Improved cross-study prediction through batch effect adjustment

Roman Hornung

Joint work with David Causeur and Anne-Laure Boulesteix

LMU Munich
Department of Medical Informatics, Biometry and Epidemiology

March, 17th, 2015

Lack of applied high-dimensional prediction rules

- Context: Prediction of phenotypes based on high-dimensional biomolecular data
- Very common in biostatistical/bioinformatical literature
- In contrast: respective prediction rules hardly applied in medical practice
- Such prediction rules could assist medical practitioners in their decision making.

Problem: Batch effects increase prediction error

- In pratice, prediction rules are commonly applied to data ("test data") from different sources than the training data (cross-study prediction).

  ⇒ Batch effects strike!

  ⇒ Potentially high prediction error ↯

- Batch effects: Systematic distortions between different sources of data for reasons unrelated to biological signal of interest.

Batch effect removal for prediction

- Idea: Make the test data more similar to the training data used to obtain the prediction rule.
  ⇒ Smaller prediction error (?)

- Approach: Use (alternated versions of) batch effect removal methods (Luo et al., 2010).

- Restricting requirement: Test data has to come in groups — no batch effect removal for single observations possible.

- We are interested in comparing our recently developed method FAbatch with other methods in this respect.

Simple batch effect removal methods

- Mean-centering: Batchwise centering of the variables
- Standardization: Mean-centering with additional batchwise scaling of the variables to unity
- Ratio-A: Batchwise dividing of the variables by their arithmetic means
- Ratio-G: Batchwise dividing of the variables by their geometric means

ComBat: Location-and-scale adjustment (Johnson et al., 2007)

Model:

$$X_{ijg} = \mu_g + \gamma_{jg} + \delta_{jg}\epsilon_{ijg}, \quad \epsilon_{ijg} \sim N(0, \sigma_g^2)$$

i observation, j Batch, g variable (e.g. gene)

Before batch effect adjustment:

$$\mathbb{E}(X_{ijg}) = \mu_g + \gamma_{jg}, \ \mathrm{Var}(X_{ijg}) = \delta_{jg}^2\sigma_g^2,$$
$$\mathrm{Corr}(X_{ijg_1}, X_{ijg_2}) = \rho_{g_1 g_2}$$

After batch effect adjustment:

$$\mathbb{E}(\widetilde{X}_{ijg}) = \mu_g, \ \mathrm{Var}(\widetilde{X}_{ijg}) = \sigma_g^2, \ \mathrm{Corr}(\widetilde{X}_{ijg_1}, \widetilde{X}_{ijg_2}) = \rho_{g_1 g_2}$$

(frozen) SVA: adjustment for latent factors (Parker et al., 2013)

Model:

$$X_{ijg} = \mu_g + \sum_{l=1}^{m} b_{gl} Z_{ijl} + \epsilon_{ijg}, \quad \mathrm{Var}(\epsilon_{ijg}) = \sigma_g^2,$$

$$\epsilon_{ijg} \ \text{ independent}, \quad Z_{ij1}, \ldots, Z_{ijm} \sim F_{ij} \ \text{ latent factors}$$

Before batch effect adjustment:

$$\mathbb{E}(X_{ijg}) = \mu_g, \ \ \mathrm{Var}(X_{ijg}) = \sigma_{ijg}^2, \ \ \mathrm{Corr}(X_{ijg_1}, X_{ijg_2}) = \rho_{ijg_1g_2}$$

After batch effect adjustment:

$$\mathbb{E}(\widetilde{X}_{ijg}) = \mu_g, \ \ \mathrm{Var}(\widetilde{X}_{ijg}) = \sigma_g^2, \ \ \mathrm{Corr}(\widetilde{X}_{ijg_1}, \widetilde{X}_{ijg_2}) = 0$$

New method FAbatch — based on ComBat and SVA

Model:

$$X_{ijg} = \mu_g + \gamma_{jg} + \sum_{l=1}^{m_j} b_{jgl} Z_{ijl} + \delta_{jg} \epsilon_{ijg}, \quad \epsilon_{ijg} \sim N(0, \sigma_g^2)$$

$$Z_{ij1}, \ldots, Z_{ijm_j} \overset{iid}{\sim} N(0,1), \quad \epsilon_{ijg} \ \text{ independent}$$

Before batch effect adjustment:

$$\mathbb{E}(X_{ijg}) = \mu_g + \gamma_{jg}, \ \mathrm{Var}(X_{ijg}) = \sum_{l=1}^{m} b_{jgl}^2 \delta_{jg}^2 \sigma_g^2,$$

$$\mathrm{Corr}(X_{ijg_1}, X_{ijg_2}) = \sum_{l=1}^{m} b_{jg_1 l} b_{jg_2 l}$$

After batch effect adjustment:

$$\mathbb{E}(\widetilde{X}_{ijg}) = \mu_g, \ \ \mathrm{Var}(\widetilde{X}_{ijg}) = \sigma_g^2, \ \ \mathrm{Corr}(\widetilde{X}_{ijg_1}, \widetilde{X}_{ijg_2}) = 0$$

Protection of biological signal of interest

“Problem”: Due to the class signal we actually have (assuming a two-class prediction problem):

$$\mathbb{E}(X_{ijg}) = \alpha_g + \beta_g cl_i := \mu_{cl_i g}, \quad cl_i \in \{0, 1\}$$

NOT as written before $\mathbb{E}(X_{ijg}) = \mu_g$.

$\Rightarrow$ When assuming a constant mean while estimating and removing the factor influences $\sum_{l=1}^{m_j} b_{jgl} Z_{ijl}$ we remove (part of) the biological signal of interest. ↯

Protection of biological signal of interest

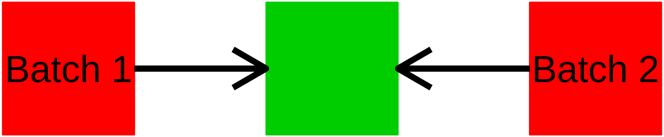
- Class cl_i naturally not known on the test data.

  $\Rightarrow$ Cannot be used in the estimation. ↯

- Solution for FAbatch ✓: Using penalized logistic regression we estimate the probabilities $\mathrm{P}(cl_i = 1)$ and use these for the actual classes $cl_i \in \{0, 1\}$ in the FAbatch estimation algorithm.

Difference to conventional batch effect removal

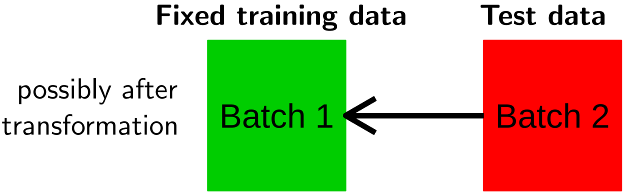
Conventional batch effect removal:



Batch 1

Batch 2

Batch effect removal for prediction purposes:



Fixed training data

Test data

possibly after transformation

Batch 1

Batch 2

Batch effect removal methods in prediction

- Mean-centering, standardization, ratio-A and ratio-G do not have to be altered for prediction, because they do not consider information across batches.

- ComBat and FAbatch do, since they involve the batch-unspecific parameters μ_g (or μ_{cl_ig} resp.) and σ_g^2. In the context of prediction we take the means and variances of the training data to be these parameters.

- For SVA there exists a method called “frozen SVA” designed for prediction.

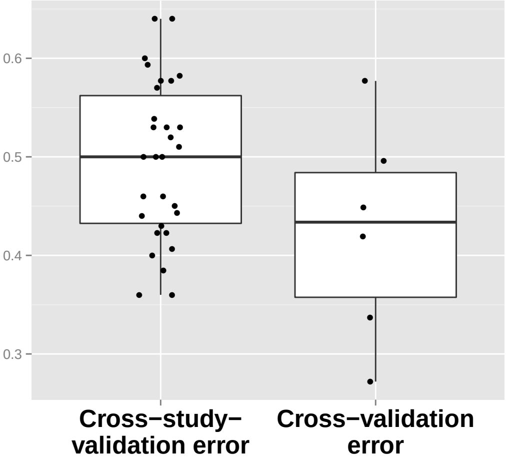
Real data comparison study

- 6 independent breast-cancer microarray datasets (with dichotomized survival times; excluding censorings)
- Sample sizes between 90 and 100 observations, 11,108 variables (after variable filtering)
- Methods: FAbatch, ComBat, frozen SVA, Mean centering, standardization, ratio-A, ratio-G, no batch effect removal

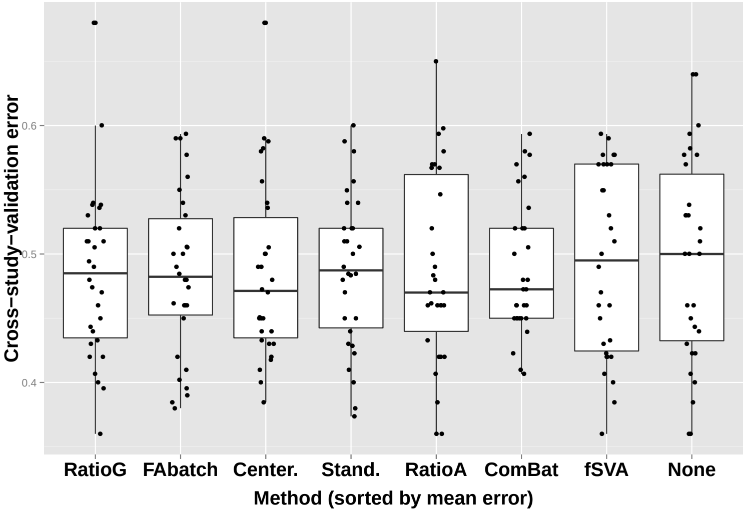
Real data comparison study

- Cross-study validation (see Bernau et al., 2014): Consider all pairs of datasets. In each pair use one dataset as training and the other test set. Then switch the roles of training and test set.
- Classification method: Linear Discriminant Analysis on Partial Least Squares components
- Performance metric: misclassification error rate

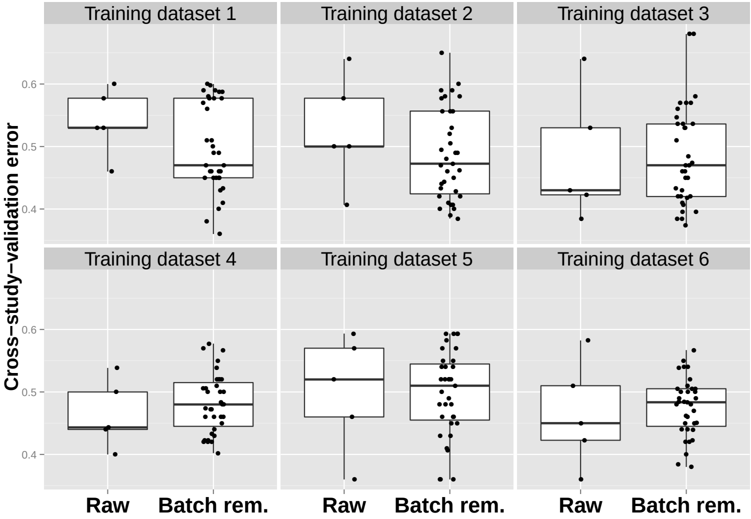
Cross-study valid. error vs. Cross-validation error



Cross-study validation error separate after method



Different improvement for different training data



Conclusion & Outlook

- Empirical study suggests only limited overall reduction of cross-study prediction error through batch effect removal.
- FAbatch performed not clearly better than other methods — has however benefit to keep original range of the data — other than e.g. mean centering
- Outlying training data sets seem to benefit more from batch effect removal.
- Outlook: Prediction rules obtained on several datasets simultaneously may have better cross-study prediction performance, because they incorporate a greater heterogeneity.

Thank you for your attention!

Background

Batch effect removal methods

New method FAbatch

Batch effect removal for prediction

Real data study

Conclusion & Outlook

References

Bernau, C., Riester, M., Boulesteix A.-L., Parmigiani, G., Huttenhower, C., Waldron, L. et al. (2014).
Cross-study validation for the assessment of prediction algorithms.
Bioinformatics **30**, i105—i112.

Friguet, C., Kloareg, C., and Causeur, D. (2009).
A Factor Model Approach to Multiple Testing Under Dependence.
Journal of the American Statistical Association **104**, 1406–1415.

Johnson, W.E., Rabinovic, A., and Li, C. (2007).
Adjusting batch effects in microarray expression data using Empirical Bayes methods.
Biostatistics **8**, 118–127.

Luo, J., Schumacher, M., Scherer, A., Sanoudou, D., Megherbi, D., Davison, T. et al. (2010).
A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data.
The Pharmacogenomics Journal **10**, 278–291.

Parker, H.S., Bravo, H.C., and Leek, J.T. (2013).
Removing batch effects for prediction problems with frozen surrogate variable analysis.
arXiv/1301.3947.