ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Capturing confounding sources of variation in DNA methylation data by spatiotemporal independent component analysis

Emilie Renard[1], Andrew E. Teschendorff[2] and Pierre-Antoine 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

Abstract.
Confounding sources of variation, which are often either unknown or known with error, are widespread in genomic datasets, and failing to adjust for them may adversely impact statistical inference. In this context, we propose a "spatiotemporal" independent component analysis method that possesses a novel invariance property, and we show that that spatiotemporal aspect may increase the ability of the method to model confounding sources of variation.

1 Introduction

We address the problem of clearing (as much as possible) large genomic datasets from confounding sources of variation, known as *batch effects*, without (too much) removing biological variations of interest. Batch effects 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 variations that are unknown or only partly known through limited information, such as batch number and processing date. Removing those confounding factors from genomic data 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 DNA methylation data, which have been generating much interest in view of associations with diseases such as cancer, diabetes, and Alzheimer's [4]. Each database thus takes the form of a p-by-n feature-by-sample matrix X, where p (the number of methylation sites) is typically around 20 000 and n (the number of individuals in the dataset) a few hundreds.

A popular approach to address batch effects, as well as other technical and biological artefacts, is surrogate variable analysis (SVA) [1]. Recent developments in SVA have shown that replacing the underlying principal component

*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.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

analysis (PCA) by independent component analysis (ICA)—yielding a method dubbed ISVA—improves identification of confounders and subsequent statistical inference [3].

Applying ICA methods (see, e.g., [5, 6, 7] for the concepts or [8, 9, 10, 11] for applications in genetics) to a feature-by-sample matrix X yields a decomposition

$$X \approx AB^T = \sum_{k=1}^{K} A_{:,k} B_{:,k}^T$$

where $A_{:,k}$ 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. Both options are justifiable a priori [12]. However, if there are reasons to believe that the technical factors and the factors of interest are less conditionally dependent given the feature, resp. given the sample, then the "A", resp. "B" option, looks more promising. In genomic data, the former seems a priori more realistic than the latter, since large databases aggregate cohorts of individuals rather than groups of genes. Moreover, in earlier times, the very vertical shape of matrix X in genetic datasets—small number of individuals compared to the number of genes—naturally pointed to the "A" option, since A contains many more observations (rows) than B on which an estimator of mutual information can be applied; but the validity of this argument has waned with the rise of larger databases.

In [12], a continuum between the "A" and "B" options was investigated on gene expression databases using a "spatiotemporal" ICA method based, in the spirit of the JADE method [13, 14], on joint diagonalization of cumulant matrices. The method was validated in [12] by assessing if known biological pathways were enriched in columns of A, and it was concluded that a markedly better enrichment may be observed at intermediate points of the continuum. However, biological pathways represent fuzzy objects and so are not ideal for the purpose of method evaluation.

In this work, building on [12] and [4, §7], we study how the tradeoff between favoring independence on A versus B may affect the ability of an ICA method to model confounding factors. Specifically, we investigate how the tradeoff affects the correlation between each column of B and the beadchip. We focus on beadchip effects because this provides a more objective framework in which to evaluate ICA methods. This is because it is known which samples have been done on which beadchip and beadchip effects normally affect all samples on the chip. Hence, the confounder is known and the modeling of it by the BSS algorithm can be more objectively assessed. The presence of a high correlation between the beadchip and certain components suggests that cleaner data may be obtained by removing those components from the data.

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

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

2 An unbiased orthogonal spatiotemporal ICA method

We now present the ICA method that we use to generate matrices A and B from the data matrix $X \in \mathbb{R}^{p\times n}$. The algorithm depends on a *spatiotemporal parameter* $\alpha \in [0,1]$ that allows it to explore a continuum between imposing independence solely on A ($\alpha = 0$) and solely on B ($\alpha = 1$). The term "spatiotemporal" comes from the pixel-by-time data in medical imaging for which the concept was introduced [15]. For lack of space, we abundantly refer to previous literature on spatiotemporal ICA and we emphasize the novel unbiased orthogonal aspect.

The first step consists of centering the feature-by-sample data matrix X by subtracting the row and column means, followed by a dimensionality reduction by means of a K-truncated SVD, yielding a new matrix $\tilde{X} = U_K D_K V_K^T$. All the possible two-factor decompositions of $\tilde{X}$ are given by $\tilde{X} = AB^T$ where $A = U_K D_K W^{-1}$ and $B^T = W V_K$ with W a $K \times K$ invertible matrix. In [12] and [4, §7], as is customary in ICA methods, W was restricted to the orthogonal group $\mathrm{O}(K) = \{W \in \mathbb{R}^{K\times K} : W^T W = I\}$. A drawback of this restriction is that further imposing decorrelatedness of the columns of A (i.e., imposing $A^T A$ diagonal) reduces very much the freedom in W, whereas this property is a natural requirement when independence is sought in A. We remedy this drawback by considering instead the decomposition

$$\tilde{X} = \underbrace{U_K D_K^{\alpha} W^{-1}}_{=:A} \underbrace{W D_K^{1-\alpha} V_K^T}_{=:B^T}, \quad W \in \mathrm{O}(K). \tag{1}$$

Consequently, the columns of A, resp. B, are structurally decorrelated when $\alpha = 0$, resp. $\alpha = 1$.

As in [4, §7], in the spirit of the JADE ICA algorithm [14], we seek W in $\mathrm{O}(K)$ that minimizes a finite-data contrast function of the form

$$f_\alpha(W) = \alpha \sum_i \mathrm{Off}(C_i(B^T)) + (1-\alpha) \sum_i \mathrm{Off}(C_i(A^T)),$$

where A and B depend on W through (1), $\mathrm{Off}(Y)$ returns the sum of squares of the off-diagonal elements of Y, and the C_i's are fourth-order cumulant matrices, satisfying the property $C_i(WM) = W C_i(M) W^T$. The minimization of f_α is thus a joint approximate diagonalization problem, which we address as in JADE using Jacobi rotations. We initialize the Jacobi algorithm with $W = I$, ensuring that both A and B initially have decorrelated columns.

3 Validation

We consider various DNA methylation databases where the samples are distributed over different beadchips with a maximum of 12 samples per chip. For each database, we feed the p-by-n feature-by-sample matrix X, the spatiotemporal parameter α, and the number of components K to the spatiotemporal ICA

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

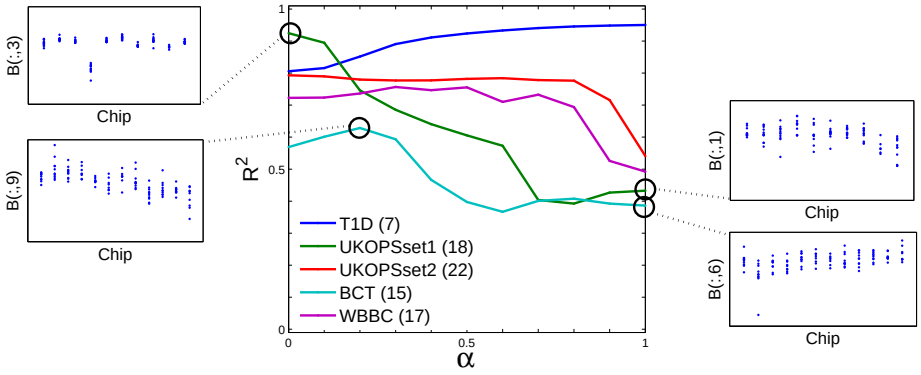

Fig. 1: Central plot: for the various databases, maximal R^2 value with beadchip number as independent variable, as a function of α. The number K of components extracted is indicated between parentheses. The smaller graphics show how some R^2 values were achieved.

algorithm outlined in Section 2. The algorithm returns matrices $A \in \mathbb{R}^{p\times K}$ and $B \in \mathbb{R}^{n\times K}$.

In order to appraise the ability of the ICA algorithm to model confounding sources of variations, we measure the correlation between each column of B and a known source of confounding, namely beadchip variations. To this end, let $c(s)$ denote the beadchip on which sample s was profiled. This assigns a category to each sample, with up to 12 samples per category for the considered databases. For $i = 1, \ldots, K$, let $B_{:,i}$ denote the ith column of B, $\bar{B}_{c(s),i}$ denote the mean value of $\{B_{s',i} : c(s') = c(s)\}$, and $\bar{B}_{:,i}$ denote the mean value of $\{B_{s,1} : s = 1, \ldots, n\}$. The correlation between $B_{:,i}$ and beadchip number is then defined to be $R^2(B_{:,i}, p) = 1 - \frac{\sum_s \left(B_{s,i} - \bar{B}_{c(s),i}\right)^2}{\sum_s \left(B_{s,i} - \bar{B}_{:,i}\right)^2}$. An R^2 close to 1 means a high correlation between $B_{:,i}$ and beadchip number, revealing that $B_{:,i}$ is strongly affected by beadchip and thus presents a potential for modeling beadchip-related confounding sources of variation.

Tests were performed on DNA methylation datasets UKOPSset1, UKOPSset2, T1D, WBBC [4] and BCT [16]. In the central plot of Figure 1, for each database, the best R^2 value found over the K columns of B is given versus the value of the spatiotemporal parameter α. The R^2 value is high in most cases, showing that at least one of the columns of B strongly correlates with beadchip. The trend is for R^2 to decrease as α goes from 0 to 1, as could be expected (see Section 1). However, the opposite trend is observed for database T1D, revealing that imposing independence on B has value for this database.

The smaller plots on Figure 1 plot the column of B that achieves the largest R^2 value against beadchip number, for a dataset and a value of α that can be read on the central plot. These smaller plots allow us to visualize how the R^2 value was achieved.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

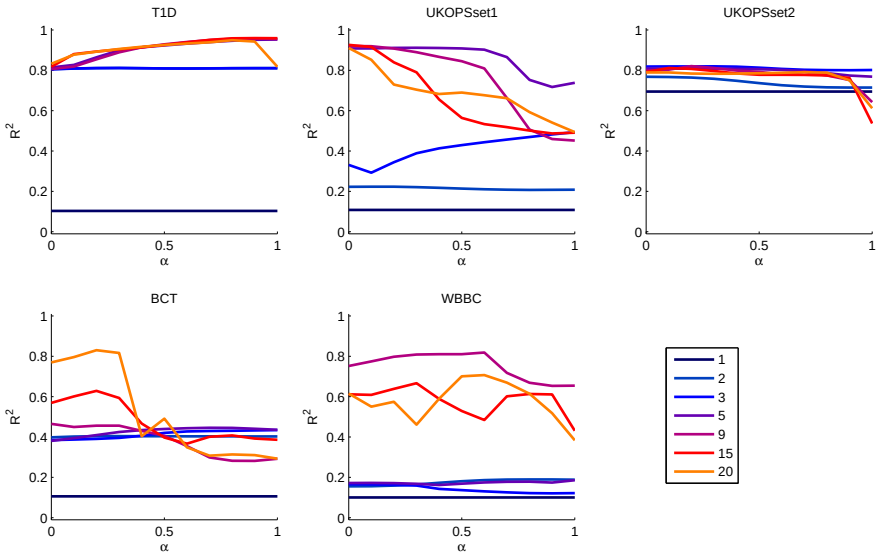

Fig. 2: Maximal R^2 values of the columns of B in a linear ANOVA model with beadchip number as independent variable, plotted versus α for different real datasets, with the number K of components varying.

In Figure 1, the number K of components was estimated based on random matrix theory, as done in [4]. In other experiments reported in Figure 2, we investigate how the maximal R^2 value is influenced by the number K of components extracted. We observe that R^2 for fixed α is seldom monotonically increasing with respect to K. As a consequence, choosing K adequately is an important issue, and a comparison between Figures 1 and 2 reveals that random matrix theory produces a reasonable choice of K in our experiments. One also observes that R^2 does not necessarily depend monotonically on α, suggesting that the continuum between imposing independence fully across genes and across samples is worth exploring.

4 Conclusions and perspectives

This paper has contributed beyond [12] and [4, §7] chiefly in two ways. On the algorithmic side, we have eliminated the time- and space-biases discussed in [12] while staying within the orthogonal ICA framework. As a consequence, the new algorithm enjoys the following invariance property that is seemingly not present in other spatiotemporal ICA methods: if the input (X, α) yields the output (A, B), then the input $(X^T, 1-\alpha)$ yields the output (B, A). In other words, the temporal and spatial flavors are treated on an equal footing. On the experimental side, we have observed that the strongest correlations are

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

not systematically obtained with the "A" option. In forthcoming work, we will thus investigate how the "B" option perform within the ISVA framework. Alternatives to the JADE approach to ICA, or even other matrix factorization methods such as the ones described in [17], are also worth testing.

References

[1] J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, 3(9):1724–1735, 2007.

[2] J. T. Leek, R. B. Scharpf, H. C. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry. 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. Accepted.

[5] P. Comon. Independent component analysis, A new concept? *Signal Processing*, 36:287–314, 1994.

[6] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[7] P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Independent Component Analysis and Applications Series. Elsevier Science, 2010.

[8] W. Liebermeister. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18(1):51–60, 2002.

[9] S.-I. Lee and S. Batzoglou. Application of independent component analysis to microarrays. *Genome Biology*, 4:R76, 2003.

[10] A. E. Teschendorff, M. Journée, P.-A. Absil, R. Sepulchre, and C. Caldas. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Comput. Biol.*, 3(8):1539–1554, 2007.

[11] W. Kong, C. R. Vanderburg, H. Gunshin, J. T. Rogers, and X. Huang. A review of independent component analysis application to microarray gene expression data. *BioTechniques*, 45:501–502, 2008.

[12] M. Sainlez, P.-A. Absil, and A. E. Teschendorff. Gene expression data analysis using spatiotemporal blind source separation. In *ESANN*, pages 159–164, 2009.

[13] J.F. Cardoso and A. Souloumiac. Blind beamforming for non-gaussian signals. *IEE Proceedings - F*, 140(6):362–370, 1993.

[14] J.-F. Cardoso. High-order contrasts for independent component analysis. *Neural Comput.*, 11(1):157–192, 1999.

[15] J. V. Stone, J. Porrill, N. R. Porter, and I. D. Wilkinson. Spatiotemporal independent component analysis of event-related fMRI data using skewed probability density functions. *NeuroImage*, 15(2):407–421, 2002.

[16] J. Zhuang et al. The dynamics and prognostic potential of dna methylation changes at stem cell gene loci in women's cancer. *PLoS Genet*, 8(2), 2012.

[17] A.V. Kossenkov and M.F. Ochs. Matrix factorisation methods applied in microarray data analysis. *International Journal of Data Mining and Bioinformatics*, 4(1):72–90, 2010.