100 repetitions.
Covariance Adjustment for Batch Effects in Gene Expression Data

Simulation

Simulated data

PC1(20%)  PC2(4%)  PC3(4%)

PC1(20%)  PC2(4%)  PC3(4%)

PC1(20%)  PC2(4%)  PC3(4%)
Adjusted data - PC plot

Before

PC1 (20%)
PC2 (4%)  

Batch1
Batch2  

MC

PC1 (9%)
PC2 (4%)  

Batch1
Batch2  

DWD

PC1 (13%)
PC2 (4%)  

Batch1
Batch2  

Z-score

PC1 (10%)
PC2 (4%)  

Batch1
Batch2
Adjusted data - PC plot
Adjusted data - eigenvalues

Before

MC

DWD

Z-score
Adjusted data - eigenvalues

- EB
- XPN
- MBCA
- MBCA(B1)
Nearest distance to the opposite batch

![Box plot showing distances for different methods and batches](image-url)
Real Data

Data I: Breast cancer data from Wang et al. (2005)
- Three batches
- Bio-label: Estrogen receptor status (ER+, ER-)

Data II: Lung cancer data from Sheddon et al. (2008)
- Four batches
- Bio-label: Overall survival (Live, Die)

Both data sets have 4634 genes and 22 factors based on GO.
## Description of data sets

<table>
<thead>
<tr>
<th>Batch</th>
<th>n</th>
<th>ER−</th>
<th>ER+</th>
<th>Batch</th>
<th>n</th>
<th>Live</th>
<th>Die</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE2034</td>
<td>286</td>
<td>77</td>
<td>209</td>
<td>HLM</td>
<td>79</td>
<td>19</td>
<td>60</td>
</tr>
<tr>
<td>GSE4922</td>
<td>245</td>
<td>34</td>
<td>211</td>
<td>UM</td>
<td>178</td>
<td>76</td>
<td>102</td>
</tr>
<tr>
<td>GSE7379</td>
<td>198</td>
<td>64</td>
<td>134</td>
<td>CAN/DF</td>
<td>82</td>
<td>47</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MSK</td>
<td>104</td>
<td>65</td>
<td>39</td>
</tr>
</tbody>
</table>
We will consider

1. Similarity of covariances
   Test statistics by Srivastava (2008)
2. Cross-batch prediction
### Equality of Covariance

<table>
<thead>
<tr>
<th>Method</th>
<th>$Q_K^2$ Breast cancer data</th>
<th>$p$-value Breast cancer data</th>
<th>$Q_K^2$ Lung cancer data</th>
<th>$p$-value Lung cancer data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>7.7697</td>
<td>0.0206</td>
<td>8.0182</td>
<td>0.0456</td>
</tr>
<tr>
<td>MC</td>
<td>7.7697</td>
<td>0.0206</td>
<td>8.0182</td>
<td>0.0456</td>
</tr>
<tr>
<td>DWD</td>
<td>7.7697</td>
<td>0.0206</td>
<td>8.0182</td>
<td>0.0456</td>
</tr>
<tr>
<td>Stan.</td>
<td>4.8767</td>
<td>0.0873</td>
<td>19.4446</td>
<td>0.0002</td>
</tr>
<tr>
<td>EB</td>
<td>0.7093</td>
<td>0.7014</td>
<td>3.3300</td>
<td>0.3435</td>
</tr>
<tr>
<td>XPN</td>
<td>0.2802</td>
<td>0.8693</td>
<td>1.8651</td>
<td>0.6009</td>
</tr>
<tr>
<td>MBCA</td>
<td>0.2421</td>
<td>0.8860</td>
<td>0.1585</td>
<td>0.9840</td>
</tr>
</tbody>
</table>
## Targeted adjustment

<table>
<thead>
<tr>
<th>Batch 1 vs. Others</th>
<th>Breast cancer data</th>
<th>Lung cancer data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Q_K^2$</td>
<td>p-value</td>
</tr>
<tr>
<td>MBCA(B1)</td>
<td>0.0553</td>
<td>0.8141</td>
</tr>
<tr>
<td>EB</td>
<td>0.8670</td>
<td>0.3518</td>
</tr>
<tr>
<td>XPN</td>
<td>2.4537</td>
<td>0.1172</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Batch 2 vs. Others</th>
<th>Breast cancer data</th>
<th>Lung cancer data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Q_K^2$</td>
<td>p-value</td>
</tr>
<tr>
<td>MBCA(B2)</td>
<td>0.0611</td>
<td>0.8047</td>
</tr>
<tr>
<td>EB</td>
<td>5.5035</td>
<td>0.0190</td>
</tr>
<tr>
<td>XPN</td>
<td>7.2880</td>
<td>0.0069</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Batch 3 vs. Others</th>
<th>Breast cancer data</th>
<th>Lung cancer data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Q_K^2$</td>
<td>p-value</td>
</tr>
<tr>
<td>MBCA(B3)</td>
<td>1.1364</td>
<td>0.2864</td>
</tr>
<tr>
<td>EB</td>
<td>1.4948</td>
<td>0.2215</td>
</tr>
<tr>
<td>XPN</td>
<td>0.5219</td>
<td>0.4700</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Batch 4 vs. Others</th>
<th>Breast cancer data</th>
<th>Lung cancer data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Q_K^2$</td>
<td>p-value</td>
</tr>
<tr>
<td>MBCA(B4)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EB</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>XPN</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Measure for cross-batch prediction

\[ MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} , \]

where TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively. The MCC value of 1 indicates perfect prediction, $-1$ reverse prediction, and 0 neutral prediction.
Cross-batch prediction

MCC by ridge linear discriminant analysis

<table>
<thead>
<tr>
<th>Method</th>
<th>Breast cancer data</th>
<th>Lung cancer data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(23)1</td>
<td>(13)2</td>
</tr>
<tr>
<td>Before</td>
<td>0.626</td>
<td>0.559</td>
</tr>
<tr>
<td>MC</td>
<td>0.695</td>
<td>0.579</td>
</tr>
<tr>
<td>DWD</td>
<td>0.713</td>
<td>0.614</td>
</tr>
<tr>
<td>Stan.</td>
<td>0.671</td>
<td>0.612</td>
</tr>
<tr>
<td>EB</td>
<td>0.662</td>
<td>0.515</td>
</tr>
<tr>
<td>XPN</td>
<td>0.708</td>
<td>0.596</td>
</tr>
<tr>
<td>MBCA</td>
<td>0.710</td>
<td>0.587</td>
</tr>
</tbody>
</table>
Targeted adjustment

MCC by ridge linear discriminant analysis

<table>
<thead>
<tr>
<th>Method</th>
<th>(234)1</th>
<th>(134)2</th>
<th>(124)3</th>
<th>(123)4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>0.049</td>
<td>-0.055</td>
<td>0.041</td>
<td>0.193</td>
</tr>
<tr>
<td>MBCA(B1)</td>
<td>0.141</td>
<td>0.169</td>
<td>0.102</td>
<td>0.263</td>
</tr>
<tr>
<td>MBCA(B2)</td>
<td>0.065</td>
<td>0.165</td>
<td>0.102</td>
<td>0.178</td>
</tr>
<tr>
<td>MBCA(B3)</td>
<td>0.067</td>
<td>0.181</td>
<td>0.038</td>
<td>0.168</td>
</tr>
<tr>
<td>MBCA(B4)</td>
<td>0.053</td>
<td>0.146</td>
<td>0.074</td>
<td>0.199</td>
</tr>
</tbody>
</table>
Conclusion

- More advanced than gene-wise approaches
- Obtains better homogeneity among batches
- Especially when there is a better, or ideal batch
Thank you for your attention!