Batch effect removal using spatio-temporal Independent Component Analysis

Emilie Renard EMILIE.RENARD@UCLOUVAIN.BE
Université catholique de Louvain, ICTEAM Institute
Avenue Georges Lemaître, 4 B-1348 Louvain-la-Neuve (Belgium)

Samuel Branders SAMUEL.BRANDERS@TOOLS4PATIENT.COM
Tools4Patient, 48 rue Auguste Piccard, B-6041 Gosselies (Belgium)

P.-A. Absil PA.ABSIL@UCLOUVAIN.BE
Université catholique de Louvain, ICTEAM Institute
Avenue Georges Lemaître, 4 B-1348 Louvain-la-Neuve (Belgium)

Keywords: Batch effect removal, expression data, spatio-temporal independent component analysis

Abstract

When merging gene expression datasets, experimental and data processing conditions proper to each dataset generally influence the expression values and can hide the biological effect of interest. It is then important to normalize the bigger merged dataset regarding those batch effects, as failing to adjust for them may adversely impact statistical inference. In this context, we present a method proposed in (Renard et al., 2016) that uses a "spatio-temporal" independent component analysis to model the influence of those unwanted effects and remove them from the data.

1. Context

Nowadays, the development of sequencing technologies allows to measure gene expression levels at a reasonable cost. However due to the limited number of samples that can be processed at the same time in an experiment, the size of such datasets is often limited in samples. As statistical inferences need a high number of samples to be robust enough and generalizable to other data, merging and combining different datasets appears as a simple solution to increase the number of samples analyzed and to potentially improve the relevance of the biological information extracted.

Expression levels of genes are the result of interactions between different biological processes, which can increase or decrease the expression level measured, and noise may also be added at each step of data acquisition. When merging different datasets, some of the main factors influencing expression levels are typically due to the fact that the samples were not processed in exactly the same conditions from one experiment/dataset to another. Those batch effects can be quite large and hide the effects related to the biological process of interest. It is then important to be able to combine datasets from different sources while removing the unwanted variations such as batch effects. Different methods exist to tackle the problem of batch effect removal, each having its advantages and weaknesses (Chen et al., 2011; Lazar et al., 2013). Here we present recent work (Renard et al., 2016) building on (Alter et al., 2000) and (Renard et al., 2014) where we propose to use spatio-temporal Independent Component Analysis (ICA) to remove batch effects when combining microarray datasets.

2. Approach

We use spatio-temporal Independent Component Analysis (ICA) to remove batch effects when combining microarray datasets. We call *aggregated dataset* the bigger dataset resulting of the concatenation of the smaller datasets, and *sub-datasets* or *batches* the smaller datasets.

Appearing in *Proceedings of Benelearn 2016*. Copyright 2016 by the author(s)/owner(s).

We consider the aggregated dataset as a gene-by-sample matrix X, where $X_{i,j}$ indicates the value of gene i in sample j. Applying the ICA method developed in (Renard et al., 2014) to matrix X yields a decomposition $X \approx AB^T = \sum_{k=1}^{K} A_{:,k} B_{:,k}^T$ where component $A_{:,k}$ can be interpreted as the gene activation pattern of component k and component $B_{:,k}$ as the weights of this pattern in the samples.

After factorization of the aggregated dataset, we select the components $B_{:,k}$ that correlate with the batch number. This correlation is estimated by means of the R^2 value that measures how well a variable x can predict a variable y in a linear model. If a component presents some correlation with the sub-datasets, then this component is removed from the initial aggregated dataset X to obtain a final dataset $X_n = X - \sum_{k\ selected} A_{:,k} B_{:,k}^T$, hopefully cleaned from the main batch effects.

The advantage of a matrix factorization approach is that the removed components are interpretable: it is easy to check that they do not correlate with some biological information of interest. ICA was shown to be able to model different sources of variation in microarray experiments (Teschendorff et al., 2011), so we propose here an ICA based approach. We use a spatio-temporal ICA method with a trade-off parameter α that allows to explore a continuum between imposing independence solely on A ($\alpha = 0$) and solely on B ($\alpha = 1$). This parameter allows to adjust between modeling of variation sources and redundancy between components. The term "spatio-temporal" comes from the pixel-by-time data in medical imaging for which the concept was introduced.

3. Results

We combined different breast cancer expression datasets to test our normalization method. With those datasets come different pieces of information as the age of the patient, grade and size of the tumor,... We compared the performances of an SVM to predict estrogen-receptor status (ER) after different normalization methods (no normalization, simple centering-scaling by sub-dataset, our proposition, replacing the ICA by an SVD, and ComBat (Johnson et al., 2007)). We can see on Figure 1 that the ICA allows a better modeling of the external information. The case without normalization appears to have a larger variance. ComBat has a smaller variance than the case without normalization, but bigger than the factorization methods. ICA and SVD are closer, ICA being slightly higher. More details on the experiments and comparison of the different methods can be found in Section 3 of (Renard et al., 2016).

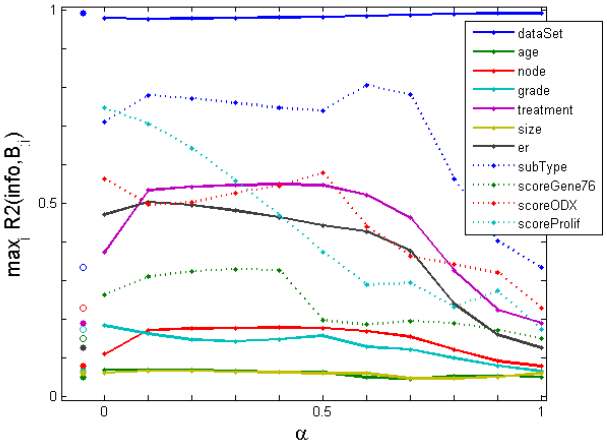


Figure 1. Recovering of external information by the spatio-temporal ICA and SVD

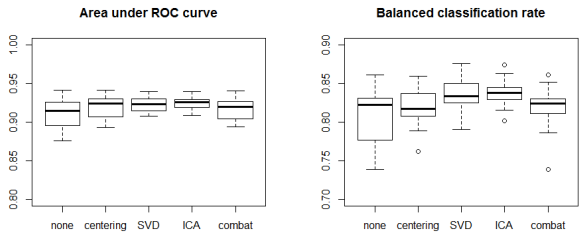


Figure 2. Area under the ROC curve and balanced classification rate for the validating sets.

4. Conclusion

The factorization approach enables to better understand what is removed in the cleaning process; our method can recover external information better than using a simple singular value decomposition. The spatio-temporal parameter α allows to adjust between modeling of external information and redundancy between components. Compared with other normalization methods, results of a classification task on a real dataset shows a slight improvement for the ICA based one.

References

Alter, O., Brown, P., & Botstein, D. (2000). Singular value decomposition for genome-wide expression

data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, pp. 10101–10106.

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., & Liu, C. (2011). Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PloS one*, *6*, e17238.

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

Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., Weiss-Solís, D. Y., Duque, R., Bersini, H., & Nowé, A. (2013). Batch effect removal methods for microarray gene expression data integration: a survey. *Briefings in bioinformatics*, *14*, 469–490.

Renard, E., Branders, S., & Absil, P.-A. (2016). Independent component analysis to remove batch effects from merged microarray datasets. *16th Workshop on Algorithms in Bioinformatics*.

Renard, E., Teschendorff, A. E., & Absil, P.-A. (2014). Capturing confounding sources of variation in dna methylation data by spatiotemporal independent component analysis. *22nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2014)*.

Teschendorff, A. E., Zhuang, J., & Widschwendter, M. (2011). Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics*, *27*, 1496–1505.