

# Chapter 17

## Supervised Normalization of Large-Scale Omic Datasets Using Blind Source Separation

Andrew E. Teschendorff, Emilie Renard and Pierre A. Absil

**Abstract** Biotechnological advances in genomics have heralded in a new era of quantitative molecular biology whereby it is now possible to routinely measure over tens of thousands of molecular features (e.g., gene expression levels) in hundreds if not thousands of patient samples. A key statistical challenge in the analysis of such large omic datasets is the presence of confounding sources of variation, which are often either unknown or only known with error. In this chapter, we present a supervised normalization method in which Blind Source Separation (BSS) is applied to identify the sources of variation, and demonstrate that this leads to improved statistical inference in subsequent supervised analyses. The statistical framework presented here will be of interest to biologists, bioinformaticians and signal processing experts alike.

### 17.1 Introduction

Omic and sequencing technologies have revolutionized the biomedical field [40]. With these technologies, it is now possible, at a reasonable economic cost, to measure the levels of molecular entities, for instance, gene expression, genome-wide,

---

A. E. Teschendorff (✉)

Statistical Cancer Genomics, UCL Cancer Institute, 72 Huntley Street, London WC1E 6BT, UK  
e-mail: a.teschendorff@ucl.ac.uk

A. E. Teschendorff

CAS-MPG Partner Institute for Computational Biology, Chinese Academy of Sciences, Shanghai Institute for Biological Sciences, 320 Yue Yang Road, Shanghai 200031, China

E. Renard · P. A. Absil

Department of Mathematical Engineering, ICTEAM Institute, Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium  
e-mail: emilie.renard@uclouvain.be

P. A. Absil

e-mail: absil@inma.ucl.ac.be

in cellular specimens from large numbers of patients [8]. Analysis of these large genomic, more generally referred to as “omic”, datasets promises to provide the advances and biomarkers, which are urgently needed in the biomedical field, heralding in the new age of personalized medicine [34]. However, a serious obstacle in translating these mammoth amounts of data into biomedical advances is the presence of confounding factors, both technical and biological [21]. Recent studies [21, 43] have shown that technical confounding factors, generally referred to as batch effects, for instance the date in which a sample was processed, are omnipresent in omic datasets, affecting even some of the highest-profile studies such as The Cancer Genome Atlas [46], or the 1,000 Genomes Project [7]. Some estimates indicate that in any given study up to 80% of measured molecular features can correlate with unwanted technical factors [21]. Furthermore, not adjusting for confounding factors can adversely impact statistical inference, compromising sensitivity and specificity [20, 45].

There are many reasons why these batch effects arise. Specially, in the case of large-scale studies profiling hundreds to thousands of samples, samples will inevitably have been processed on either different dates, by different laboratories or personnel, or on different plates or chips. Laboratory conditions can vary between dates affecting the biological measurements, or the quality of the profiling technology may also vary significantly from batch to batch. Moreover, profiled samples may come from patients treated at different medical centers, and therefore the way samples were handled (e.g., time from sampling to storage) may introduce further variation (see e.g., [25]). All of these factors have been shown to introduce unwanted variation in the data, and since “*the more you measure the more can go wrong*”, it is clear that large scale studies are particularly vulnerable to such confounding factors. On the other hand, it is worth pointing out that large-scale studies are also much better placed than small sample-size studies at adjusting for confounding factors. For instance, it is easier to detect and subsequently correct for a single chip/plate effect if there are many other chips/plates in the study that have performed well since the latter can then serve as controls.

The statistical design of a study is of critical importance in trying to prevent the potentially adverse effects of confounding factors on downstream statistical inference. Clearly, the statistical design of a study must be such so as to ensure that a number of specific research questions can be properly addressed. This typically requires that samples be distributed randomly across batches, ensuring balanced numbers of specific phenotypes across them. Thus, in comparing phenotypes A and B, one would randomize these across batches ensuring balanced numbers of A and B in each batch. However, it is not unusual for unbalanced designs to arise as a result of samples dropping out, in turn caused by logistical or quality control issues. This is particularly true for large-scale studies where logistical or quality control issues almost inevitably arise. These unbalanced designs can then have a dramatic negative impact on statistical inference if adjustment for the technical sources of variation is not performed. Thus, (large-scale) studies with an initial perfect study design may still be hampered by confounding factors.

There are a number of other key issues to mention in connection with confounding factors. First, it is clear that the potential impact of confounding factors will depend on the signal-to-noise ratio. This in turn depends on numerous study-specific factors, including the phenotype of interest, the nature of the confounding variation and the tissue type being profiled. For instance, if one is measuring DNA methylation, a covalent modification of DNA that can affect the activity of nearby genes [9], and if the comparison is between normal and cancer tissue, then it is likely that batch effects can be ignored, since DNA methylation changes associated with cancer are generally of a large magnitude (high signal-to-noise ratio limit) [46]. On the other hand, if the Epigenome-wide Association Study (EWAS) [31] measuring DNA methylation is being conducted in whole blood tissue [24], then this is likely to involve small effect sizes in relation to the technical sources of variation like chip effects, or biological factors such as age. For instance, in Rakyan et al. [31], the authors report a genomic site with a DNA methylation pattern in whole blood that correlates with smoking status, involving small 5–10% shifts in average methylation between cases and controls. Such 5–10% shifts could in principle be also caused by batch/chip effects. Similarly, such small shifts in average DNA methylation levels could be due to relatively small changes in blood cell type composition, which in turn could be caused by differences in the age of the sampled individuals [43]. Thus, techniques like Singular Value Decomposition (SVD) are specially useful for omic data since they easily allow approximate relative quantification of the variance associated with different sources of variation [43].

A second important issue is that the way in which statistical inference is affected strongly depends on how the confounders are correlated to the phenotype of interest (POI) [19]. Clearly, a confounding factor which is anti-correlated to a POI will dampen the statistical significance, while positive correlations will lead to overoptimistic results. An orthogonal confounder of large variability in relation to the POI signal will similarly compromise the statistical significance and lead to a large false negative rate (FNR). Thus, when analyzing omic data it is important to be aware of these different potential scenarios and generation of *P*-value histograms is strongly recommended as a means of detecting the strength and type of confounding [19].

Last but not least, confounding sources of variation can be of a very different nature, directly influencing the type of statistical adjustment procedure to be used. For instance, some confounders like plate or date, are examples of known confounders in the sense that we know exactly on which date and on which plate a given sample was processed, as these are factors that are normally recorded in an experiment. In this case, adjustment with (Bayesian) regression models, which use the confounders as explicit covariates, is possible and indeed fairly popular [16]. However, surprisingly often confounders are only known with uncertainty or error. For instance, in DNA methylation studies conducted with the Illumina Infinium beadchips, samples need to be preprocessed using a bisulfite conversion step, which translates epigenetic changes into genetic ones allowing these to be measured on the beadchip [4]. This conversion step is variable between samples and although the conversion efficiency can be measured using control probes on the beadchip, this measurement is subject to error. As another example, we have observed components of variation in DNA

methylation data associated with the season in which samples were collected. Season can be viewed as a surrogate for temperature, which is the more likely causal factor, yet the exact temperature to which the samples were exposed to during transportation from medical centers to the central processing lab was not recorded. At the other extreme, we may have confounders which are completely unknown, or there is no correlated known factor that could be used as surrogate. All these considerations are important in the context of this chapter, because clearly in the latter two scenarios, explicit adjustment for confounders is neither advisable or possible. Hence, BSS techniques are needed to infer these confounders from the data itself. On the other hand, as we shall see, known confounders also become useful in the BSS context, since they can be used to objectively evaluate the quality of blind source separation.

It is paramount to stress again the importance of adjusting for confounding factors, as not doing so can seriously reduce the effective power of the studies, or lead to unacceptably large false discovery rates [21, 45]. Thus, there is an urgent need for powerful statistical methods to be applied in the biomedical field to help address these significant challenges. To further motivate a BSS-based approach to statistical inference, we emphasize that it is only natural to view any biological omic dataset as an interference pattern, with some sources of variation reflecting the biological phenotype of interest, and others reflecting the effects of technical factors. Therefore, BSS methods are optimally placed to infer such sources of variation.

Indeed, BSS methods have already been extensively applied to omic data, but only as a means of performing dimensional reduction to identify *biological* sources of variation [12, 18, 22, 23, 28, 42, 49], and, secondly, as a means of performing feature selection and classification [14]. Specific popular BSS algorithms include Independent Component Analysis (ICA) [15] and non-negative matrix factorisation (NMF) [13], which have been applied to diverse data types, from gene expression [42] to DNA methylation data [51], including even mutational data [1] and multidimensional cancer genomic profiles [50]. The earliest studies already demonstrated that BSS methods like ICA and NMF lead to substantial improvements in modeling biological sources of variation and that these improvements are mainly due to the sparse (supergaussian) nature of the underlying biological sources [18, 42].

In contrast, relatively few BSS applications have focused on the problem of artifact removal in biomedical data, which is surprising given that technical sources of variation are omnipresent in such data and that they can so negatively affect statistical inference. We would also argue that the application of BSS methods to identify and remove technical artifacts in real omic data provides a substantially better framework in which to objectively evaluate BSS algorithms. There are several reasons for this. First, biological sources of variation such as activity of a molecular signaling pathway are “fuzzy” objects and only rarely can be used as defining a ground truth. On the other hand, technical artifacts are sometimes well known to the experimentalist performing the study and hence, as explained above, these can be exploited to assess the quality of BSS separation. Indeed, we recently demonstrated the feasibility of this conceptual framework for assessing BSS methods in a proof-of-principle study, analyzing both DNA methylation and gene expression data [45]. In that work, we proposed an algorithm called Independent Surrogate Variable Analysis

(ISVA), based on ICA, for performing supervised normalization in the presence of confounding factors [45], demonstrating its superiority over non-BSS based alternatives. The main purpose of this chapter is therefore to demonstrate that BSS methods can lead to substantial improvements in statistical inference in large omic datasets, thanks to a more efficient deconvolution of the confounding sources of variation. Our secondary aim is to increase the awareness among the BSS community of the importance of this fairly novel BSS application to artifact removal in biomedical omic data, and thus provide a fertile ground for interdisciplinary cross-pollination.

This chapter is organized as follows. First, because most of the examples considered in this chapter are drawn from studies in DNA methylation, we provide the reader with a brief introduction to DNA methylation and the Illumina Infinium Beadarray technology, a technology that allows genome-wide measurements of this epigenetic mark. In the subsequent section, we provide a number of examples of confounding variation in omic data and describe their negative impact on downstream statistical inference, including examples where methods based on explicit adjustment of confounders cannot be applied. In Sect. 17.3, we describe the problem of performing supervised analysis in the background of confounding factors, introducing and reviewing the SVA framework of Leek et al. [19, 20]. We argue theoretically why SVA may break down and why a BSS method is needed to avoid the pitfalls associated with SVA. This motivates the ISVA algorithm [45], which we review in the next subsection. In Sect. 17.4, we validate ISVA on simulated data and demonstrate the need for adjustment of confounding factors. In Sect. 17.5, we compare ISVA to SVA in modeling beadchip effects in real omic data. Section 17.6 provides a rigorous evaluation of ISVA on eight real omic datasets, using the non-BSS SVA method as well as another method based on explicit adjustment as benchmarks. In the final section, we briefly explore the performance of a generalized BSS algorithm in modeling beadchip effects. We end with conclusions and suggestions for further research.

## 17.2 DNA Methylation and the Illumina Infinium Beadarray Technology

DNA methylation refers to the covalent attachment of a methyl  $\text{CH}_3$  group to DNA cytosines, normally, but not exclusively, in the context of a CG dinucleotide, referred to as a CpG [9]. There are about 30 million of such CpG sites in the human genome, most of which are methylated. These 30 million CpG sites represent in fact an underenrichment of CpGs in the human genome. In some genomic regions however, the density of CpGs is much higher than normal, and these are referred to as CpG islands. Roughly, about 60% of gene promoters fall within CpG islands and most of these are normally unmethylated. Thus, whereas most of the genome is methylated, many of the promoter CpG islands are unmethylated in the normal state.

DNA methylation is important for a number of reasons. It is not only essential for embryonic development, but is also key in developmental processes [9]. Very recently, it has been demonstrated that differentially methylated regions between diverse normal cell types are enriched for transcription factor binding sites, supporting the view that DNA methylation is associated with how accessible the DNA is to transcription factors. Thus, hypomethylation, i.e., loss of DNA methylation, allows transcription factor proteins to more easily bind to DNA in order to initiate developmental differentiation programs. The DNA methylation state at the gene promoter is also a key determinant of the gene's activity, i.e., its gene expression level, with promoter hypermethylation normally associated with gene silencing [9]. DNA methylation is particularly important in diseases like cancer, where it is significantly altered [11, 17]. Indeed, a key cancer hallmark is the hypermethylation of CpG island promoters, whilst most of the cancer genome undergoes widespread hypomethylation. These deregulations in DNA methylation may lead respectively, to underexpression/silencing of key tumor suppressor genes, or overexpression of oncogenes (tumor promoting genes).

DNA methylation can be measured fairly accurately using a number of different technologies. In this chapter, we will be considering DNA methylation data generated using the Infinium beadarray technology from Illumina [4]. In particular, we will be considering a version of this technology, called Infinium 27k, that allows measurement of DNA methylation at over 27,000 CpG sites, mostly located within gene promoters of approximately 14,000 genes. The beadarray consists of a set of probes that interrogate the methylation state at each of these 27,000 sites. For each CpG site, there are two sets of probes, one designed to match the methylated version of the allele, while the other matches the unmethylated version. This is made possible by treating the DNA with bisulfite, prior to hybridisation to the beadarray. During bisulfite conversion, unmethylated cytosines are converted into uracil and then thymine upon DNA amplification (i.e.,  $uC \rightarrow T$ ), whereas methylated cytosines are protected and remain cytosines (i.e.,  $mC \rightarrow C$ ). Thus, an epigenetic difference can be translated into a genetic one, which is then easily measured using probes on the beadarray as described. While the methylation state of a given CpG site in a given diploid cell can take only three values (0 = both alleles unmethylated, 1 = only one of the alleles is methylated, 2 = both alleles are methylated), in practice, measurement is taken over many thousands of cells, with the methylation state also being potentially variable between cells. Hence, methylation at a single CpG site in a given sample taken from an individual is quantified in terms of a  $\beta$ -distributed quantity,  $\beta = M/(U + M)$ , where  $M$  and  $U$  denote the intensities of the methylated and unmethylated versions of the allele, as estimated from the respective probes on the array. By construction, this  $\beta$ -value lies between 0 (unmethylated) and 1 (fully methylated).

A number of important features of the Illumina methylation beadarrays are worth mentioning. First, a maximum of 12 samples can be measured on any given beadchip. As with any technology, the quality of beadchips can vary from batch to batch. Also, the DNA quality of a sample can vary significantly, which would subsequently affect  $\beta$ -value estimates. For these reasons, the beadchips are equipped with a number

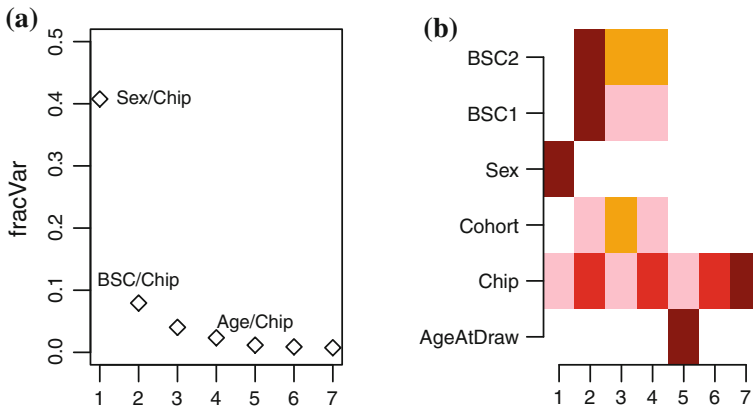
of control probes, each designed to measure the quality of a particular aspect of the assay. For instance, bisulfite conversion efficiency (BSC) could vary between samples, causing biases in the  $\beta$ -values, and this can be assessed using built-in control probes which measure the efficiency of bisulfite conversion.

### 17.3 Confounding Factors in Large-Scale Omic Studies

In order to illustrate the nature and impact of the problem posed by confounding factors, we consider two examples. These examples are taken from two separate DNA methylation studies generated with the Infinium 27k technology. Let us consider our first example. This is a DNA methylation dataset of whole blood samples from 187 individuals with type-1 diabetes, including both sexes, and with individuals drawn from two underlying cohorts. This particular dataset was used to test if DNA methylation changes correlate with the age of the individual at sample draw, thus age is here the POI [44]. The 187 samples were distributed over 17 different beadchips with at most 12 samples per beadchip. A SVD of the  $27,578 \times 187$  row-centered (rows label CpGs) data matrix was performed to assess the nature of the largest sources of variation. As can be seen in Fig. 17.1, it is only the fifth component of variation that correlates with the POI (i.e., age), with the top components correlating with other factors such as sex, BSC and (bead)chip. Furthermore, it can be seen that the fifth component also correlates with chip indicating that this could be a potential confounder. This example further illustrates that technical or other biological variation can be of orders of magnitude larger than the effect size of interest.

As a second example, we consider a DNA methylation dataset of 48 samples, consisting of 30 normal samples from the cervix and 18 representing an intraepithelial cervical neoplasia of grade 2 or higher (CIN2+) (a preinvasive cancer condition). Here too, a SVD on the row-centered data matrix, reveals that it is only the third, fourth, and fifth components that correlate with biological factors such as age or CIN2+ status (Fig. 17.2a–b). Furthermore, unsupervised clustering of the samples does not lead to segregation of the samples according to CIN2+ status, as one would have expected on biological grounds (Fig. 17.2c). This example also illustrates that the top component of variation is correlating with an unknown factor, possibly spatial artifacts on the chips but which are also largely independent of chip. The key point to appreciate here is that there is no surrogate known factor that we can use to model this confounding source of variation, and hence explicit adjustment for this confounder using a multivariate regression model in which the confounder is included as a covariate is not possible [16].



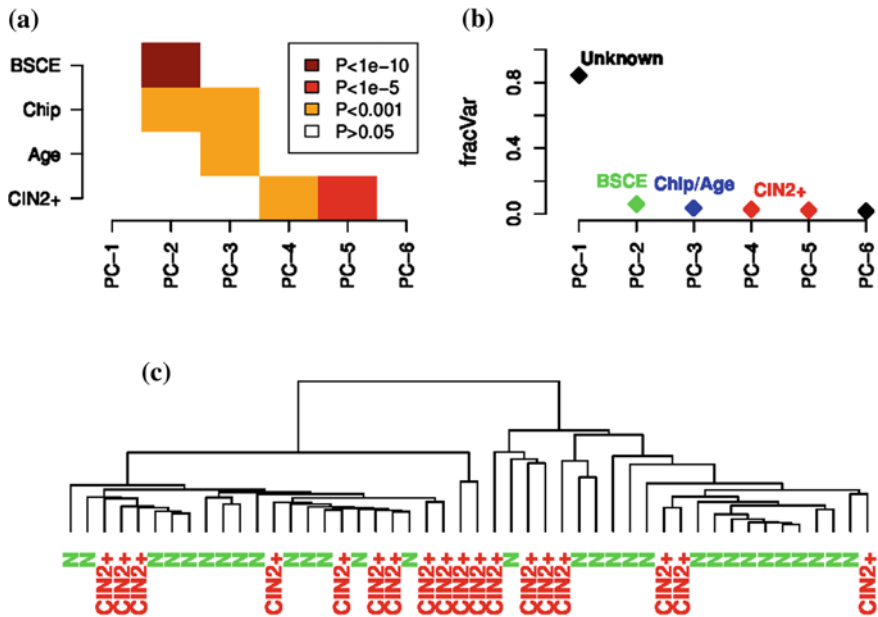


**Fig. 17.1** **a** Relative fraction of variation carried by each of the seven significant singular vectors of a SVD, as measured relative to the total variation in the data. Number of significant singular vectors was estimated using Random Matrix Theory (RMT) [45]. Some of the singular values are labeled according to which confounders the corresponding singular vectors are correlated to, as shown in panel **b**. **b** Heatmap of  $P$ -values of association between the seven significant singular vectors and the phenotype of interest (here age at sample draw) and confounding factors (Chip, cohort, sex, and bisulphite conversion (BSC) efficiency controls 1 and 2).  $P$ -values were estimated using linear ANOVA models in the case of chip, cohort and sex, while linear regressions were used for age and BSC efficiency. Color codes:  $P < 1e-10$  (brown),  $P < 1e-5$  (red),  $P < 0.001$  (orange),  $P < 0.05$  (pink),  $P > 0.05$  (white)

## 17.4 Supervised Normalization by SVA and ISVA

The previous examples illustrate some of the difficulties that confounding factors can pose in statistical analyses. One of the common tasks in omic data analysis is to perform a supervised analysis in which we seek to identify features associated with a phenotype of interest. Clearly, such task may be compromised by the presence of confounding factors, specially if the confounder is unknown or if it is only known subject to error, since in these cases we can't adjust for them explicitly. Thus, one desires a statistical framework in which to perform supervised analysis (i.e., feature selection) in the presence of uncertain or unknown confounding factors. We refer to this supervised analysis problem as “supervised normalization” in the sense that the normalization of the data is performed as part of the supervised analysis and is therefore dependent on the phenotype of interest. So far, only two algorithms, SVA [19, 20] and ISVA [45] have been proposed to address this problem in the context of omic data, where by definition the number of features is relatively large.





**Fig. 17.2** Confounding variation in a DNA methylation dataset of 30 normal cervical samples and 18 cervical intraepithelial neoplasias of grade 2 or higher (CIN2+). **a** Relative fraction of variation carried by each of the six significant singular vectors of a SVD, as measured relative to the total variation in the data. Number of significant singular vectors was estimated using Random Matrix Theory (RMT) [45]. Some of the singular values are labeled according to which confounders the corresponding singular vectors are correlated to, as shown. **b** Heatmap of  $P$ -values of association between the six significant singular vectors and the phenotypes of interest (here CIN2+ status and age at sample draw) and confounding factors (Chip and bisulphite conversion efficiency (BSCE)).  $P$ -values were estimated using linear ANOVA models in the case of chip and CIN2+ status, while linear regressions were used for age and BSC efficiency. Color codes:  $P < 1e - 10$  (brown),  $P < 1e - 5$  (red),  $P < 0.001$  (orange),  $P > 0.05$  (white). **c** Hierarchical clustering of the 48 samples over the 5,000 most variable probes

### 17.4.1 Surrogate Variable Analysis

Leek and Storey proposed an ingenious solution to the problem posed above, known as SVA [19, 20], which we now describe. Let us assume that we have a data matrix,  $X_{ij}$ , with  $i$  ( $i = 1, \dots, p$ ) labeling the features (genes, CpGs,...) and  $j$  ( $j = 1, \dots, n$ ) labeling the samples, with  $p \gg n$ . Furthermore, we assume that each row of  $X$  has been mean centered, and that we have a POI encoded by a vector  $\mathbf{y} = \{y_1, \dots, y_n\}$ . As in [20] we may allow for a general function of the phenotype vector, so that the starting model for SVA takes the form

$$X_{ij} = f_i(y_j) + \varepsilon_{ij}. \tag{17.1}$$

Typically,  $f_i(y)$  would be a function of the form  $f_i = b_i F(y)$  with  $b_i$  a feature specific regression parameter (to be estimated) and  $F$  representing a general link function. Thus, SVA starts by performing univariate regressions, leading to estimates  $\hat{b}_i$  as well as an estimate of the error matrix  $\epsilon$ , which we shall call the residual variation matrix,  $R \equiv \hat{\epsilon}$ . Componentwise,  $R_{ij} \equiv X_{ij} - \hat{f}_i(y_j)$ . SVA then proceeds by performing a SVD of the residual variation matrix

$$R = UDV^T. \quad (17.2)$$

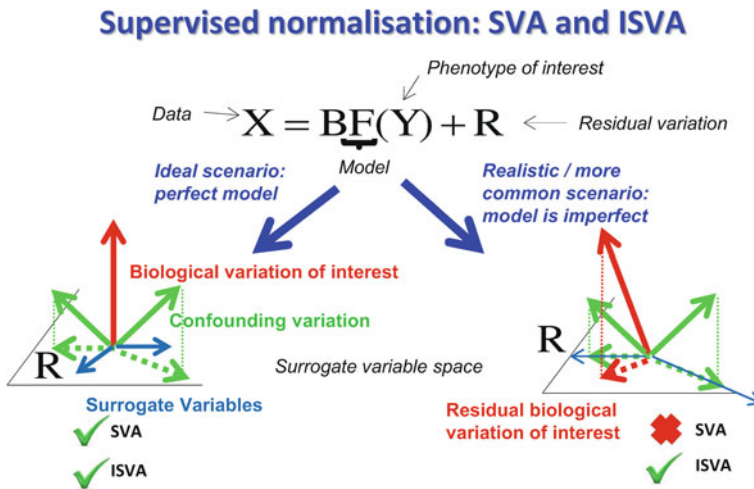
Thus, the singular vectors of the SVD capture variation which is orthogonal to the variation associated with the POI. This residual variation is therefore likely to be associated with other biological factors, not of direct interest, or with experimental factors, all of which constitute potential confounders. SVA provides a prescription for the construction of surrogate variables,  $v_k$  ( $k = 1, \dots, K$  with  $K < n$ ), in terms of the singular vectors (i.e., the column vectors of  $V$ ) of this SVD [20]. In the final step, feature selection is performed using the modified regression model

$$X_{ij} = f_i(y_j) + \sum_{k=1}^K \lambda_{ki} v_{kj} + \epsilon'_{ij}. \quad (17.3)$$

with the rows of  $\epsilon'$  now uncorrelated [19].

In the above framework, it is key to realize that SVA hinges on a big assumption, which is that we have a perfect, or at least a sufficiently accurate model  $F(y)$  describing the data, such that the residual variation encapsulated by the matrix  $R$  does not contain any biological variation of interest (see left part of Fig. 17.3). In this case, the only requirement on the surrogate variables describing the confounding variation is that they span the residual variation space. We note that there is in fact no requirement for the surrogate variables (SVs) to align with (i.e., precisely model) the confounding factors.

However, now consider an alternative, and, as we shall see later, a more realistic scenario, where model  $F(y)$  is imperfect. For instance, we may be using a linear function  $F$  when the relation between data and POI is highly nonlinear. In this case, residual biological variation of interest may be present in  $R$  (see right part of Fig. 17.3). In such a scenario, we would want our SVs to align with the confounding factors and not with the residual biological variation, since otherwise inclusion of this in the subsequent adjusted supervised analysis (Eq. 17.3) would lead to a reduced biological signal. Later we shall see examples of this happening. Hence, in this more realistic scenario, we need to choose SVs that span a *subspace* of  $R$ , i.e., one that is also orthogonal to the residual biological variation. This in turn means that we need an algorithm that can more accurately deconvolve the confounding sources from the residual biological variation. As one might expect (and we shall see examples of this later), the SVD used in SVA can not accurately deconvolve these different sources of variation. This motivates the introduction of BSS methods in the context of supervised normalization.



**Fig. 17.3** Surrogate Variable Analysis (SVA) begins by performing a regression of the data matrix,  $X$ , against the phenotype of interest,  $Y$ , specified through a possibly nonlinear function  $F(Y)$ . In the equation above,  $B$  denotes regression parameters, whereas  $R$  denotes the residual variation, i.e., the variation in the data not explained by the phenotype of interest under the specified model  $F$ . Under such a model, there are two possible scenarios. In the ideal scenario (*left pointing arrow*),  $F(Y)$  models the data perfectly in the sense that the residual variation space, depicted by the plane  $R$ , contains no residual biological variation of interest. In this case, the surrogate variables, which are estimated from a SVD of  $R$ , and are indicated by blue arrows, don't need to align with the confounding factors (*green arrows*), as they are only required to span the same plane  $R$ . However, in the more realistic scenario, there could be imperfections in the model  $F(Y)$  (e.g., using a linear model when the relationship between  $X$  and  $Y$  is nonlinear), which in turn could lead to residual biological variation (*red arrow*) in the residual variation space  $R$ . In this case, we need to choose surrogate variables that align with the confounders and “avoid” the residual biological variation of interest, since otherwise using the whole space  $R$  in the subsequent adjustments will lead to loss of biological signal. Thus, in this scenario, we need to select an appropriate subspace of  $R$  and only use this subspace for the subsequent adjustments and supervised analysis. ISVA uses ICA instead of PCA/SVD in the decomposition of  $R$ , thus allowing to infer surrogate variables that better model the confounding sources of variation. Geometrically, this means that the independent surrogate variables align significantly better with the confounders and the residual biological variation, thus allowing an appropriate subspace of  $R$  to be selected. This subspace should not contain any residual biological variation and ICA is key to achieving this

### 17.4.2 Independent Surrogate Variable Analysis

Motivated by the discussion above, we seek a BSS method that can more accurately infer the sources of variation in the estimated residual matrix  $R$ . The generalization of SVA in which a BSS method is used to decompose  $R$  is called ISVA [45]. Although many BSS methods exist, in [45] we considered one of the simplest versions of ICA, the “fastICA” algorithm [15]. Thus, as with SVA, there are three parts to the ISVA algorithm: (i) detection of confounding/unmodeled factors (steps 1–4),

(ii) construction of surrogate variables (SVs) (steps 5–10), and (iii) final feature selection using the SVs as covariates.

In detail, the steps in ISVA are:

1. Construction of the residual variation matrix by removing the variation associated with the phenotype of interest:  $R_{ij} \equiv X_{ij} - \hat{f}_i(y_j)$ .
2. We estimate the intrinsic dimensionality,  $K$ , of the residual variation matrix using RMT [29]. This gives the number of components as input to the ICA algorithm.
3. Perform ICA on  $R$ :  $R = SA + \epsilon$ , with  $S$  a  $p \times K$  source matrix and  $A$  a  $K \times n$  mixing matrix. We point out that in this formulation of ICA, the statistical independence requirement is imposed on the columns of  $S$ . We denote the columns of  $S$  and rows of  $A$  by  $S_k$  and  $A_k$ , respectively.
4. We regress  $A_k$  to each  $X_i$  ( $i = 1, \dots, p$ ) and calculate  $P$ -values of association  $p_i$ .
5. From this  $P$ -value distribution, we estimate the FDR using the  $q$ -value method [38] and select the features with  $q < 0.05$ . If the number of selected features is less than 500, we select the top 500 features (based on  $P$ -values). Let  $r_k$  denote the number of selected features.
6. We construct the reduced  $r_k \times n$  data matrix  $X_r$  obtained by selecting the features in previous step.
7. Perform ICA on  $X_r$  using  $K$  independent components:  $X_r = S_r A_r + \epsilon_r$ . Find the column  $k^*$  of  $A_r$  that best correlates (absolute correlation) with  $A_k$ .
8. Set the SV  $v_k = (A_r)_{k^*}$ . The purpose of steps 4–8 is to regularize the estimates and thus avoid overfitting as explained in [20].
9. Repeat steps 4–8 for each significant independent component,  $A_k$ , obtained in step-3.
10. Perform SV subspace selection using a SV selection criterion. Let  $K^*$  denote the set of selected SVs.
11. Finally, we run the model

$$X_{ij} = f_i(y_j) + \sum_{k \in K^*} \lambda_{ki} v_{kj} + \epsilon'_{ij}. \quad (17.4)$$

and perform feature selection using a FDR ( $q$ -value) estimation procedure [38] and a nominal  $q$ -value threshold of say 0.05.

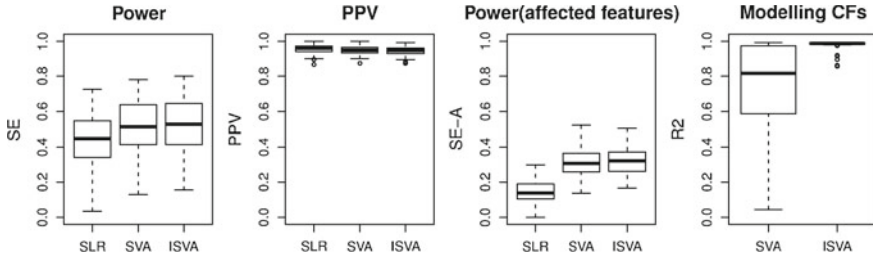
As formulated above, there are three differences between ISVA [45] and SVA [19]. First, ISVA uses RMT to estimate the dimensionality, in contrast to SVA which uses an explicit randomization procedure [20]. This difference is, however, not of major consequence [45]. Second, ISVA uses ICA in step-3 instead of SVD. Third, ISVA incorporates a SV subspace selection step (step-10) using a SV selection criterion that we shall discuss in detail in Sect. 17.7.4. This step is absolutely key to the improved inference that ISVA offers, and we point out here that the use of a BSS method in step-3 is also key to facilitating the choice of SV subspace in step-10. Finally, we remark that any BSS technique could be used to model the sources of variation in

$R$  (step-3), and thus the ISVA framework can be easily generalized to incorporate more sophisticated BSS algorithms.

## 17.5 Validation of SVA and ISVA on Simulated Data

Before exploring the SVA and ISVA algorithms in the context of real data, it is illuminating to first compare their performance on simulated data. The simulation model is exactly the one considered in [45], and for completeness we provide full details here again in the appendix. Briefly, we generated synthetic data matrices with 2,000 features and 50 samples and considered the case of two confounding factors (CFs) in addition to the primary POI. The primary phenotype is a binary variable  $y$  with 25 samples in one class ( $y = 0$ ) and the other half with  $y = 1$ . Similarly, each confounding factor is assumed to be a binary variable affecting one half of the samples (randomly selected). We further assume 10% of features (200 features) to be true positives (TPs) discriminating the two phenotypic classes. We model the confounding factors as follows: each confounding factor is assumed to affect 10% of features with a 25% overlap with the TPs (i.e., 50 of the 200 TPs are confounded by each factor). Without loss of generality, noise is modeled by a Gaussian of mean zero and unit variance  $N(0, 1)$ . We further assume that the POI is associated with an effect size  $e_y (= \Delta\mu/\sigma)$  of 1, i.e., the difference in the means between the phenotypes,  $\Delta\mu$ , equals the standard deviation,  $\sigma$ , within each group. Effect sizes of the two confounders are assumed to be equal to  $e_{CF}$  and we define the relative effect size as  $e_R \equiv e_{CF}/e_y = e_{CF}$ . We here consider the case  $e_R = 2$  corresponding to a situation where the confounding factors are associated with a larger variance than the POI. The simulation model is run a total of 100 times and for each run we record the following measures (using an estimated FDR threshold of 0.05): the sensitivity (SE), the positive predictive value (PPV), the sensitivity of TPs specifically affected by the confounding factors (SE-A), and the overall correlation ( $R^2$ -values) to the CFs. For the first three measures, we also compare SVA and ISVA to a simple linear regression method that does not do any adjustment for the confounding factors (LR). Results are shown in Fig. 17.4.

From this figure, we can make the following observations. First, the PPV is high for all methods, and is in line with the estimated FDR ( $=1-PPV$ ) of 0.05 used in performing feature selection. Second, we can see that the power of the study is reduced if no adjustment is made for the confounding factors. Indeed, we can see that, focusing on those true positive features which are corrupted by confounding variation, the sensitivity to retrieve these features is improved approximately twofold by using SVA or ISVA. Third, ISVA and SVA perform similarly on simulated data, despite the fact that ISVA reconstructs the confounding factors at substantially higher  $R^2$  values. Thus, the simulated data nicely illustrates the “perfect model” scenario depicted in the left side of Fig. 17.3. Since the data are simulated with the same model that is subsequently used to run the univariate regression, the residual variation matrix  $R$  contains no residual biological variation, hence it does not matter if the SVs align

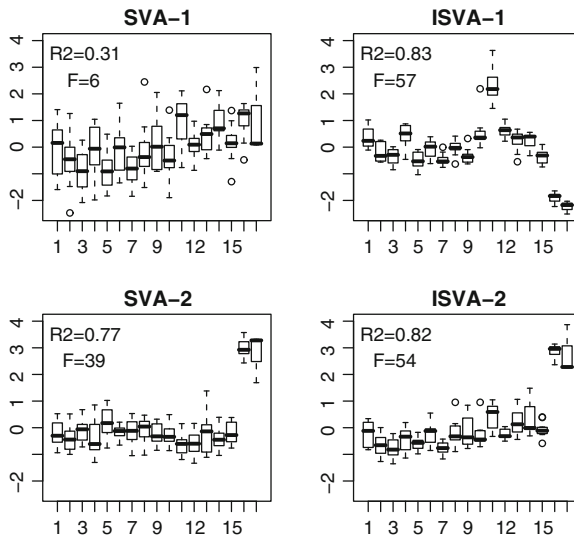


**Fig. 17.4** Feature selection performance metrics of different algorithms over 100 runs of the synthetic data ran with  $e_R = 2$ . The algorithms for feature selection are SVA, ISVA, and a simple linear regression without adjustment for confounders (SLR). For a given estimated FDR threshold of 0.05, we compare the sensitivity/power (SE), the positive predictive value (PPV), the sensitivity to detect true positives which are affected by confounders (SE-A), and the average  $R^2$ -value between confounders and the best correlated surrogate variable. See Appendix for further details of simulation model

with the confounders. The main requirement is for the SVs to span the space  $R$ , and hence similar results are obtained using the SVs from SVA or ISVA, since in both cases, the SVs span the same space.

## 17.6 Improved Modeling of Confounding Factors in Omic Data by BSS Methods

In the previous section, we have seen how ISVA models the confounding factors much better than SVA. The aim of this section is to demonstrate that ISVA also leads to improved modeling of the confounding sources of variation in real data. Later, in the subsequent section, we shall see how this translates into improved feature selection. Once again, we consider DNA methylation data and as confounding factor we consider the beadchip. Illumina Infinium beadchips can accommodate up to 12 samples per chip, hence there are enough samples for beadchip effects to be assessed. Importantly, it is always known which samples were profiled on which beadchip, hence this is an example of a known confounder and thus it can be used to objectively assess the quality of blind source separation. As a benchmark we consider SVA which uses SVD/PCA to decompose the residual variation matrix. As shown in Fig. 17.5, the surrogate variables inferred using ISVA model the beadchip effects substantially better than those inferred using SVA, as indicated by the significantly higher  $R^2$  values. For further examples, we refer the reader to [45].



**Fig. 17.5** Comparison of ISVA to SVA in identifying beadchip effects in the DNA methylation dataset from [3]. The weights (y-axis) of the two surrogate variables that most significantly associated with beadchip effects are plotted against beadchip number (x-axis), for SVA and ISVA separately. To compare the identifiability of beadchip effects, we provide the  $R^2$  and F-statistics of a linear ANOVA model with beadchip number as the independent variable

## 17.7 Improved Feature Selection Using ISVA

We have seen that ISVA can model confounding sources of variation substantially better than SVA. This in turn should lead to improved statistical inference, e.g., feature selection, at least in those scenarios where it is necessary to select a surrogate variable subspace, as explained in Sect. 17.3. To demonstrate this, we first provide a number of real data examples where SVA breaks down. Subsequently, we show how ISVA circumvents the problem, leading to substantially improved statistical inference.

### 17.7.1 SVA Breakdown in mRNA Expression Data

In order to demonstrate that SVA can break down, we consider a real dataset with a known biological signature: it is well known that many genes implicated in cell proliferation and the cell-cycle are differentially expressed between high and low grade cancers [26, 32, 36, 41]. The grade of a cancer refers to the level of differentiation of the cancer cells, with high-grade cancers exhibiting a less differentiated state, whilst low-grade cancers are more differentiated in the sense that they are



more similar to normal (healthy) tissue, which is a highly differentiated state compared to the undifferentiated stem cells that they are derived from. Thus, high-grade cancers are generally more aggressive and correspondingly are also characterized by a higher expression of cell proliferation and cell-cycle genes. This cell proliferation gene expression signature is a universal signature, able to distinguish high grade from low-grade cancers, irrespective of tissue type [26, 32, 36, 41]. Thus, given a gene expression dataset of high and low grade cancers, selecting features (genes) that best discriminate low and high grade cancers should lead to significant enrichment of genes implicated in the cell-cycle and cell proliferation. The enrichment of a top ranked list of discriminatory genes for any gene ontology can be assessed using a Fisher's exact test, as done for instance in [43], a procedure known generally as Gene Set Enrichment Analysis (GSEA) [39]. If a feature selection method were to not yield significant enrichment for cell-cycle or cell proliferation genes, one would conclude that the feature selection procedure has failed to retrieve the known biological signature. Thus, in what follows we consider "grade" as the POI and we aim to show that SVA breaks down, not being able to retrieve the cell proliferation/cell-cycle enrichment due to the presence of confounding factors.

Specifically, we consider the case of breast cancer. There are two main subtypes of breast cancer: estrogen receptor positive (ER+) and estrogen receptor negative (ER-) breast cancer [48]. This stratification of breast cancers reflects the levels of expression of the estrogen receptor gene, *ESR1*, with ER- breast cancers showing absent expression of *ESR1*. Thus, in ER+ breast cancer, *ESR1* expression and activity is high, which results in the overexpression of genes within the *ESR1* signaling pathway. We note that these *ESR1* signaling genes are different from the cell-cycle/cell-proliferation ones. Now, it is well known that most ER- breast cancers are of high grade, whilst ER+ breast cancers can be either high or low grade [41]. Thus, if the aim is to identify genes whose expression correlates with grade, ER-status may be seen as a biological confounder, since the distribution of ER+ and ER- tumors will differ between low and high-grade cancers. Furthermore, it is also well known that low and high grade ER+ breast cancers do not differ in terms of the level of *ESR1* expression and ER-signaling [26, 36, 41]. Hence, this means that in the task of identifying genes that are associated with grade, any gene set enrichment must be specific to cell-cycle and should not include terms involved in ER-signaling. In other words, if feature selection for grade associated genes also leads to enrichment of ER-signaling genes, then this indicates confounding by ER-status. Although here the confounder is biological, this does not matter for the sake of comparing algorithms, and indeed the biological framework considered here provides a nice testing ground for the SVA and ISVA algorithms.

As expression data, we consider the data from four independent breast cancer studies [5, 26, 35, 36], as used in [45]. In these datasets, besides ER-status, we also consider tumor size as a potential biological confounder. We note that in these datasets potential technical confounders such as batch effects are unknown. The *P*-values of the GSEA of the top ranked grade-associated genes against cell-cycle and ER-signaling terms are given in Table 17.1 for genes selected using SVA and a

**Table 17.1** Grade associated expression differences: in each mRNA expression dataset and for each method (LR+CFs, SVA, ISVA) we give the number of confounding factors (CFs) or SVs used as covariates in the regression analysis, the number of genes differentially expressed with histological grade (nDEGs) at a false discovery rate threshold of 0.05 ( $FDR < 0.05$ ), and the  $P$ -value of enrichment (Hypergeometric/Fisher test) of cell-cycle and estrogen upregulated gene (ESR1-UP) categories among these differentially expressed genes

	LR + CF	SVA	ISVA
<i>Dataset(Sotiriou)</i>			
nCF/SV	2	4	4
nDEGs	491	0	607
$P$ -value(Cell-cycle)	<b>6e-18</b>	1	<b>5e-16</b>
$P$ -value(ESR1-UP)	0.03	1	0.14
<i>Dataset(Loi)</i>			
nCF/SV	2	19	5
nDEGs	829	0	146
$P$ -value(Cell-cycle)	<b>5e-37</b>	1	<b>7e-24</b>
$P$ -value(ESR1-UP)	0.90	1	0.61
<i>Dataset(Schmidt)</i>			
nCF/SV	2	27	15
nDEGs	2364	0	451
$P$ -value(Cell-cycle)	<b>3e-25</b>	1	<b>5e-19</b>
$P$ -value(ESR1-UP)	<b>7e-4</b>	1	0.14
<i>Dataset(Blenkiron)</i>			
nCF/SV	2	20	8
nDEGs	1292	1	829
$P$ -value(Cell-cycle)	<b>2e-25</b>	1	<b>7e-27</b>
$P$ -value(ESR1-UP)	<b>7e-4</b>	1	0.31

Confounding factors here are ER status and tumor size. In bold face we indicate those  $P$ -values that are significant after adjustment for multiple-testing

feature selection method that uses ER–status and tumor size as explicit covariates in the linear regression model (LR + CF).

Based on this table, we can make two important observations. First, in three datasets, SVA predicts no differentially expressed genes between low and high grade breast cancer, a result which is in complete disagreement with extensive biological knowledge [26, 32, 41]. As a result of this, none of the biological terms cell-cycle or ER–signaling are enriched. Second, performing feature selection using a multivariate linear regression model with ER–status and size as explicit covariates (LR + CF) leads to many differentially expressed genes (DEGs) in every dataset. Correspondingly, we observe strong enrichment of the cell-cycle term among these genes, consistent with biological knowledge. However, we also observe that ER–signaling is significantly enriched in 2 out of 4 studies, hence the enrichment for cell-cycle genes is nonspecific. This means that explicit adjustment for the confounders has not fully eliminated the effect of one confounder (ER–status) and hence we can conclude that the list of DEGs contains many false positives associated with ER–signaling. This contamination of ER–signaling genes is likely to be due to the

**Table 17.2** Age-associated CpGs: in each dataset and for each method (LR + CFs, SVA, ISVA) we give the number of CFs or SVs used as covariates in the regression analysis, the number of CpGs differentially methylated with age (nDMCs) (FDR < 0.05 for Datasets T1D and UKOPS1, FDR < 0.3 for Datasets UKOPS2 and WBBC), the number of these that are hypermethylated with age and that map to polycomb group targets (nPCGTs), and the *P*-value of PCGT enrichment among age-hypermethylated CpGs (Hypergeometric test)

Dataset(T1D)	LR + CF	SVA	ISVA
nCF/SV	4	4	6
nDMCs	440	688	902
nPCGTs	96	110	148
<i>P</i> -value	<b>4e-32</b>	<b>7e-26</b>	<b>2e-34</b>
<i>Dataset(UKOPS1)</i>			
nCF/SV	3	18	6
nDMCs	267	4	232
nPCGTs	75	1	59
<i>P</i> -value	<b>4e-24</b>	0.27	<b>2e-19</b>
<i>Dataset(UKOPS2)</i>			
nCF/SV	3	21	8
nDMCs	20	201	225
nPCGTs	4	15	29
<i>P</i> -value	0.001	0.01	<b>3e-7</b>
<i>Dataset(WBBC)</i>			
nCF/SV	3	15	6
nDMCs	564	185	469
nPCGTs	84	19	64
<i>P</i> -value	<b>7e-22</b>	0.01	<b>3e-11</b>

The CFs in each dataset are described in Appendix. In bold-face we indicate those *P*-values that are significant after adjustment for multiple-testing

fact that the immunohistochemically determined ER—status of the samples is only approximate, i.e., the confounder is subject to error. Thus, neither method, SVA or LR + CF, succeeds in yielding specific enrichment of cell-cycle genes among the genes associated with grade.

### 17.7.2 SVA Breakdown in DNA Methylation Data

As a second example, we consider DNA methylation data. A large number of studies have now unequivocally demonstrated that promoter DNA methylation of a specific class of genes, known generally as PolyComb Group Targets (PCGTs), increases with the age of the tissue (see e.g., [27, 30, 44]). Hence, feature selection for CpGs in gene promoters undergoing age-associated increases in DNA methylation should be enriched of PCGTs. Table 17.2 shows the results of applying SVA and a linear regression method that uses confounders as explicit covariates (LR + CF).

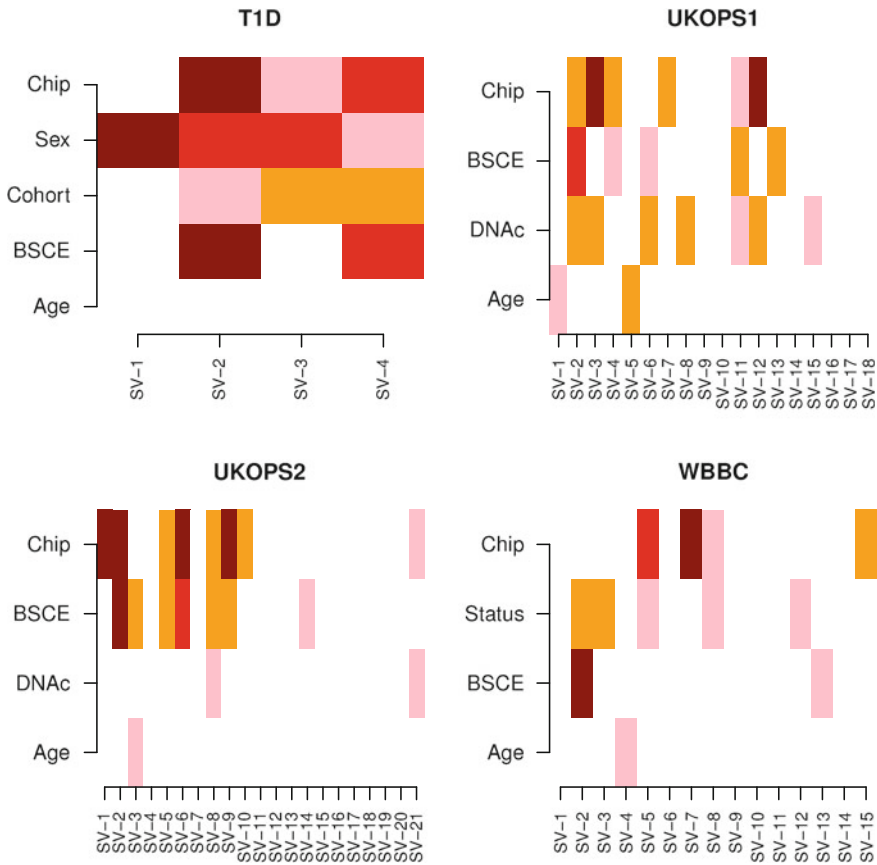
We can see that in only one of the four datasets (T1D set), does SVA convincingly retrieve the age-PCGT DNA methylation signature. In the other three datasets, the  $P$ -value of enrichment is either not significant or would fail significance after correction for multiple testing. In contrast, linear regression with explicit adjustment for confounders (see Appendix for the nature of the explicit confounders) convincingly captures the biological signature in 3 out of 4 datasets.

### ***17.7.3 Residual Biological Variation***

The results presented above clearly demonstrate a pitfall of the SVA algorithm: it can fail to retrieve a well-known and extensively validated association between a molecular signature and a phenotype of interest. The most plausible explanation for why this happens is that residual biological variation is being interpreted as confounding variation leading to a “dampening” of the biological signal (see Fig. 17.3). To show that this is indeed what is happening we can study the correlations between the surrogate variables and the biological as well as confounding factors. The statistical significance of these correlations is best shown as a heatmap. This is shown for the four DNA methylation datasets considered in Table 17.2 in Fig. 17.6. From this figure and Table 17.2 we can see that in all three datasets where SVA fails to clearly capture the age-PCGT DNA methylation signature, that in all three of them there is residual variation correlating with age. Conversely, in the one dataset where there is no residual variation correlating with age (i.e., T1D set), SVA retrieves the biological signature. Thus, this example clearly illustrates that the scenario of residual biological variation arising due to imperfections in the modeling, as depicted in Fig. 17.3, is indeed fairly common.

### ***17.7.4 The Need for Surrogate Variable Subspace Selection***

The above two examples in gene expression and DNA methylation data demonstrate the need to perform adjustment on a surrogate variable subspace, since otherwise one risks “peeling” away biological variation of interest. In the case where there is no residual biological variation it should be clear that it does not matter what basis (i.e., surrogate variables) we use to span the surrogate variable subspace. In other words, it should not matter whether we use SVs constructed from principal components (SVA) or from the independent components (ISVA). However, in the scenario where biological variation of interest is present in the residual variation matrix  $R$ , we need to select surrogate variables that “align” with the true confounders and which avoid as much as possible the directions defined by the residual biological variation. This then requires a BSS method to better deconvolute the effects of the confounders and this residual biological variability. However, application of a BSS method to  $R$  only yields a decomposition of  $R$  into a number of independent “sources” and does not,



**Fig. 17.6** Heatmap of  $P$ -values of association between the surrogate variables (SVs) inferred using SVA and the confounders and phenotype of interest (age).  $P$ -values were estimated from linear ANOVA in the case of categorical confounders (e.g Chip, Sex, Cohort) and from linear regressions in the case of continuous variables (age, BSC efficiency-BSCE and DNA concentration-DNAc). Color codes:  $P < 1e - 10$  (darkred),  $P < 1e - 5$  (red),  $P < 0.001$  (orange),  $P < 0.05$  (pink),  $P > 0.05$  (white)

on its own, provide a prescription for subspace selection. Hence, how do we select this subspace?

The previous example discussed in Table 17.2 and Fig. 17.6 provides a possible prescription for how to perform the subspace selection, namely, only those SVs should be included that do not correlate significantly with the phenotype of interest. But what if SVs correlate significantly with both the POI and a confounder? In this scenario, it is unclear whether to include these SVs in the final feature selection procedure (i.e., step-11). The surrogate variable selection step therefore remains an outstanding problem.

**Table 17.3** Surrogate Variable Selection: there are four possible case scenarios to consider depending on the  $R_{vf}^2$  values between surrogate variable  $v$  and factor  $f$ , as shown

Scenarios	POI( $f = b$ )	CF( $f = t$ )	ISVA
Case-1	$P_{vb} < 0.001$	$P_{vt} > 0.001$	Exclude
Case-2	$P_{vb} > 0.001$	$P_{vt} < 0.001$	Include
Case-3	$P_{vb} < 0.001$	$P_{vt} < 0.001$	Include if $R_{vb}^2 < R_{vt}^2$
Case-4	$P_{vb} > 0.001$	$P_{vt} > 0.001$	Normally include

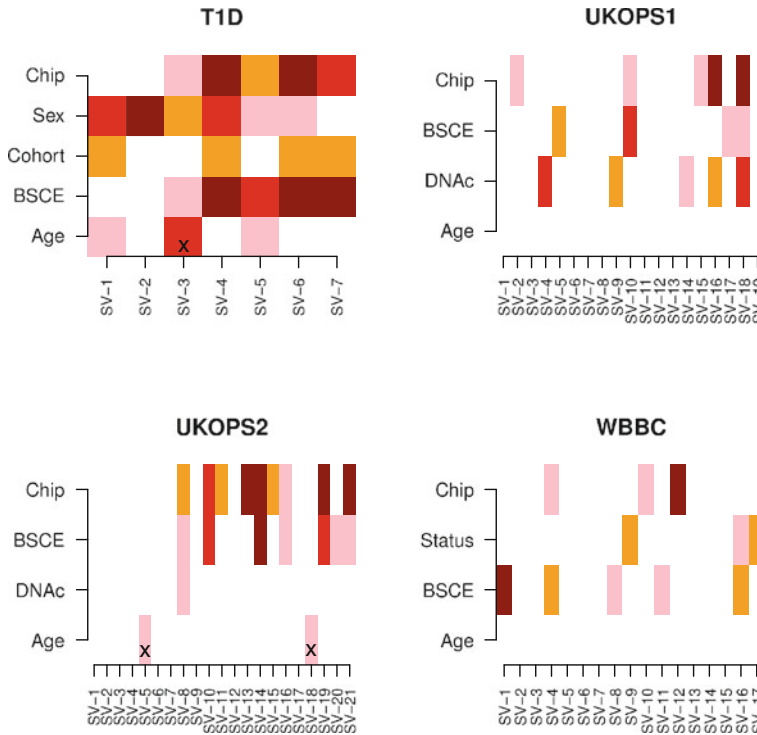
POI phenotype of interest ( $f = b$ ), CF technical confounder ( $f = t$ ).  $P_{vf}$  denotes the  $P$ -value of the association between SV  $v$  and factor  $f$ . Final column indicates whether the SV  $v$  should be included in the final adjustment step of ISVA or not. A conservative Bonferroni threshold of 0.001 is used to call statistical significance since the number of SVs is typically on the order of  $\sim 10$

Here we propose a simple heuristic to the subspace selection problem, which we can only justify a posteriori, by showing that it leads to successful retrieval of the known biological signatures. For each of the SVs and for each factor (biological or technical) we first compute a model fit  $R^2$  value, using an appropriate linear or nonlinear model framework. Let  $R_{vf}^2$  denote the  $R^2$  value between surrogate variable  $v$  and factor  $f$ . Further, let  $b$  denote the POI factor, and  $t$  denote a generic technical factor. Then, there are four possible cases to consider, as indicated in Table 17.3. In case-1, the surrogate variable correlates significantly only with the POI, and hence it ought to be excluded as remarked earlier. Conversely, if the surrogate variable correlates significantly with a technical factor but not with the POI, then the corresponding SV should be included. In the third case, where the SV correlates significantly with both the POI and a technical CF, we use the model selection criterion

$$R_{vb}^2 < R_{vt}^2 \quad (17.5)$$

to include only those where the correlation with the technical factor is stronger. The rationale for this criterion is that if the variation described by  $v$  correlates more strongly with the POI, then it is more likely that this variation is genuinely associated with the POI, and hence this component should be excluded. The final case corresponds to a scenario where the SV does not correlate with any known factor, in which case it is also unclear whether to include the SV or not. In principle, one must allow for the possibility of complete unknown (i.e., hidden) factors, in which case the SV should be included. On the other hand, exclusion could be argued on grounds of small variability and inaccuracies in dimensionality estimation.

Before demonstrating that the simple procedure presented in Table 17.3 works, we need to discuss further what may seem as a serious drawback with the above heuristic, as it requires some knowledge of the technical confounding factors. Given that BSS methods are ideally suited to the scenario where sources of variation are unknown, does this then pose an intrinsic limitation to the ISVA method? The answer is no. To understand this, we first note that BSS methods are useful also in circumstances where confounders are only known with error, since in such cases it would be better to model the effects of the confounders from the data itself. In this case, the simple



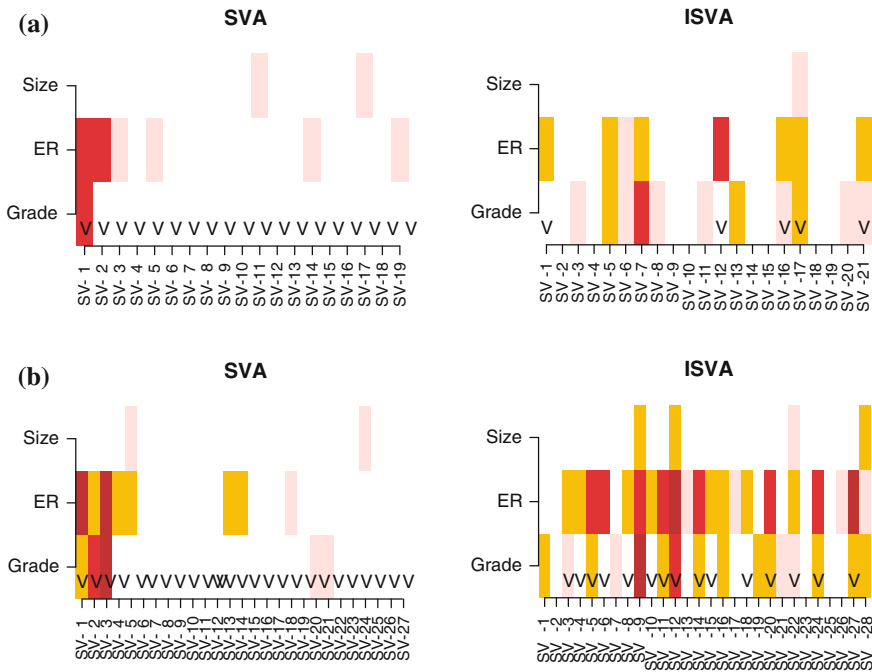
**Fig. 17.7** Heatmap of  $P$ -values of association between the surrogate variables (SVs) inferred using ISVA and the confounders and phenotype of interest (age).  $P$ -values were estimated from linear ANOVA in the case of categorical confounders (e.g., Chip, Sex, Cohort) and from linear regressions in the case of continuous variables (age, BSC efficiency-BSCE and DNA concentration-DNAc). Color codes:  $P < 1e - 10$  (darkred),  $P < 1e - 5$  (red),  $P < 0.001$  (orange),  $P < 0.05$  (pink),  $P > 0.05$  (white)

SV subspace selection step described above can be applied. Second, the scenario where confounders are known, or only known subject to error, constitutes the most common scenario. Last but not least, SVs not correlating with any factor (case-4) may still be included in the adjustment, as the main requirement is to avoid including SVs that correlate strongly with the POI.

### 17.7.5 The ISVA Solution

Let us now see how ISVA resolves the problematic issues that we encountered earlier with SVA. We first consider the four DNA methylation datasets considered in Table 17.2 and Fig. 17.6. In Fig. 17.7 we show the heatmap of associations between SVs constructed from ISVA with the same confounders.





**Fig. 17.8** Heatmap of  $P$ -values of association between the surrogate variables (SVs) inferred using SVA and ISVA and the confounders (ER—status and tumor size) and the phenotype of interest (Grade). **a** Dataset Loi, **b** Dataset Schmidt.  $P$ -values were estimated from linear regressions. Color codes:  $P < 1e - 10$  (darkred),  $P < 1e - 5$  (red),  $P < 0.001$  (orange),  $P < 0.05$  (pink),  $P > 0.05$  (white). “V” indicates SVs selected for adjustment in SVA or ISVA

Note how in two datasets (UKOPS1 and WBBC) there is no residual biological variability associated with age (the POI). In the UKOPS2 set, there are two SVs that correlate marginally with age, and importantly they do not correlate with any other factor, hence these are not included in step-11 of ISVA. In the T1D set, there are three SVs that correlate with age, but only one of these (SV-3) is excluded, because the other two (SV-1 and SV-5) correlate more strongly with potential confounders such as Sex, Cohort, BSCE, and Chip. As seen in Table 17.2, ISVA with the above prescription for SV subspace selection, leads to significant enrichment of PCGTs in all four DNA methylation datasets. Thus, using ISVA the known biological signature is successfully retrieved in all sets.

It could be argued that the key step is the SV subspace selection, and not the BSS algorithm *per se*. To show how the use of ICA facilitates the SV subspace selection, we return to the example of mRNA expression data with grade as the POI and ER—status playing the role of confounder. Table 17.1 shows the results obtained by ISVA. In comparison to SVA, we can see that ISVA leads to specific enrichment of cell-cycle genes (i.e., ER—signaling genes are not enriched), clearly indicating that confounding by ER—status has been successfully removed. As we can see from Fig. 17.8, this improved feature selection can be attributed to a more accurate

deconvolution of residual variation associated with grade from that associated with ER–status. As illustrated in Fig. 17.8a, SV-1 in SVA is equally strongly correlated with grade and ER–status, indicating inaccurate deconvolution. In contrast, with ISVA, the SVs correlating most strongly with ER (SV-12) and grade (SV-7) are distinct, thus facilitating SV subspace selection and subsequently allowing improved feature selection. Similarly, in Fig. 17.8b, SV-3 in SVA is selected for adjustment yet it correlates very strongly with grade. In contrast, in ISVA the SV correlating most strongly with grade (SV-9) does so much more strongly than with ER–status, and hence this SV is not included in the subsequent adjustment. The effect of ER in the residual variation space is captured by other SVs (SV-12, 20, 24, 27) which do not correlate as strongly with grade, and these are therefore included in the adjustment. Thus, in these two examples, the BSS method is key since it allows more accurate deconvolution of the different sources of variation in the residual variation space. Even if a SV subspace selection step is incorporated into SVA (using the same heuristic criterion as for ISVA), we would still select problematic SVs since PCA does not allow accurate deconvolution of the different sources of variation (see [45] for results of this modified SVA).

## 17.8 Modeling of Confounding Factors with Generalized BSS Algorithms

In the previous sections, we have seen how a simple BSS method (fastICA) can lead to substantial improvements in modeling confounding factors as well as to an improved deconvolution of the biological and confounding factors, both of which are important, and which subsequently lead to improved feature selection in supervised analysis problems. We have also provided an objective evaluation framework in which to assess and compare the different algorithms.

It is therefore of interest to consider more sophisticated BSS methods, since these might offer further improvements in statistical inference. In doing so, the first question to address is whether modeling of confounders is improved using these more advanced BSS methods. One particular generalization of ICA which is of interest to study concerns the statistical independence assumption, which so far has been applied to the columns of the source matrix  $S$ . In other words, given the residual matrix  $R$  of dimension  $p \times n$ , we applied ICA in the context

$$R = S_1 A + \epsilon \tag{17.6}$$

with the inference required to minimize a residual sum of squares subject to the constraint that the  $K$   $p$ –dimensional column vectors of  $S_1$  be as statistically independent as possible. However, as shown in previous studies [37, 47], a dual interpretation/implementation is possible, whereby statistical independence is imposed on the rows of the mixing matrix  $A$ . This dual problem can be expressed as:

$$\begin{aligned} R^T &= A^T S_1^T + \epsilon \\ &= \tilde{S}_2 \tilde{A} + \epsilon \end{aligned} \quad (17.7)$$

where statistical independence is now imposed on the columns of  $\tilde{S}_2$  which is a matrix of dimensionality  $n \times K$ . As shown in [2, 33, 37, 47], it is possible to formulate a “spatio-temporal” or bi-dimensional ICA,

$$R = S_1 S_2^T + \epsilon \quad (17.8)$$

in which statistical independence is favored across both features (“time”) and samples (“space”), by means of an overall cost function,  $C_f$ , defined as a weighted linear combination of the cost functions used to solve Eqs. 17.6 and 17.7, i.e.,

$$C_f = (1 - a)C_{f_1} + aC_{f_2} \quad (17.9)$$

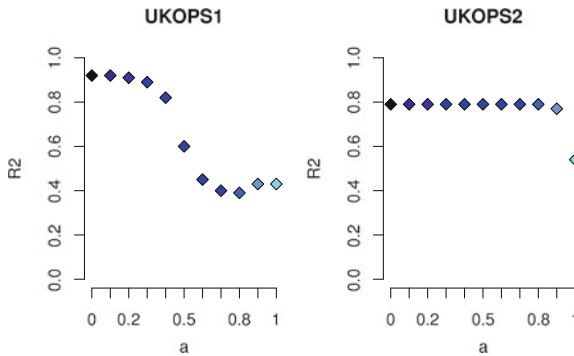
More formally, the specific bi-dimensional ICA algorithm we consider here [2, 33, 47] starts with a SVD of the row and column centered (residual) data matrix  $R$ , so  $R = U D V^T$ , with corresponding estimation of the dimensionality  $K$  (using as before RMT). One then constructs the reduced matrix  $R_K = U_K D_K V_K^T$  where the first  $K$  columns of  $U$  and  $V$  have been selected corresponding to the top  $K$  singular values of  $D$ . This reduced matrix can then be rewritten as

$$R_K = \underbrace{U_K D_K W^{-1}}_{S_1} \underbrace{W V_K^T}_{S_2^T} \quad (17.10)$$

with  $W$  an invertible matrix of size  $K \times K$ . Finally, we seek to optimize the matrix  $W$  such that the fourth-order cumulants of  $S_1$  and  $S_2$  are as diagonal as possible, i.e., minimizing

$$C_f(W) = \left( a \sum_i \text{Off} \left( C_i(S_2^T) \right) + (1 - a) \sum_i \text{Off} \left( C_i(S_1^T) \right) \right) \quad (17.11)$$

where  $\text{Off}(Y)$  returns the sum of squares of the off-diagonal elements of  $Y$ , and the  $C_i$  are fourth-order cumulants. Imposing that  $W$  is orthogonal leads to a formulation which can be solved by means of the JADE algorithm [6]. We note however that this formulation of bi-dimensional ICA differs slightly from that of [33, 47], as the second term in the contrast function involves  $(C_i(S_1^T))$  instead of  $(C_i(S_1^T))^{-1}$ . Minimizing one or the other pursues the same goal, namely statistical independence for columns of  $S_1$ . This novel formulation however allows us to treat both extreme cases on an equal footing:  $a = 1$  corresponds to JADE applied on  $R_K^T = S_2 S_1^T$  whereas  $a = 0$  corresponds to JADE applied on  $R_K = S_1 S_2^T$ . Thus, the cost function can be interpreted as a weighted linear combination of two ‘jade-like’ cost functions.



**Fig. 17.9** Modeling of beadchip effects by bi-dimensional ICA in two DNA methylation datasets. y-axis labels the  $R^2$  value of the component correlating best with the beadchip as assessed using a linear ANOVA model. x-axis labels the parameter  $a$  in Eq. 17.11

Given the above formulation of bi-dimensional ICA, it is of interest to study the effect of the parameter  $a$  on the quality of BSS. Since beadchip effects provide an objective framework in which to assess the quality of the BSS, we focus on how well these effects are modeled by the family of bi-dimensional ICA algorithms above. For simplicity, we consider the unsupervised problem in which the ICA decomposition is done on the data matrix  $X$  itself.<sup>1</sup> Figure 17.9 shows the results, indicating that in terms of modeling beadchip effects, ICA is best run with values of  $a$  close to zero. This corresponds to imposing statistical independence of the sources across features, as implemented in the fastICA version of the ISVA algorithm.

## 17.9 Conclusions

In this chapter, we have presented and discussed the problem that confounding factors pose in large omic datasets. Since feature selection is a common task in the analysis of such large datasets, it is paramount to have statistical methods in place that can perform supervised analysis and feature selection in the background of such confounding factors, specially when these are uncertain or unknown. We have seen how BSS methods are necessary in this context, since there is a requirement to accurately model confounding factors and to deconvolve these from variation associated with the phenotype of interest. We have presented an algorithm, ISVA, which uses a BSS technique (ICA) to perform a supervised normalization of the data and have shown that it offers a more sound statistical framework in which to perform feature selection than a competing non-BSS tool based on PCA.

As mentioned earlier, it is possible to consider any BSS algorithm within the ISVA framework. One of the most straightforward generalizations of the fastICA algorithm used in our ISVA implementation is to relax the statistical independence

<sup>1</sup> Instead of the residual variation matrix  $R$  which requires specification of the POI and is thus supervised.

assumption, but to simultaneously impose partial statistical independence along the dual “sample”-space, resulting in a bi-dimensional ICA. However, we have seen that, at least in terms of modeling beadchip effects, that the original implementation (i.e., imposing statistical independence across features) is optimal. This could be due to the sources across features being well described by sparse distributions or by the fact that statistical independence is best assessed using the larger feature space.

Although the bi-dimensional ICA did not lead to improved modeling of beadchip effects, it is nevertheless of interest to investigate this and other BSS algorithms in the ISVA context. For instance, it could well be that other types of confounding factors are best modeled using bi-dimensional ICA or ICA algorithms that also allow for skewed sources of variation [37, 47]. Exact known confounders (like beadchip effects) allow for objective assessment of BSS in real data, yet unfortunately, not many such factors exist. On the other hand, the number of beadchips in studies can vary substantially, thus allowing assessment of the BSS methods at least in relation to statistical properties such as kurtosis, which would vary for beadchip effects depending on the overall sample size of the study. Thus, a beadchip effect affecting 12 samples out of 120 samples (10 beadchips) will exhibit different statistical properties to one in a study of only 36 samples.

Besides the detailed modeling of the sources, another key challenge faced in ISVA is the SV subspace selection step. Although we have presented a simple heuristic selection criterion, which, as we have seen, successfully retrieves the known biological signatures in diverse real datasets, the criterion itself is not applicable to the case where confounders are complete unknowns (i.e., hidden). In fact, this remains an outstanding statistical challenge since (1) the presence of biological variation of interest in the matrix of residuals is almost always inevitable and (2) it is entirely plausible that some of this variation is driven by hidden confounding factors and hence that the associated SVs should be included in the final regression model.

The results on eight real datasets presented here however, conclusively demonstrate that a SV selection step is absolutely necessary to arrive at the correct biological conclusion, yet in other datasets where the biological truth is unknown, the SV selection criterion used here could falter due to hidden confounding factors. In other words, in the eight real datasets considered here we can be fairly certain that the data is not subject to substantial hidden (i.e., completely unknown) confounding variation, since otherwise our SV selection criterion would not have led to the retrieval of the known biological signatures.

With this chapter we hope to engage biologists, bioinformaticians, and signal processing experts alike. The problem that confounding factors pose in the statistical analysis of omic data is both challenging and critical to the ultimate success of large-scale genomic and epigenomic studies aiming to identify the much needed disease biomarkers. Further research in this area is therefore urgently needed.

**Acknowledgments** AET was supported by a Heller Research Fellowship. This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Program initiated by the Belgian Science Policy Office.

## Appendix

### *Simulated Data*

We simulated data matrices with 2,000 features and 50 samples and considered the case of two confounding factors (CFs) in addition to the primary phenotype of interest. The primary phenotype is a binary variable  $I_1$  with 25 samples in one class ( $I_1 = 0$ ) and the other half with  $I_1 = 1$ . Similarly, each confounding factor is assumed to be a binary variable affecting one half of the samples (randomly selected). For a given sample  $s$  we thus have a 3-tuple of indicator variables  $I_s = (I_{1s}, I_{2s}, I_{3s})$  where  $I_2$  and  $I_3$  are the indicators for the two confounding factors. Thus, samples fall into 8 classes. For instance, if  $I_s = (0, 0, 0)$  then this sample belongs to phenotype class 1 and is not affected by the two confounding factors. Similarly,  $I_s = (0, 1, 0)$  means that the sample belongs to class 1 and is affected by the first confounding factor but not the second.

We assume 10% of features (200 features) to be TPs discriminating between the two phenotypic classes. We model the confounding factors as follows: each confounding factor is assumed to affect 10% of features with a 25% overlap with the TPs (i.e. 50 of the 200 TPs are confounded by each factor). Let  $J_g$  denote the indicator variable of feature  $g$ , so  $J_g$  is a 3-tuple  $(J_{1g}, J_{2g}, J_{3g})$  with  $J_{1g}$  an indicator for the feature to be a true positive, and  $J_{2g}$  ( $J_{3g}$ ) an indicator for the feature to be affected by the first (second) confounding factor. Thus, the space of features is also divided into eight groups. Furthermore, let  $(e_1, e_2, e_3)$  denote the effect sizes of the primary variable and the two confounding factors respectively, where we assume for simplicity that  $e_2 = e_3$ . Without loss of generality, we further assume that noise is modeled by a Gaussian of mean zero and unit variance  $N(0, 1)$ . Thus, for a given sample  $s$  we draw data values for the various feature groups as follows:

1.  $J_g = (0, 0, 0)$ : null unaffected features

$$p(x|I_s) \sim \delta_{J_g,000}N(0, 1)$$

2.  $J_g = (0, 1, 0)$  or  $(0, 0, 1)$ : null features affected by only one CF

$$\begin{aligned} p(x|I_s) \sim & \delta_{J_g,010} \{ \delta_{I_s,x1z} N(e_2, 1) \\ & + \delta_{I_s,x0z} N(0, 1) \} \\ & + \delta_{J_g,001} \{ \delta_{I_s,xy1} N(e_3, 1) \\ & + \delta_{I_s,xy0} N(0, 1) \} \end{aligned}$$

3.  $J_g = (0, 1, 1)$ : null features affected by the two CFs

$$\begin{aligned}
 p(x|I_s) &\sim \delta_{J_g,011} \{ \delta_{I_s,x11} N(e_2 + e_3, 1) \\
 &\quad + \delta_{I_s,x01} N(e_3, 1) \\
 &\quad + \delta_{I_s,x10} N(e_2, 1) \\
 &\quad + \delta_{I_s,x00} N(0, 1) \}
 \end{aligned}$$

4.  $J_g = (1, 0, 0)$ : true positives not affected by CFs

$$\begin{aligned}
 p(x|I_s) &\sim \delta_{J_g,100} \{ \delta_{I_s,0yz} N(0, 1) \\
 &\quad + \delta_{I_s,1yz} (\pi_{-1} N(-e_1, 1) + \pi_1 N(e_1, 1)) \}
 \end{aligned}$$

5.  $J_g = (1, 0, 1)$  or  $(1, 1, 0)$ : true positives affected by one CF

$$\begin{aligned}
 p(x|I_s) &\sim \delta_{J_g,101} \{ \delta_{I_s,0y0} N(0, 1) + \delta_{I_s,0y1} N(e_3, 1) \\
 &\quad + \delta_{I_s,1y0} (\pi_{-1} N(-e_1, 1) + \pi_1 N(e_1, 1)) \\
 &\quad + \delta_{I_s,1y1} (\pi_{-1} N(-e_1 + e_3, 1) \\
 &\quad + \pi_1 N(e_1 + e_3, 1)) \} \\
 &\sim \delta_{J_g,110} \{ \delta_{I_s,00z} N(0, 1) + \delta_{I_s,01z} N(e_2, 1) \\
 &\quad + \delta_{I_s,10z} (\pi_{-1} N(-e_1, 1) + \pi_1 N(e_1, 1)) \\
 &\quad + \delta_{I_s,11z} (\pi_{-1} N(-e_1 + e_2, 1) \\
 &\quad + \pi_1 N(e_1 + e_2, 1)) \}
 \end{aligned}$$

6.  $J_g = (1, 1, 1)$ : true positives affected by all CFs

$$\begin{aligned}
 p(x|I_s) &\sim \delta_{J_g,111} \{ \delta_{I_s,000} N(0, 1) \\
 &\quad + \delta_{I_s,010} N(e_2, 1) + \delta_{I_s,001} N(e_3, 1) \\
 &\quad + \delta_{I_s,011} N(e_2 + e_3, 1) \\
 &\quad + \delta_{I_s,101} (\pi_{-1} N(-e_1 + e_3, 1) \\
 &\quad + \pi_1 N(e_1 + e_3, 1)) \\
 &\quad + \delta_{I_s,110} (\pi_{-1} N(-e_1 + e_2, 1) \\
 &\quad + \pi_1 N(e_1 + e_2, 1)) \\
 &\quad + \delta_{I_s,111} (\pi_{-1} N(-e_1 + e_2 + e_3, 1) \\
 &\quad + \pi_1 N(e_1 + e_2 + e_3, 1)) \}
 \end{aligned}$$



where in the above  $\delta_{x'y'z',xyz}$  denotes the triple Kronecker delta:  $\delta_{x'y'z',xyz} = 1$  if and only if  $x' = x$ ,  $y' = y$  and  $z' = z$ , otherwise  $\delta_{x'y'z',xyz} = 0$ , and  $(\pi_{-1}, \pi_1)$  are weights satisfying  $\pi_{-1} + \pi_1 = 1$ . In our case, we used  $\pi_1 = \pi_{-1} = 0.5$ .

### ***DNA Methylation Data (Whole Blood Tissue)***

In all datasets, age is the phenotype of interest. (i) T1D: this DNAm dataset consists of 187 blood samples from patients (94 women and 93 men) with type-1 diabetes. This set served as validation for a DNAm signature for aging [44]. We take BSCE, beadchip, cohort, and sex as potential confounding factors. Samples were distributed over 17 beadchips; (ii) UKOPS1: this DNAm set consists of 108 blood samples from healthy postmenopausal women which served as controls for the UKOPS study [43]. Confounding factors in this study include BSCE, beadchip and DNA concentration (DNAc). Samples were distributed over 10 beadchips; (iii) UKOPS2: This is similar to Dataset2 but consists of 145 blood samples from healthy postmenopausal women distributed over 36 beadchips (i.e., approximately four healthy samples per chip, the other eight blood samples per chip were from cancer cases) [43]; (iv) WBBC: This dataset consists of whole blood samples from a total of 84 women (49 healthy and 35 women with breast cancer). Samples were distributed over seven beadchips, and confounders are BSCE, status (cancer/healthy), and beadchip.

### ***Breast Cancer mRNA Expression Data***

The mRNA expression profiles are all from primary breast cancers and three of the datasets were profiled on Affymetrix platforms, while another was profiled on an Illumina Beadchip. Normalized data were downloaded from GEO (<http://ncbi.nlm.nih.gov/>), and probes mapping to the same Entrez ID identifier were averaged. Sotiriou: 14,223 genes and 101 samples [36]; Loi: 15,736 genes and 137 samples [26]; Schmidt: 13,292 genes and 200 samples [35]; Blenkiron: 17,941 genes and 128 samples [5]. In these datasets, we take histological grade as the phenotype of interest and consider estrogen receptor status and tumor size as potential confounders. Cell-cycle-related genes are known to discriminate low and high grade breast cancers irrespective of estrogen receptor status [26, 36]. Therefore, we compare the algorithms in their ability to detect specifically cell-cycle-related genes and not estrogen-regulated genes. To this end, we focused attention on two gene sets, one representing cell-cycle-related genes from the Reactome <http://www.reactome.org>, and another representing estrogen receptor (*ESR1*) upregulated genes [10]. The cell-cycle set showed negligible overlap with the *ESR1* gene set, however, we removed the few overlapping genes to ensure mutual exclusivity of the cell-cycle and *ESR1* sets.

## References

1. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J., Stratton, M.R.: Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**(1), 246–259 (2013)
2. Baufays, H.: Unification de techniques de sparation aveugle de sources avec application l'analyse de l'expression des gnes. Ecole Polytechnique de Louvain, Master thesis with Prof. P.-A. Absil (2011)
3. Bell, C.G., Teschendorff, A.E., Rakyan, V.K., Maxwell, A.P., Beck, S., Savage, D.A.: Genome-wide dna methylation analysis for diabetic nephropathy in type 1 diabetes mellitus. *BMC Med. Genomics* **3**, 33 (2010)
4. Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R., Gunderson, K.L.: Genome-wide DNA methylation profiling using the infinium assay. *Epigenomics* **1**(1), 177–200 (2009)
5. Blenkiron, C., Goldstein, L.D., Thorne, N.P., Spiteri, I., Chin, S.F., Dunning, M.J., Barbosa-Morais, N.L., Teschendorff, A.E., Green, A.R., Ellis, I.O., Tavar, S., Caldas, C., Miska, E.A.: Microrna expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biol.* **8**(10), R214 (2007)
6. Cardoso, J.F.: High-order contrasts for independent component analysis. *Neural Comput.* **11**(1), 157–192 (1999)
7. Consortium 1000 Genomes Project, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A.: An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**(7422), 56–65 (2012)
8. Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., Grf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Watson, P., Markowitz, F., Murphy, L., Ellis, I., Purushotham, A., Brresen-Dale, A.L., Brenton, J.D., Tavar, S., Caldas, C., Aparicio, S.: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**(7403), 346–352 (2012)
9. Deaton, A.M., Bird, A.: CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010–1022 (2011)
10. Doane, A.S., Danso, M., Lal, P., Donaton, M., Zhang, L., Hudis, C., Gerald, W.L.: An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen. *Oncogene* **25**(28), 3994–4008 (2006)
11. Feinberg, A.P., Vogelstein, B.: Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **301**(5895), 89–92 (1983)
12. Frigyesi, A., Veerla, S., Lindgren, D., Hoglund, M.: Independent component analysis reveals new and biologically significant structures in micro array data. *BMC Bioinformatics* **7**, 290 (2006)
13. Gao, Y., Church, G.: Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics* **21**(21), 3970–3975 (2005)
14. Huang, D.S., Zheng, C.H.: Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* **22**(15), 1855–1862 (2006)
15. Hyvaerinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley, New York (2001)
16. Johnson, W.E., Li, C., Rabinovic, A.: Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* **8**(1), 118–127 (2007)
17. Jones, P.A., Baylin, S.B.: The epigenomics of cancer. *Cell* **128**(4), 683–692 (2007)
18. Lee, S.I., Batzoglou, S.: Application of independent component analysis to microarrays. *Genome Biol.* **4**(11), R76 (2003)
19. Leek, J.T., Storey, J.D.: A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci. USA* **105**(48), 18, 718–18, 723 (2008)
20. Leek, J.T., Storey, J.D.: Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**(9), 1724–1735 (2007)

21. Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., Irizarry, R.A.: Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**(10), 733–739 (2010)
22. Liao, J.C., Boscolo, R., Yang, Y.L., Tran, L.M., Sabatti, C., Roychowdhury, V.P.: Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl. Acad. Sci. USA* **100**(26), 15,522–15,527 (2003)
23. Liebermeister, W.: Linear modes of gene expression determined by independent component analysis. *Bioinformatics* **18**(1), 51–60 (2002)
24. Liu, Y., Aryee, M.J., Padyukov, L., Fallin, M.D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., Shchetynsky, K., Scheynius, A., Kere, J., Alfredsson, L., Klareskog, L., Ekström, T.J., Feinberg, A.P.: Epigenome-wide association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.* **31**(2), 142–147 (2013)
25. Liu, N.W., Sanford, T., Srinivasan, R., Liu, J.L., Khurana, K., Aprelikova, O., Valero, V., Bechert, C., Worrell, R., Pinto, P.A., Yang, Y., Merino, M., Linehan, W.M., Bratslavsky, G.: Impact of ischemia and procurement conditions on gene expression in renal cell carcinoma. *Clin. Cancer Res.* **19**(1), 42–49 (2013)
26. Loi, S., Haibe-Kains, B., Desmedt, C., Lallemand, F., Tutt, A.M., Gillet, C., Ellis, P., Harris, A., Bergh, J., Foekens, J.A., Klijn, J.G., Larsimont, D., Buyse, M., Bontempi, G., Delorenzi, M., Piccart, M.J., Sotiriou, C.: Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J. Clin. Oncol.* **25**(10), 1239–1246 (2007)
27. Maegawa, S., Hinkal, G., Kim, H.S., Shen, L., Zhang, L., Zhang, J., Zhang, N., Liang, S., Donehower, L.A., Issa, J.P.: Widespread and tissue specific age-related dna methylation changes in mice. *Genome Res.* **20**(3), 332–340 (2010)
28. Martoglio, A.M., Miskin, J.W., Smith, S.K., MacKay, D.J.: A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. *Bioinformatics* **18**(12), 1617–1624 (2002)
29. Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A., Guhr, T., Stanley, H.E.: Random matrix approach to cross correlations in financial data. *Phys. Rev. E Stat. Nonlinear Soft Matter Phys.* **65**(6), 066,126 (2002)
30. Rakyan, V.K., Down, T.A., Maslau, S., Andrew, T., Yang, T.P., Beyan, H., Whittaker, P., McCann, O.T., Finer, S., Valdes, A.M., Leslie, R.D., Deloukas, P., Spector, T.D.: Human aging-associated dna hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res.* **20**(4), 434–439 (2010)
31. Rakyan, V.K., Down, T.A., Balding, D.J., Beck, S.: Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.* **12**(8), 529–541 (2011)
32. Rhodes, D.R., Chinnaiyan, A.M.: Integrative analysis of the cancer transcriptome. *Nat. Genet.* **37**, S31–S37 (2005)
33. Sainlez, M., Absil, P.-A., Teschendorff, A. Gene expression data analysis using spatiotemporal blind, source separation. In: *Proceedings of ESANN'2009*, pp. 159–164. (2009)
34. Sawyers, C.L.: The cancer biomarker problem. *Nature* **452**(7187), 548–552 (2008)
35. Schmidt, M., Bhm, D., von Trone, C., Steiner, E., Puhl, A., Pilch, H., Lehr, H.A., Hengstler, J.G., Kibl, H., Gehrman, M.: The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res.* **68**(13), 5405–5413 (2008)
36. Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Van de Vijver, M.J., Bergh, J., Piccart, M., Delorenzi, M.: Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst.* **98**(4), 262–272 (2006)
37. Stone, J.V., Porrill, J., Porter, N.R., Wilkinson, I.D.: Spatiotemporal independent component analysis of event-related fmri data using skewed probability density functions. *Neuroimage* **15** (2002)

38. Storey, J.D., Tibshirani, R.: Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**(16), 9440–9445 (2003)
39. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**(43), 15, 545–15, 550 (2005)
40. Swanton, C., Caldas, C.: From genomic landscapes to personalized cancer management—is there a roadmap? *Ann. N. Y. Acad. Sci.* **1210**, 34–44 (2010)
41. Teschendorff, A.E., Naderi, A., Barbosa-Morais, N.L., Caldas, C.: Pack: profile analysis using clustering and kurtosis to find molecular classifiers in cancer. *Bioinformatics* **22**(18), 2269–2275 (2006)
42. Teschendorff, A.E., Journe, M., Absil, P.A., Sepulchre, R., Caldas, C.: Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Comput. Biol.* **3**(8), e161 (2007)
43. Teschendorff, A.E., Menon, U., Gentry-Maharaj, A., Ramus, S.J., Gayther, S.A., Apostolidou, S., Jones, A., Lechner, M., Beck, S., Jacobs, I.J., Widschwendter, M.: An epigenetic signature in peripheral blood predicts active ovarian cancer. *PLoS ONE* **4**(12), e8274 (2009)
44. Teschendorff, A.E., Menon, U., Gentry-Maharaj, A., Ramus, S.J., Weisenberger, D.J., Shen, H., Campan, M., Noushmehr, H., Bell, C.G., Maxwell, A.P., Savage, D.A., Mueller-Holzner, E., Marth, C., Kocjan, G., Gayther, S.A., Jones, A., Beck, S., Wagner, W., Laird, P.W., Jacobs, I.J., Widschwendter, M.: Age-dependent dna methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.* **20**(4), 440–446 (2010)
45. Teschendorff, A.E., Zhuang, J., Widschwendter, M.: Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* **27**(11), 1496–1505 (2011)
46. The Cancer Genome Atlas Research Network: Integrated genomic analyses of ovarian carcinoma. *Nature* **474**(7353), 609–615 (2011)
47. Theis, F., Gruber, P., Keck, I., Meyer-Bäse, A., Lang, E.: Spatiotemporal blind source separation using double-sided approximate joint diagonalization. In: *Proceedings of EUSIPCO 2005*, Antalya, Turkey (2005)
48. Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Yu, J., Jatkoa, T., Berns, E.M., Atkins, D., Foekens, J.A.: Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**(9460), 671–679 (2005)
49. Zhang, X.W., Yap, Y.L., Wei, D., Chen, F., Danchin, A.: Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis. *Eur. J. Hum. Genet.* **13**(12), 1303–1311 (2005)
50. Zhang, S., Liu, C.C., Li, W., Shen, H., Laird, P.W., Zhou, X.J.: Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.* **40**(19), 9379–9391 (2012)
51. Zhuang, J., Widschwendter, M., Teschendorff, A.E.: A comparison of feature selection and classification methods in dna methylation studies using the illumina infinium platform. *BMC Bioinformatics* **13**, 59 (2012)