ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Spatiotemporal ICA improves the selection of differentially expressed genes

Emilie Renard[1], Andrew E. Teschendorff[2,3] and P.-A. Absil[1] *

1- Université catholique de Louvain - ICTEAM Institute
Avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve - Belgium

2- University College London - Cancer Institute
72 Huntley Street, London WC1E 6BT - United Kingdom

3- PI Computational Systems Genomics - CAS-MPG Partner Institute
for Computational Biology - 320 Yue Yang Road - Shanghai 200031 - China

Abstract. Selecting differentially expressed genes with respect to some phenotype of interest is a difficult task, especially in the presence of confounding factors. We propose to use a spatiotemporal independent component analysis to model those factors, and to combine information from different spatiotemporal parameter values to improve the set of selected genes. We show on real datasets that the proposed method allows to significantly increase the proportion of genes related to the phenotype of interest in the final selection.

1 Introduction

We address the problem of selecting genes differentially expressed with some phenotype of interest in large genomic datasets. The approach presented here uses matrix factorization to model the different sources of variations, including both biological and confounding sources of variation. The latter are known as *batch effects* and occur in large-scale genomic datasets that aggregate measurements obtained under different technical conditions such as reagent quality, laboratory temperature, or chip. Batch effect removal is particularly challenging due to the many possible sources of variation that are unknown or only partly known through limited information, such as batch number and processing date. Including those confounding factors in the modelling is however of critical importance, as not doing so may adversely affect the validity of biological conclusions drawn from the datasets [1, 2, 3, 4].

The genomic data we focus on here are gene expression data, which are now more and more used to explore biological questions such as detecting differentially expressed genes, or predicting sample class. Each database thus takes the form of a p-by-n feature-by-sample matrix X, where p (the number of genes) is typically around 20 000 and n (the number of individuals in the dataset) a few hundred.

*This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office.

A popular approach to model batch effects, as well as other technical and biological artefacts, is to use independent component analysis (ICA) [5, 3]. Applying ICA methods to a feature-by-sample matrix X yields a decomposition

$$X \approx AB^T = \sum_{k=1}^{K} A_{:,k} \left(B_{:,k}\right)^T \tag{1}$$

where $A_{:,k}$ (the kth column of A) can be interpreted as the gene activation pattern of component k and $B_{:,k}$ as the weights of this pattern in the samples.

When computing this decomposition, the question arises whether one should minimize the mutual information between the columns of A or those of B. In [6, 4], a continuum between the "A" and "B" options was investigated using a "spatiotemporal" ICA method based on joint diagonalization of cumulant matrices. The method was validated in [6] by assessing if known potential confounding sources were correlating with $B_{:,k}$ components. Depending on the dataset and on the confounding factors, a better recovering of those factors could be attained by modifying the spatiotemporal trade-off α.

The question is then: if there is no optimal value of α, can we aggregate information from different α values? Moreover, an important aspect in such methods is stability: a good algorithm should give similar results in similar conditions. In this work, building on [6] and [4], we study how combining information from different α values may help to improve the list of selected genes.

The paper is organized as follows. Section 2 presents the different steps of the method used, which is validated in Section 3, and conclusions are drawn in Section 4.

2 Selection of differentially activated genes

Combining ideas from [6] and [4], we now present the method that we use to select differentially expressed genes. As in [4], the underlying model assumes that the expression level of a gene is the result of interactions between different phenomena, biological or physical, where one in particular is of interest (the phenotype of interest, or POI) and the others are not. We cannot know exactly all those phenomena, but we can have an idea of the behavior of some of them by means of available external information (confounding factors, or CF) such as chip type and reagent quality. Under the assumption that those interactions are linear, the model can be written as:

$$\underbrace{X_{g,:}}_{\substack{activation\\ levels\\ of\ gene\ g}} \quad = \quad f_g(\underbrace{\mathbf{y}^T}_{\substack{phenotype\\ of\ interest}}) \quad + \sum_{k=1}^{K} \mathbf{A}_{gk} h_k(\underbrace{\mathbf{r}_k^T}_{\substack{confounding\\ factors}}) \quad + \quad \underbrace{\epsilon_g,}_{noise} \qquad g = 1, ..., p.$$

The $f(x)$ notation, used throughout this paper, stands for a linear prediction model based on the knowledge of variables x. Since we do not know exactly the impacts $h_k(\mathbf{r}_k^T)$ of the confounding phenomena, we try instead to find a basis of surrogate variables v_k that spans the same space, with possible insights from CFs:

$$X_{g,:} \;=\; f_g(\mathbf{y}^T) \;+\; \sum_{k=1}^{K} \mathbf{A}_{gk} \underbrace{\mathbf{v}_k^T}_{\substack{surrogates\\ variables}} \;+ \epsilon.$$

Under the assumption of linear interactions, variables v_k can be inferred from X and y by means of a matrix factorization which is done here using the spatiotemporal ICA developed in [6]. The interest is that, as shown in [4], varying the spatiotemoral tradeoff represented by the parameter α can potentially improve the modeling of confounding factors.

Once the surrogate variables are built from X in an unsupervised way, we can check if they do not correlate too much with the phenotype of interest. If the p-value of association between a surrogate variable and the POI is below some threshold (typically under 0.05), and if the surrogate variable is not more strongly associated with a confounding factor, then this surrogate variable is discarded.

The final objective is to find a subset of genes differentially expressed with the POI. If gene g is differentially expressed with the POI y, then a good model of its behavior has to include y in the variables. That is we must have that $\hat{X}^{(1)}_{g,:} = f(y, v_1, ..., v_K)$ gives a better approximation of $X_{g,:}$ than $\hat{X}^{(2)}_{g,:} = f(v_1, ..., v_K)$. The significance of the difference between both predictions is evaluated by computing associated p-values. As many genes implies many statistical tests, the p-values are corrected in the corresponding q-values [7].

In this paper, to improve the stability of the list of selected genes, we compare all sets of selected genes obtained for different values of the tradeoff parameter α. We retain as final selection the genes present in most of those lists.

3 Validation

To validate our approach, we tested it on breast cancer expression. We combined different datasets which can be accessed under GEO numbers GSE2034 [8], GSE5327 [9], GSE7390 [10], GSE2990 [11], GSE3494 [12], GSE6532 [13] and from [14, 15].[1] As in [3], we took histological grade as the phenotype of interest, and oestrogen receptor status and tumor size as potential confounders. We removed all samples with missing information about grade, oestrogen receptor status or tumor size; which gives a combined dataset of 1473 samples for 22 282 genes.

To evaluate the final output, that is the list of selected genes, we compared it with lists of genes known to be differentially activated with the POI (or CF) using a hypergeometric test. If we know the total number of genes N and

[1] We used the dataset from the *breastCancerNKI* R package, available on http://bioconductor.org/packages/breastCancerNKI/

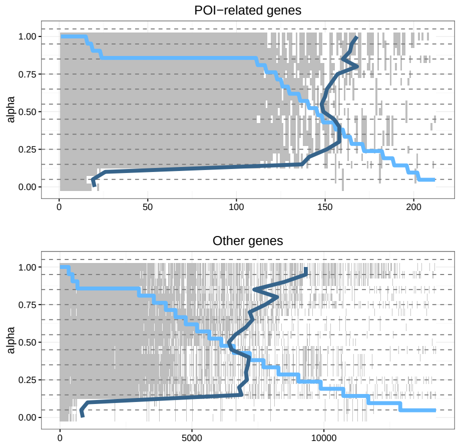

Fig. 1: Selection of POI-related and other genes, depending on α (showing only genes selected at least once). A grey tile indicates that the gene was selected for the corresponding α value. The light blue curve shows the number of selections of each gene (in %). The dark blue curve shows the number of genes selected for each α value.

among them the total number of POI-related genes K, the hypergeometric test computes the probability that we observe k or more POI-related selected genes if we select n genes at random.

Figure 1 shows the list of genes with a q-value under 0.001 for 21 equispaced values of the spatiotemporal parameter α between $\alpha = 0$ (mutual information minimized solely on the first factor A in Equation 1) and $\alpha = 1$ (solely on the second factor B). The main observation is that the list of selected genes differs from one α value to another, but with significant overlaps. Moreover, this overlap is proportionally bigger for POI-related genes than for others genes: $\sim 43\%$ of all POI-related genes are selected in 85% of cases, against $\sim 14\%$ for other genes. If we compare the number of genes selected in 85% of cases to the median over α of the number of genes selected (dark blue line), the proportions are respectively about 72% and 32% for POI-related and other genes. So returning the most stable genes across α values clearly improves the gene selection: the number of POI-related selected genes is reduced, but the number of non POI-related selected genes decreases much more. The proportion of selected genes that are POI-related nearly doubles with this method, which we term $\texttt{aggr}_1$.

Figure 2 shows a comparison of the proposed $\texttt{aggr}_1$ method with three other methods:
(i) `ISVA`, the method proposed in [4], which returns the list obtained with $\alpha = 0$;
(ii) `LR`, where the gene selection is based on a linear regression with the POI only;
(iii) `LR+CF`, where the gene selection is based on a linear regression with the POI and the CFs.

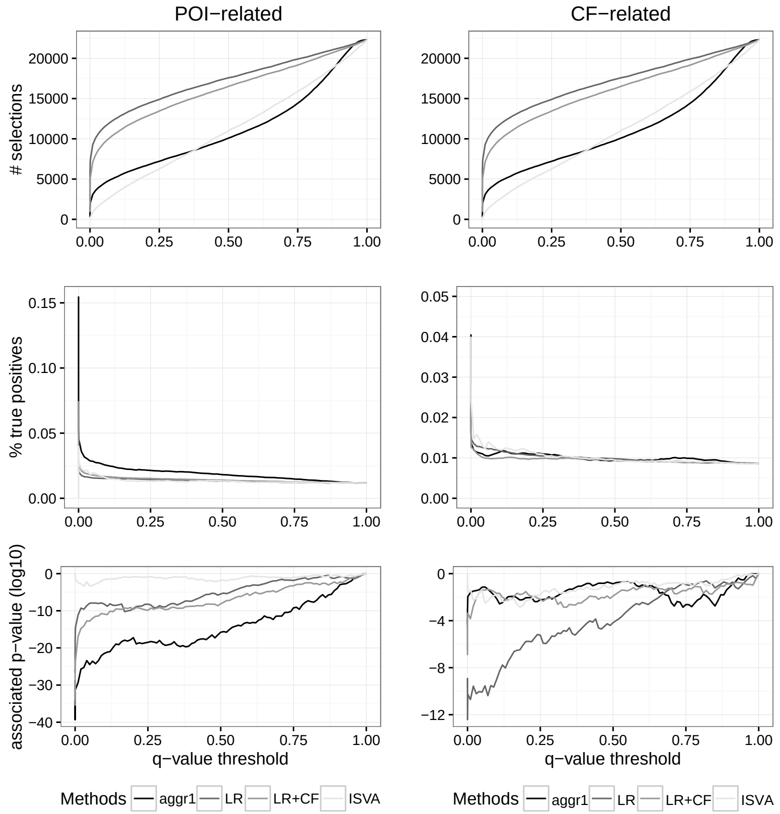

Fig. 2: Number of genes selected, proportions of POI or CF-related genes and associated p-values depending on the selection threshold.

For a same q-value threshold, the methods select different numbers of genes. Unsurprisingly, $\texttt{aggr}_1$ selects among the lowest number of genes. All methods select similar proportions of CF-related genes for q-values thresholds larger than 0.001. However, due to the high number of selected genes for LR, this proportion implies a much smaller p-value. Clearly, adding CF information in the model will improve the POI/CF distinction as can be seen when comparing p-values of the different models. We can see that for thresholds between 0.001 and 0.3, the proportion of POI-related genes among the selected genes is between 100% and 40% higher for our method compared to the others. This higher proportion is translated in a smaller associated p-value.

4 Conclusion

Building on the approach of [4] and the spatiotemporal ICA proposed in [6], we proposed to keep only the intersection of lists obtained from different values of the spatiotemporal parameter α to improve the set of selected genes. We tested our method on a large dataset of breast cancer expressions and showed that the proportion of genes selected for all α values is bigger for POI-related values than for other genes. It can then be used to decrease the proportion of non-POI related selected genes.

References

[1] J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):e161, 2007.

[2] J. T. Leek et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet.*, 11(10):733–739, 2010.

[3] A. E. Teschendorff, J. Zhuang, and M. Widschwendter. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics*, 27(11):1496–1505, 2011.

[4] A. E. Teschendorff, E. Renard, and P.-A. Absil. Supervised normalisation of large-scale omic data sets using blind source separation. In Ganesh R. Naik and Wenwu Wang, editors, *Advances in Modern Blind Source Separation Techniques: Theory and Applications*. Springer, 2013.

[5] Kong et al. A review of independent component analysis application to microarray gene expression data. *Biotechniques*, 45(5):501, 2008.

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

[7] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[8] Y. Wang et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460):671–679, 2005.

[9] A. J Minn et al. Lung metastasis genes couple breast tumor size and metastatic spread. *Proceedings of the National Academy of Sciences*, 104(16):6740–6745, 2007.

[10] C. Desmedt et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clinical cancer research*, 13(11):3207–3214, 2007.

[11] C. Sotiriou et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, 98(4):262–272, 2006.

[12] L. D Miller et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38):13550–13555, 2005.

[13] S. Loi et al. Definition of clinically distinct molecular subtypes in estrogen receptor–positive breast carcinomas through genomic grade. *Journal of clinical oncology*, 25(10):1239–1246, 2007.

[14] Van De Vijver et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.

[15] Van't Veer et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536, 2002.