Spectral analysis of two-signed microarray expression data

Article *in* Mathematical Medicine and Biology · July 2007

DOI: 10.1093/imammb/dql030 · Source: PubMed

3 authors, including:

Gabriela Kalna
Beatson Institute for Cancer Research
151 PUBLICATIONS **3,570** CITATIONS

SEE PROFILE

Mathematical Medicine and Biology Page 1 of 18
doi:10.1093/imammb/dql030

Spectral analysis of two-signed microarray expression data

DESMOND J. HIGHAM† AND GABRIELA KALNA
Department of Mathematics, University of Strathclyde, Glasgow G1 1XH, UK

AND

J. KEITH VASS
The Beatson Institute for Cancer Research, Glasgow G61 1BD, UK

[Received on 22 June 2005; accepted on 23 June 2006]

We give a simple and informative derivation of a spectral algorithm for clustering and reordering complementary DNA microarray expression data. Here, expression levels of a set of genes are recorded simultaneously across a number of samples, with a positive weight reflecting up-regulation and a negative weight reflecting down-regulation. We give theoretical support for the algorithm based on a biologically justified hypothesis about the structure of the data, and illustrate its use on public domain data in the context of unsupervised tumour classification. The algorithm is derived by considering a discrete optimization problem and then relaxing to the continuous realm. We prove that in the case where the data have an inherent 'checkerboard' sign pattern, the algorithm will automatically reveal that pattern. Further, our derivation shows that the algorithm may be regarded as imposing a random graph model on the expression levels and then clustering from a maximum likelihood perspective. This indicates that the output will be tolerant to perturbations and will reveal 'near-checkerboard' patterns when these are present in the data. It is interesting to note that the checkerboard structure is revealed by the first (dominant) singular vectors—previous work on spectral methods has focussed on the case of nonnegative edge weights, where only the second and higher singular vectors are relevant. We illustrate the algorithm on real and synthetic data, and then use it in a tumour classification context on three different cancer data sets. Our results show that respecting the two-signed nature of the data (thereby distinguishing between up-regulation and down-regulation) reveals structures that cannot be gleaned from the absolute value data (where up- and down-regulation are both regarded as 'changes').

Keywords: bioinformatics; cDNA; checkerboard; clustering; data mining; maximum likelihood; microarray; reordering; singular value decomposition; tumour classification; unsupervised feature extraction.

1. Introduction

Spectral algorithms for dimension reduction and clustering are known to be useful in a range of areas of science and engineering. In particular, they have found success in bioinformatics applications involving gene/protein expression and interaction data sets (Grindrod & Kibble, 2004; Kluger *et al.*, 2003; Xing & Karp, 2001). The bipartite graph framework is a natural setting for many large gene expression data sets that are currently being generated from high-throughput microarray experiments. Our aim in this work is to develop, analyse and test the basis of a spectral algorithm for clustering such a bipartite graph. Our emphasis is on the case where the entire data set, combining both up- and down-regulation, is treated as a whole. From a graph theory perspective, this corresponds to allowing both positive and negative weights, whereas typical clustering algorithms assume that edge weights are nonnegative. We are interested in

†Email: djh@maths.strath.ac.uk

the big-picture question of how a spectral decomposition can be motivated from a biological perspective and what type of information can be extracted. For fine detail issues of how to process and validate the spectral output, we refer to Handl *et al.* (2005), Kluger *et al.* (2003), Yeung & Ruzzo (2001) and the references therein.

Our data are a rectangular array $A \in \mathbb{R}^{M\times N}$. The underlying bipartite graph has a set $G = \{g_1, g_2, \ldots, g_M\}$ of nodes, each of which has an edge to each member of a set $S = \{s_1, s_2, \ldots, s_N\}$ of nodes. The entry a_{ij} indicates the weight of the edge from the ith node in set G to the jth node in set S. In the microarray setting, G is a set of M genes and S is a set of N samples, and we assume that a_{ij} represents the relative expression level of gene i in sample j. If $a_{ij} > 0$ (respectively $a_{ij} < 0$), then gene i is relatively over-expressed (respectively under-expressed) in sample j. Further details are given in Section 5.

The matrix A is potentially large: $M = 20\,000$ genes and $N = 20$ samples are typical. To extract meaningful information, it is necessary to process and summarize the data. The approach examined here is motivated by the idea of bi-clustering simultaneously the genes and samples, i.e. 'we aim to split the genes into two or more groups and to split the samples into two or more groups such that for each group of genes and each group of samples, the expression levels are similar across genes and samples'. The motivation is that genes involved in a common process are likely to be active in a common set of samples. Thus, bi-clustering is an attempt to identify sets of related genes and the corresponding samples in which they are active/inactive. Further discussion of the bi-clustering approach and the justification based on the biological literature can be found in Kluger *et al.* (2003).

We note that it is common either to take absolute values of expression levels or to treat up- and down-regulation separately. In either case, a matrix of nonnegative weights is used. For example, in Kluger *et al.* (2003), the authors begin by stating that 'we will assume that the values in the matrix A_{ij} represent absolute levels and that all entries are nonnegative' (p. 704) and then give a justification for spectral bi-clustering. Later, though, they give results of computations involving positive and negative data based on log-ratios. One of our aims here is to justify the use of spectral bi-clustering in the presence of positive and negative data. We show in Section 5 that dealing directly with the two-signed data can reveal patterns that are lost when $|a_{ij}|$ is used.

Following the discussion above, our explicit working hypothesis is that the genes and samples can both be split into two groups $G = G_1 \cup G_2$ and $S = S_1 \cup S_2$ in such a way that

- genes in G_1 tend to be up-regulated for samples in S_1,
- genes in G_2 tend to be up-regulated for samples in S_2,
- genes in G_1 tend to be down-regulated for samples in S_2 and
- genes in G_2 tend to be down-regulated for samples in S_1.

More reasonably, we hope to find large subsets of genes and samples for which this type of behaviour is approximated.

As a starting point for deriving an algorithm, we let $p \in \mathbb{R}^M$ be an indicator vector that determines whether gene i is to be placed in group G_1 (so that $p_i = 1$) or G_2 (so that $p_i = -1$). Similarly, we let $q \in \mathbb{R}^N$ be an indicator vector that determines whether sample j is to be placed in group S_1 (so that $q_j = 1$) or S_2 (so that $q_j = -1$).

On the basis that

- we aim to have $a_{ij} \geqslant 0$ when $p_i = q_j = 1$ and when $p_i = q_j = -1$,
- we aim to have $a_{ij} \leqslant 0$ when $p_i = 1, q_j = -1$ and when $p_i = -1, q_j = 1$,

it is reasonable to choose p and q as solutions to the optimization problem

$$\max_{p_i \in \{\pm 1\},\, q_j \in \{\pm 1\}} \sum_{i=1}^{M} \sum_{j=1}^{N} p_i q_j a_{ij}. \tag{1}$$

To understand (1), note that the objective function is encouraging gene i and sample j to be placed in the same groups ($p_i q_j = 1$) when $a_{ij} > 0$ and in different groups ($p_i q_j = -1$) when $a_{ij} < 0$. This discrete optimization problem will be too hard to solve for a large data set, so we will 'relax' the problem by allowing $p \in \mathbb{R}^M$ and $q \in \mathbb{R}^N$. In doing so, we must restrict the size of p and q in some way (otherwise, the solution will involve unbounded values). Hence, we consider the problem

$$\max_{p \in \mathbb{R}^M,\, q \in \mathbb{R}^N} \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} p_i q_j a_{ij}}{\| D_{\mathrm{L}}^{\mathrm{weight}} p \|_2 \, \| D_{\mathrm{R}}^{\mathrm{weight}} q \|_2}. \tag{2}$$

Here, we use $\|\cdot\|_2$ to denote the Euclidean vector norm and also the induced matrix norm, and $D_{\mathrm{L}}^{\mathrm{weight}} \in \mathbb{R}^{M \times M}$ and $D_{\mathrm{R}}^{\mathrm{weight}} \in \mathbb{R}^{N \times N}$ are fixed diagonal weight matrices, with positive diagonal elements. Different choices of diagonal weight matrices will, in general, lead to different solutions p and q. Two natural choices for $D_{\mathrm{L}}^{\mathrm{weight}}$ and $D_{\mathrm{R}}^{\mathrm{weight}}$ are

unnormalized: $D_{\mathrm{L}}^{\mathrm{weight}} = I$ and $D_{\mathrm{R}}^{\mathrm{weight}} = I$,

row/column scaled: $D_{\mathrm{L}}^{\mathrm{weight}} = D_{\mathrm{gene}}^{\frac{1}{2}}$ and $D_{\mathrm{R}}^{\mathrm{weight}} = D_{\mathrm{sample}}^{\frac{1}{2}}$.

Here, I denotes an identity matrix of the appropriate dimension and D_{gene} and D_{sample} denote the diagonal 'absolute weight sum for gene' and 'absolute weight sum for sample' matrices; so $D_{\mathrm{gene}} \in \mathbb{R}^{M \times M}$ with $(D_{\mathrm{gene}})_{ii} = \sum_{j=1}^{N} |a_{ij}|$ and $D_{\mathrm{sample}} \in \mathbb{R}^{N \times N}$ with $(D_{\mathrm{sample}})_{jj} = \sum_{i=1}^{M} |a_{ij}|$.

At this stage, we recall that any matrix $B \in \mathbb{R}^{M \times N}$ has a singular value decomposition (SVD) of the form $B = U \Sigma V^{\top}$, where $U \in \mathbb{R}^{M \times M}$ and $V \in \mathbb{R}^{N \times N}$ are orthogonal and $\Sigma \in \mathbb{R}^{M \times N}$ has its only nonzero elements $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r > 0$ ordered from high-to-low along the diagonal, and r is the rank of B (see, e.g. Horn & Johnson, 1985). The kth columns of U and V are referred to as the kth left and the kth right singular vectors of B, respectively, and we denote them by $u^{[k]}$ and $v^{[k]}$.

THEOREM 1.1 Problem (2) is solved by taking $p = (D_{\mathrm{L}}^{\mathrm{weight}})^{-1} u^{[1]}$ and $q = (D_{\mathrm{R}}^{\mathrm{weight}})^{-1} v^{[1]}$, where $u^{[1]}$ and $v^{[1]}$ are the first left and right singular vectors of $(D_{\mathrm{L}}^{\mathrm{weight}})^{-1} A (D_{\mathrm{R}}^{\mathrm{weight}})^{-1}$.

Proof. If $(D_{\mathrm{L}}^{\mathrm{weight}})^{-1} A (D_{\mathrm{R}}^{\mathrm{weight}})^{-1}$ has SVD given by $U \Sigma V^{\top}$, then making the substitutions $p = (D_{\mathrm{L}}^{\mathrm{weight}})^{-1} U x$ and $q = (D_{\mathrm{R}}^{\mathrm{weight}})^{-1} V y$ reduces Problem (2) to

$$\max_{x \in \mathbb{R}^M,\, y \in \mathbb{R}^N} \frac{\sum_{k=1}^{r} \sigma_k x_k y_k}{\|x\|_2 \, \|y\|_2}.$$

A solution is clearly found by setting $x_1 = y_1 = 1$ and all other x and y components equal to zero. This translates to $p = (D_{\mathrm{L}}^{\mathrm{weight}})^{-1} u^{[1]}$ and $q = (D_{\mathrm{R}}^{\mathrm{weight}})^{-1} v^{[1]}$. □

Theorem 1.1 shows how to solve the relaxed problem (2). If the relaxed solutions $p \in \mathbb{R}^M$ and $q \in \mathbb{R}^N$ have components that divide naturally into two groups, then this suggests a bipartitioning for the genes and samples. Moreover, regarding p and q as projections into one dimension, we may opt,instead, for a 2D projection. A similar analysis to that giving Theorem 1.1 shows that looking for

the 'next best' directions for projection (i.e. solving (1.1) over directions orthogonal to p and q) leads to $(D_{\mathrm{L}}^{\mathrm{weight}})^{-1}u^{[2]}$ and $(D_{\mathrm{R}}^{\mathrm{weight}})^{-1}v^{[2]}$. So, plotting $\{((D_{\mathrm{L}}^{\mathrm{weight}})^{-1}u^{[1]})_i, ((D_{\mathrm{L}}^{\mathrm{weight}})^{-1}u^{[2]})_i\}_{i=1}^{M}$ as points in 2D gives a way of visualizing the data—genes that appear close together can be regarded as similar. Similarly, plotting $\{((D_{\mathrm{R}}^{\mathrm{weight}})^{-1}v^{[1]})_i, ((D_{\mathrm{R}}^{\mathrm{weight}})^{-1}v^{[2]})_i\}_{i=1}^{N}$ displays the samples.

Alternatively, we may use p and q to 'reorder' the graph (Grindrod, 2002; Grindrod & Kibble, 2004; Higham, 2003). Reordering the genes according to components of p (i.e. placing gene i before gene j if and only if $p_i < p_j$) and reordering the samples according to components of q lead to a new matrix in which nearby rows and columns should exhibit similar behaviour.

In Sections 2 and 5, we illustrate these uses of the singular vectors.

The idea of using the SVD for bi-clustering appears in Dhillon (2001) and Kluger *et al.* (2003) and has close connections with principal component analysis (Wall *et al.*, 2003; Yeung & Ruzzo, 2001). The spectral approach can be used to preprocess data before applying some other clustering algorithm, although empirical tests on gene expression data in Yeung & Ruzzo (2001) did not support this as a general technique. Dhillon's (2001) derivation is similar to ours in the sense that it is based on relaxation, but, as it is designed for document/word clustering, it assumes nonnegative edge weights. We believe that the derivation above has the virtue of simplicity and generality and, as we show in Section 2, it reveals an important property of the relaxed solution.

2. Recovering a perfect checkerboard structure

In this section, we show that the solution in Theorem 1.1 reveals an interesting structure whenever it is present in the data. We begin with some preparatory notation and definitions.

We let $d_{\mathrm{gene}} \in \mathbb{R}^M$ and $d_{\mathrm{sample}} \in \mathbb{R}^N$ to denote the 'absolute out-degree' and the 'absolute in-degree' vectors, i.e. $(d_{\mathrm{gene}})_i = (D_{\mathrm{gene}})_{ii}$ and $(d_{\mathrm{sample}})_j = (D_{\mathrm{sample}})_{jj}$, and, for convenience, we assume that all these quantities are strictly positive. (The case of a zero absolute in or out degree is easily handled.) We use superscripts on vectors to denote componentwise powers, so, e.g. $d_{\mathrm{gene}}^{\frac{1}{2}}$ is the vector in $\mathbb{R}^M$ with ith component $((d_{\mathrm{gene}})_i)^{\frac{1}{2}}$.

We also use $\mathrm{Diag}_{\pm}^{M\times M}$ to denote the set of all $M \times M$ diagonal matrices with entries of ± 1 on the diagonal, and let $|\cdot|$ denote the componentwise absolute values, so $(|A|)_{ij} = |a_{ij}|$.

DEFINITION 2.1 The matrix $A \in \mathbb{R}^{M\times N}$ has a 'plus–minus checkerboard structure', if there exist $D_{\mathrm{L}} \in \mathrm{Diag}_{\pm}^{M\times M}$ and $D_{\mathrm{R}} \in \mathrm{Diag}_{\pm}^{N\times N}$ such that $D_{\mathrm{L}}AD_{\mathrm{R}} = |A|$.

In words, this definition says that it is possible to scale each row by ± 1 and each column by ± 1 in such a way that all the entries become nonnegative. This corresponds to there being a perfect splitting of $G = G_1 \cup G_2$ and $S = S_1 \cup S_2$, as in our working hypothesis in Section 1.

There is an equivalent definition that helps to emphasize this correspondence. Here, we use $\mathrm{Perm}^{M\times M}$ to denote the set of all $M \times M$ permutation matrices.

DEFINITION 2.2 The matrix $A \in \mathbb{R}^{M\times N}$ has a 'plus–minus checkerboard structure', if there exist $P_{\mathrm{L}} \in \mathrm{Perm}^{M\times M}$ and $P_{\mathrm{R}} \in \mathrm{Perm}^{N\times N}$ such that, for some $1 \leqslant \widehat{i} \leqslant M$ and $1 \leqslant \widehat{j} \leqslant N$,

$$(P_{\mathrm{L}}AP_{\mathrm{R}})_{ij} \begin{cases} \geqslant 0, & \text{for } 1 \leqslant i \leqslant \widehat{i},\ 1 \leqslant j \leqslant \widehat{j}, \\ \geqslant 0, & \text{for } \widehat{i} < i \leqslant M,\ \widehat{j} < j \leqslant N, \\ \leqslant 0, & \text{otherwise.} \end{cases}$$

In words, this definition says that it is possible to permute rows and columns in such a way that the permuted matrix has a 2×2 block sign pattern:

1, 1 block: entries in rows 1 to $\widehat{i}$ and columns 1 to $\widehat{j}$ are $\geqslant 0$,

1, 2 block: entries in rows 1 to $\widehat{i}$ and columns $\widehat{j}+1$ to N are $\leqslant 0$,

2, 1 block: entries in rows $\widehat{i}+1$ to M and columns 1 to $\widehat{j}$ are $\leqslant 0$,

2, 2 block: entries in rows $\widehat{i}+1$ to M and columns $\widehat{j}+1$ to N are $\geqslant 0$.

The two equivalent definitions are connected as follows: the permutation P_{L} may be taken as that which reorders D_{L} in such a way that $P_{\mathrm{L}} D_{\mathrm{L}} \in \mathrm{Diag}_{\pm}^{M\times M}$ has $\widehat{i}$ contiguous $+1$s on the diagonal followed by $M-\widehat{i}$ contiguous -1s and, similarly, the permutation P_{R} may be taken as that which reorders D_{R} in such a way that $D_{\mathrm{R}} P_{\mathrm{R}} \in \mathrm{Diag}_{\pm}^{N\times N}$ has $\widehat{j}$ contiguous $+1$s on the diagonal followed by $N-\widehat{j}$ contiguous -1s.

We now show that the SVD approach is able to recover the checkerboard structure when it is present in the data. We follow the convention that inequalities $x \geqslant 0$ and $X \geqslant 0$ for vectors and matrices are to be interpreted componentwise.

THEOREM 2.1 If $A \in \mathbb{R}^{M\times N}$ has a plus–minus checkerboard structure, then Problem (2) has a solution such that $D_{\mathrm{L}} p \geqslant 0$ and $D_{\mathrm{R}} q \geqslant 0$.

Proof. Setting $\widehat{x} = D_{\mathrm{L}} p$ and $\widehat{y} = D_{\mathrm{R}} q$, Problem (2) becomes

$$\max_{\widehat{x}\in\mathbb{R}^M,\, \widehat{y}\in\mathbb{R}^N} \frac{\widehat{x}^{\top} D_{\mathrm{L}} A D_{\mathrm{R}} \widehat{y}}{\| D_{\mathrm{L}}^{\mathrm{weight}} \widehat{x}\|_2 \, \| D_{\mathrm{R}}^{\mathrm{weight}} \widehat{y}\|_2}.$$

Now, from Theorem 1.1, this problem is solved by $\widehat{x} = (D_{\mathrm{L}}^{\mathrm{weight}})^{-1} u^{[1]}$ and $\widehat{y} = (D_{\mathrm{R}}^{\mathrm{weight}})^{-1} v^{[1]}$, where $u^{[1]}$ and $v^{[1]}$ are the first left and right singular vectors of $(D_{\mathrm{L}}^{\mathrm{weight}})^{-1} D_{\mathrm{L}} A D_{\mathrm{R}} (D_{\mathrm{R}}^{\mathrm{weight}})^{-1}$. Since $D_{\mathrm{L}} A D_{\mathrm{R}} \geqslant 0$ and because the singular vectors of a matrix B are eigenvectors of the matrices $B^{\top}B$ and $BB^{\top}$, Perron theory (see, e.g. Horn & Johnson, 1985, Theorem 8.3.1) shows that $u^{[1]} \geqslant 0$ and $v^{[1]} \geqslant 0$. Hence, $\widehat{x} = D_{\mathrm{L}} p \geqslant 0$ and $\widehat{y} = D_{\mathrm{R}} q \geqslant 0$. □

Theorem 2.1 shows that the sign patterns in the relaxed solutions p and q match the sign patterns in $D_{\mathrm{L}} \in \mathrm{Diag}_{\pm}^{M\times M}$ and $D_{\mathrm{R}} \in \mathrm{Diag}_{\pm}^{N\times N}$. Hence, reordering or partitioning A according to the components of $u^{[1]}$ and $v^{[1]}$ will reveal the sign pattern in A by forming contiguous blocks of nonnegative and nonpositive elements.

For the row/column-scaled case, where $D_{\mathrm{L}}^{\mathrm{weight}} = D_{\mathrm{gene}}^{\frac{1}{2}}$ and $D_{\mathrm{R}}^{\mathrm{weight}} = D_{\mathrm{sample}}^{\frac{1}{2}}$, we can get further insight by explicitly identifying the first singular vectors.

LEMMA 2.1 If $A \in \mathbb{R}^{M\times N}$ has a plus–minus checkerboard structure, then $\| D_{\mathrm{gene}}^{-\frac{1}{2}} A D_{\mathrm{sample}}^{-\frac{1}{2}} \|_2 = 1$.

Proof. Using the facts that (a) $\|D_1 A D_2\|_2 = \|A\|_2$ for any $D_1 \in \mathrm{Diag}_{\pm}^{M\times M}$ and $D_2 \in \mathrm{Diag}_{\pm}^{N\times N}$ and (b) diagonal matrices commute, we have

$$\left\| D_{\mathrm{gene}}^{-\frac{1}{2}} A D_{\mathrm{sample}}^{-\frac{1}{2}} \right\|_2 = \left\| D_{\mathrm{L}} D_{\mathrm{gene}}^{-\frac{1}{2}} A D_{\mathrm{sample}}^{-\frac{1}{2}} D_{\mathrm{R}} \right\|_2 = \left\| D_{\mathrm{gene}}^{-\frac{1}{2}} D_{\mathrm{L}} A D_{\mathrm{R}} D_{\mathrm{sample}}^{-\frac{1}{2}} \right\|_2 = \left\| D_{\mathrm{gene}}^{-\frac{1}{2}} |A| D_{\mathrm{sample}}^{-\frac{1}{2}} \right\|_2.$$

Now,

$$\left\| D_{\text{gene}}^{-\frac{1}{2}} |A| D_{\text{sample}}^{-\frac{1}{2}} \right\|_2 = \left\| \begin{bmatrix} D_{\text{gene}}^{-\frac{1}{2}} & 0 \\ 0 & D_{\text{sample}}^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} 0 & |A| \\ |A|^\top & 0 \end{bmatrix} \begin{bmatrix} D_{\text{gene}}^{-\frac{1}{2}} & 0 \\ 0 & D_{\text{sample}}^{-\frac{1}{2}} \end{bmatrix} \right\|_2.$$

The matrix on the right-hand side has an eigenvalue equal to 1 corresponding to the eigenvector

$$\begin{bmatrix} d_{\text{gene}}^{-\frac{1}{2}} \\ d_{\text{sample}}^{-\frac{1}{2}} \end{bmatrix},$$

and is similar to the matrix

$$\begin{bmatrix} D_{\text{gene}}^{-1} & 0 \\ 0 & D_{\text{sample}}^{-1} \end{bmatrix} \begin{bmatrix} 0 & |A| \\ |A|^\top & 0 \end{bmatrix},$$

which has absolute row sums equal to one. The result follows. □

THEOREM 2.2 If $A \in \mathbb{R}^{M\times N}$ has a plus–minus checkerboard structure, then the largest singular value of $\left\| D_{\text{gene}}^{-\frac{1}{2}} A D_{\text{sample}}^{-\frac{1}{2}} \right\|_2$ is $\sigma_1 = 1$, and the corresponding left and right singular vectors of $D_{\text{gene}}^{-\frac{1}{2}} A D_{\text{sample}}^{-\frac{1}{2}}$ are

$$u^{[1]} = \frac{D_{\text{L}} d_{\text{gene}}^{\frac{1}{2}}}{\left\| d_{\text{gene}}^{\frac{1}{2}} \right\|_2} \quad \text{and} \quad v^{[1]} = \frac{D_{\text{R}} d_{\text{sample}}^{\frac{1}{2}}}{\left\| d_{\text{sample}}^{\frac{1}{2}} \right\|_2}.$$

Proof. Lemma 2.1 shows that $\sigma_1 = 1$. Letting $\mathbf{1}_s$ denote the vector in $\mathbb{R}^s$ with all elements equal to one, we note that

$$A D_{\text{R}} \mathbf{1}_N = D_{\text{L}} d_{\text{gene}} \quad \text{and} \quad A^\top D_{\text{L}} \mathbf{1}_M = D_{\text{R}} d_{\text{sample}}. \tag{3}$$

Hence,

$$\begin{aligned} \left(D_{\text{gene}}^{-\frac{1}{2}} A D_{\text{sample}}^{-\frac{1}{2}} \right)^\top \left(D_{\text{gene}}^{-\frac{1}{2}} A D_{\text{sample}}^{-\frac{1}{2}} \right) D_{\text{R}} d_{\text{sample}}^{\frac{1}{2}} &= D_{\text{sample}}^{-\frac{1}{2}} A^\top D_{\text{gene}}^{-1} A D_{\text{R}} D_{\text{sample}}^{-\frac{1}{2}} d_{\text{sample}}^{\frac{1}{2}} \\ &= D_{\text{sample}}^{-\frac{1}{2}} A^\top D_{\text{gene}}^{-1} A D_{\text{R}} \mathbf{1}_N \\ &= D_{\text{sample}}^{-\frac{1}{2}} A^\top D_{\text{gene}}^{-1} D_{\text{L}} d_{\text{gene}} \\ &= D_{\text{sample}}^{-\frac{1}{2}} A^\top D_{\text{L}} \mathbf{1}_M \\ &= D_{\text{sample}}^{-\frac{1}{2}} D_{\text{R}} d_{\text{sample}} \\ &= D_{\text{R}} d_{\text{sample}}^{\frac{1}{2}}. \end{aligned}$$

Thus, $v^{[1]} = D_{\text{R}} d_{\text{sample}}^{\frac{1}{2}} / \left\| d_{\text{sample}}^{\frac{1}{2}} \right\|_2$. A proof for $u^{[1]}$ follows similarly. □

COROLLARY 2.1 If $A \in \mathbb{R}^{M\times N}$ has a plus–minus checkerboard structure, then the row/column-scaled version of (2), with $D_{\mathrm{L}}^{\mathrm{weight}} = D_{\mathrm{gene}}^{\frac{1}{2}}$ and $D_{\mathrm{R}}^{\mathrm{weight}} = D_{\mathrm{sample}}^{\frac{1}{2}}$, has solution $p = D_{\mathrm{L}}\mathbf{1}_M / \left\| d_{\mathrm{gene}}^{\frac{1}{2}} \right\|_2$ and $q = D_{\mathrm{R}}\mathbf{1}_N / \left\| d_{\mathrm{sample}}^{\frac{1}{2}} \right\|_2$.

Proof. From Theorems 1.1 and 2.2, we have

$$p = (D_{\mathrm{L}}^{\mathrm{weight}})^{-1} u^{[1]} = D_{\mathrm{gene}}^{-\frac{1}{2}} D_{\mathrm{L}} \frac{d_{\mathrm{gene}}^{\frac{1}{2}}}{\left\| d_{\mathrm{gene}}^{\frac{1}{2}} \right\|_2} = \frac{D_{\mathrm{L}}\mathbf{1}_M}{\left\| d_{\mathrm{gene}}^{\frac{1}{2}} \right\|_2}$$

and

$$q = (D_{\mathrm{R}}^{\mathrm{weight}})^{-1} v^{[1]} = D_{\mathrm{sample}}^{-\frac{1}{2}} D_{\mathrm{R}} \frac{d_{\mathrm{sample}}^{\frac{1}{2}}}{\left\| d_{\mathrm{sample}}^{\frac{1}{2}} \right\|_2} = \frac{D_{\mathrm{R}}\mathbf{1}_N}{\left\| d_{\mathrm{sample}}^{\frac{1}{2}} \right\|_2}.$$

□

Corollary 2.1 shows that in the row/column-scaled case, if A has a plus–minus checkerboard structure, then the relaxed solutions p and q 'do not distinguish between individual members within the same groups', i.e. all genes get mapped to $\pm 1/\left\| d_{\mathrm{gene}}^{\frac{1}{2}} \right\|_2$ and all samples get mapped to $\pm 1/\left\| d_{\mathrm{sample}}^{\frac{1}{2}} \right\|_2$. From this point of view, the first singular vectors are 'only' concerned with the checkerboard structure and not with any other aspect of the interactions.

3. A maximum likelihood argument

In this section, we show that the optimization problem (1) may be interpreted from a maximum likelihood viewpoint. This probabilistic interpretation suggests that the structure-revealing property shown in Theorem 2.1 should be robust to noise, i.e. near-checkerboard patterns should be found if they are present in the data. The idea of interpreting spectral reordering from a maximum likelihood viewpoint was first suggested in Higham (2003), and the topic of random graph models for bioinformatics data sets was discussed in, e.g. Grindrod (2002), Grindrod & Kibble (2004), Morrison *et al.* (2006), Przulj *et al.* (2004) and Thomas *et al.* (2003).

For simplicity, we restrict our attention to the case where entries in a_{ij} are either -1, 0 or 1. The theory extends to general $a_{ij} \in \mathbb{R}$, but the details become more cumbersome.

Given indicator vectors with components $p_i \in \{\pm 1\}$ and $q_j \in \{\pm 1\}$ for $1 \leqslant i \leqslant M$ and $1 \leqslant j \leqslant N$, consider the class of random matrices $A \in \mathbb{R}^{M\times N}$ with entries $a_{ij} \in \{-1, 0, 1\}$ drawn independently with probabilities

$$\mathbb{P}(a_{ij} = x) = K \mathrm{e}^{\alpha p_i q_j x}, \tag{4}$$

where $\alpha > 0$ is a fixed parameter and $K = 1/(\mathrm{e}^{\alpha} + 1 + \mathrm{e}^{-\alpha})$ is a normalizing constant. In other words, when a_{ij} represents a gene–sample weight from the set G_1 to the set S_1 or from the set G_2 to the set S_2, we have

$$a_{ij} = \begin{cases} 1 & \text{with probability } K\mathrm{e}^{\alpha}, \\ 0 & \text{with probability } K, \\ -1 & \text{with probability } K\mathrm{e}^{-\alpha}, \end{cases}$$

whereas when a_{ij} represents a gene–sample weight from the set G_1 to the set S_2 or from the set G_2 to the set S_1, we have

$$a_{ij} = \begin{cases} 1 & \text{with probability } K\mathrm{e}^{-\alpha}, \\ 0 & \text{with probability } K, \\ -1 & \text{with probability } K\mathrm{e}^{\alpha}. \end{cases}$$

Suppose now that we are given the elements a_{ij} in an instance of such a random matrix and we wish to recover the indicator vectors p and q. The maximum likelihood solution is found by solving

$$\max_{p_i \in \{\pm 1\},\, q_j \in \{\pm 1\}} \prod_{i=1}^{M} \prod_{j=1}^{N} \mathrm{e}^{\alpha p_i q_j a_{ij}}. \tag{5}$$

Taking logs in this problem converts it to the form (1), and we conclude that the approach of using the first left and right singular vectors to reorder the matrix may be regarded as a 'relaxed maximum likelihood approach', based on the random matrix model above.

From this viewpoint, the algorithm may be regarded as hypothesizing a particular random model for the weights and, given data, finding the bipartition that best fits that data. We note that the probability of a zero weight is the same in the two cases $p_i q_j = \pm 1$, suggesting that the zero-weighted edges are equally likely to be found anywhere in the graph.

A further point of interest is that when the data values $\{a_{ij}\}$ represent log-ratios (as described in Section 5), the probability defined in (4) has a natural, linear dependence on the underlying ratios. This gives further support for the standard practice of applying logs to complementary DNA (cDNA) expression ratios.

4. Numerical tests on checkerboard structure

Our aim in this section is to illustrate the relevance of the analysis in Sections 2 and 3 in terms of recovering a checkerboard structure and also to test the performance of the SVD approach in the presence of noise.

Figure 1 illustrates the bi-clustering algorithm on some synthetic test data. In the upper left picture, we show a matrix in $\mathbb{R}^{M\times N}$, corresponding to $M = 40$ genes and $N = 20$ samples. The matrix was computed as $W = D_{\mathrm{L}} G D_{\mathrm{R}}$, where $G \in \mathbb{R}^{M\times N}$ has elements given by the absolute value of calls to a $N(0, 1)$ pseudorandom number generator, and $D_{\mathrm{L}} \in \mathrm{Diag}_{\pm}^{M\times M}$ and $D_{\mathrm{R}} \in \mathrm{Diag}_{\pm}^{N\times N}$ are chosen arbitrarily. Hence, this matrix has a perfect checkerboard structure. The matrix elements are displayed in a grey-scale from light (most negative) to dark (most positive). The picture below this shows the same matrix with genes and samples reordered according to the weighted left and right singular vectors $D_{\mathrm{gene}}^{-\frac{1}{2}} u^{[1]}$ and $D_{\mathrm{sample}}^{-\frac{1}{2}} v^{[1]}$ of $D_{\mathrm{gene}}^{-\frac{1}{2}} W D_{\mathrm{sample}}^{-\frac{1}{2}}$. Below that we show the sign pattern of the reordered matrix. Here, light is negative and dark is positive. We see that, as proved in Corollary 2.1, the first singular vectors recover the checkerboard structure. Components of the weighted left and right singular values are shown at the foot of the figure. The middle-top picture shows a matrix constructed in a similar manner, except that eight arbitrarily chosen elements of G were multiplied by -1 before the transformation to $D_{\mathrm{L}} G D_{\mathrm{R}}$ was applied. By construction, this matrix is close to having a checkerboard structure in the sense that it may be row/column permuted to have most elements following the desired pattern. Beneath this, we show a repeat of the computations described above, and we see that the spectral approach does a good job of displaying the structure. The sign pattern of the reordered matrix is violated only in eight places. Moreover, these 'ill-fitting' entries occur close to the boundary, emphasizing that

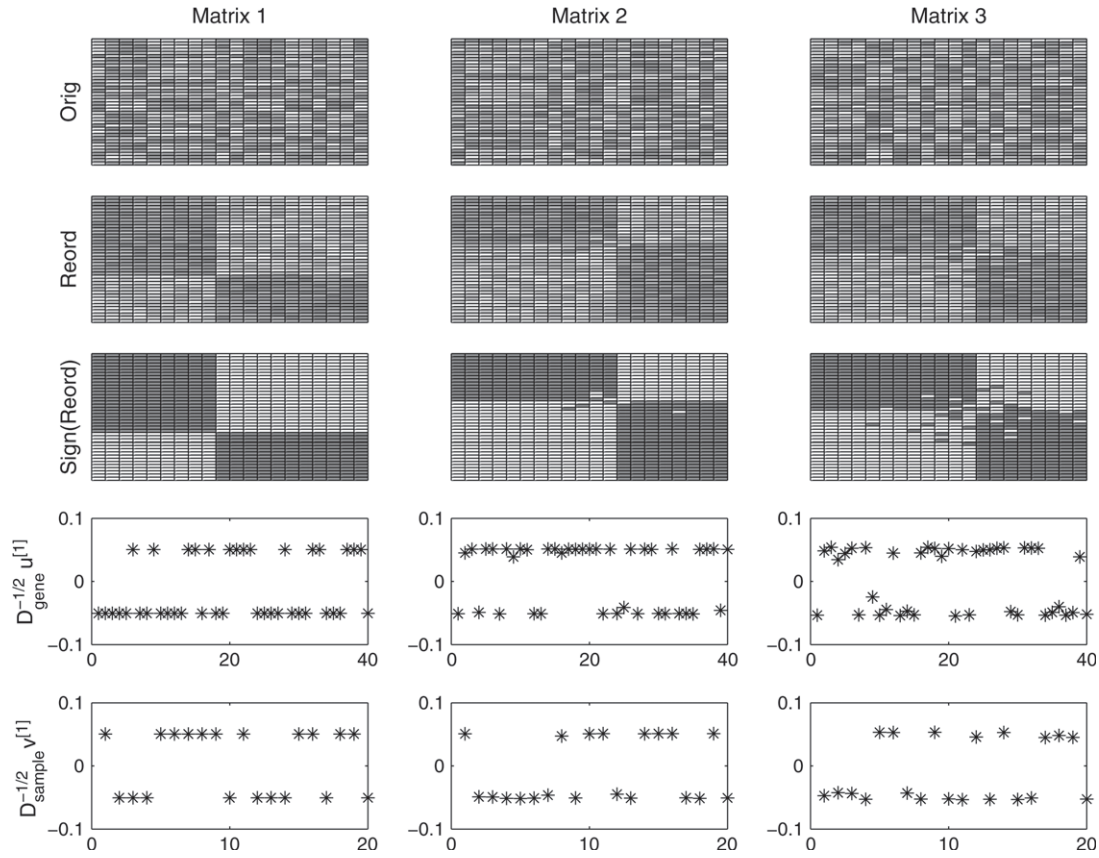

FIG. 1. Top left: a matrix with perfect plus–minus checkerboard structure. Below this: the same matrix, reordered using first singular vectors. Below this: the sign pattern of the reordered matrix. Below this: the weighted first left singular vector, $D_{\text{gene}}^{-\frac{1}{2}}u^{[1]}$. Below this: the weighted first right singular vector, $D_{\text{sample}}^{-\frac{1}{2}}v^{[1]}$. Middle and right columns: corresponding results for matrices that are close to the one having plus–minus checkerboard structure.

the corresponding genes and samples have a less clear-cut categorization. The third picture in the top row of Fig. 1 shows a similar experiment where 32 entries of G were sign-flipped. Here, the reordered sign pattern shows 32 deviations from exact checkerboard structure.

255 Next, we perform a large-scale experiment where the level of noisy information is closely controlled and the performance is quantified. We use synthetic data of the form

$$\begin{bmatrix} E & -E \\ -E & E \end{bmatrix} + sK \in \mathbb{R}^{100\times 20}, \tag{6}$$

where $E \in \mathbb{R}^{50\times 10}$ is a matrix with all elements equal to 1, $K \in \mathbb{R}^{100\times 20}$ is a matrix with every entry equal to an independent sample from the $N(0, 1)$ distribution and $s \in \mathbb{R}$ is a scaling factor. Here, with $s = 0$, the data correspond to 100 genes and 20 samples forming a perfect checkerboard structure with

equal-sized blocks. We are interested in quantifying the performance of the algorithm as s is increased from zero, causing the checkerboard structure to be compromised. To measure the information content in the first weighted left singular vector, $D_{\text{gene}}^{-\frac{1}{2}}u^{[1]}$, we assume that the signs of its components are used to assign genes to one of the two groups, and we measure the relative number of misclassifications. More precisely, our relative error measure is

$$\frac{1}{2\times 100}\min\left(\left\|\operatorname{sign}\left(D_{\text{gene}}^{-\frac{1}{2}}u^{[1]}\right)-\begin{bmatrix}\mathbf{1}_{50}\\ -\mathbf{1}_{50}\end{bmatrix}\right\|_1,\left\|\operatorname{sign}\left(D_{\text{gene}}^{-\frac{1}{2}}u^{[1]}\right)+\begin{bmatrix}\mathbf{1}_{50}\\ -\mathbf{1}_{50}\end{bmatrix}\right\|_1\right), \tag{7}$$

where sign denotes the componentwise sign function and $\|\cdot\|_1$ is the L_1 vector norm. Here, $[\mathbf{1}_{50}, -\mathbf{1}_{50}]^\top$ is an indicator vector for the target classification and the min operation allows for the fact that there are two ways to assign genes to groups based on the sign pattern of $D_{\text{gene}}^{-\frac{1}{2}}u^{[1]}$. The analogous measure was used to judge $D_{\text{sample}}^{-\frac{1}{2}}v^{[1]}$.

The upper plots in Fig. 2 show the behaviour of the error measure as the standard deviation s increases from 0 to 20. Here, for each s, we generated 10^4 data matrices. The asterisks show the average

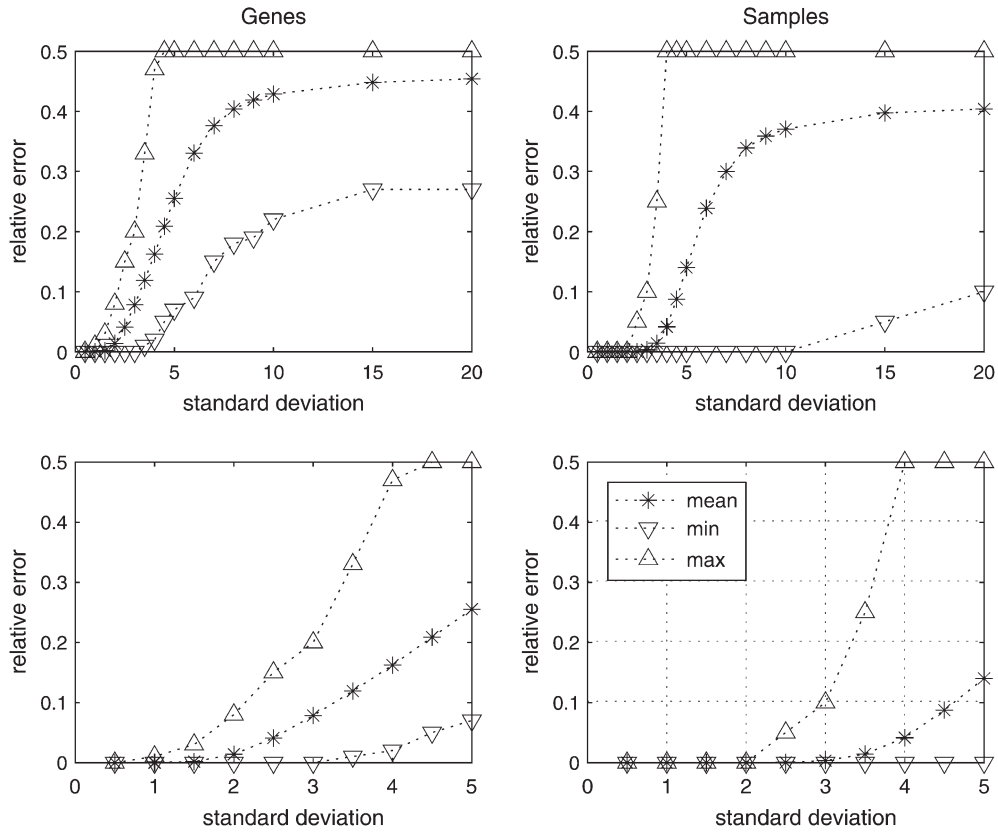

FIG. 2. The relative error measure (7) from 10^4 data matrices of the form (6). Left: genes. Right: samples. The lower pictures zoom in on a smaller range of s values.

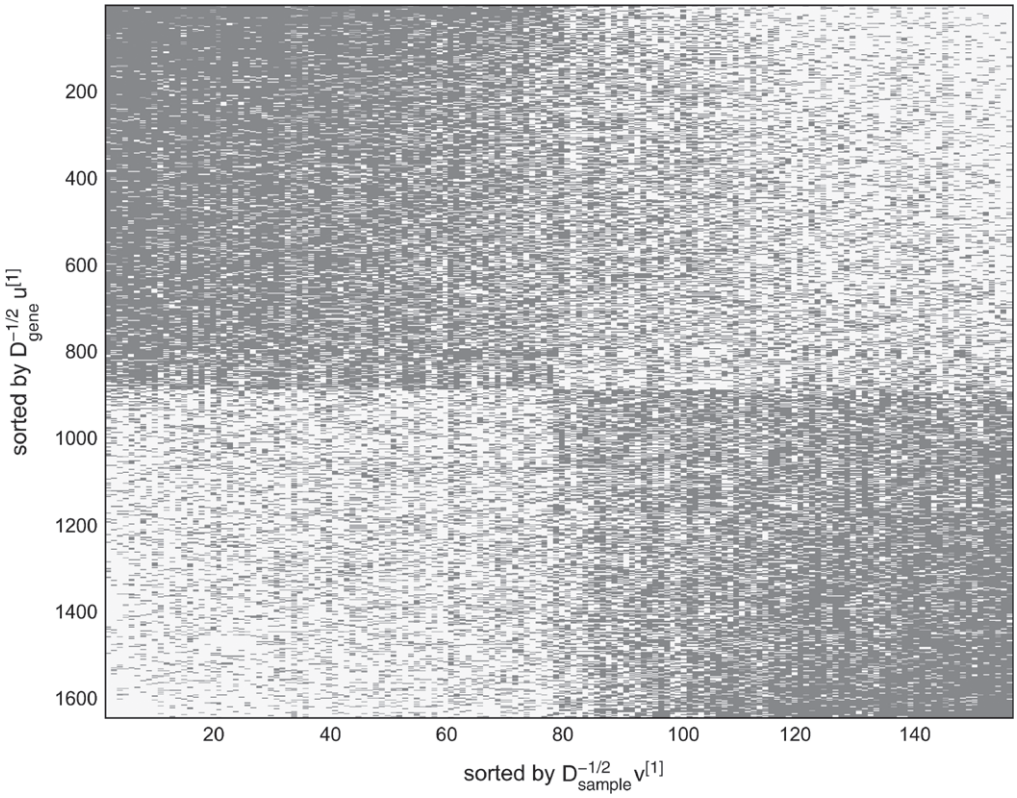

FIG. 3. Checkerboard structure in a liver cancer data set from Chen *et al.* (2002) comprising 1648 genes and 156 samples.

error over all data matrices (standard errors are smaller than the symbols used) and the triangles show the maximum and the minimum error observed for each s. The lower plots in Fig. 2 zoom in on the range $0 \leqslant s \leqslant 5$. We see that the gene classification remains accurate (with an error less than 0.1) for up to three standard deviations of added noise. As we would expect, the sample error, which is based on more information, remains accurate over a larger range of noise levels.

Figure 3 illustrates the algorithm on liver cancer microarray data from Chen *et al.* (2002), as available at http://genome-www.stanford.edu/hcc/supplement.shtml. Further details of how these type of data are generated are given in Section 5. In this data set, there are 156 samples of which 82 are hepatocellular carcinomas and 74 are nontumour liver tissues. The 1648 genes were chosen in Chen *et al.* (2002) to be useful at distinguishing between the two sample types. The figure plots the sign pattern of the reordered matrix. In this case, there are many zero elements, which are displayed as grey. The algorithm clearly reveals a near-checkerboard structure that correctly separates the two types of sample.

5. Results for tumour classification

We now illustrate the spectral clustering algorithm on three public domain cDNA microarray data sets. These involve lymphoma, prostate cancer and lung cancer. In each case, the samples correspond to

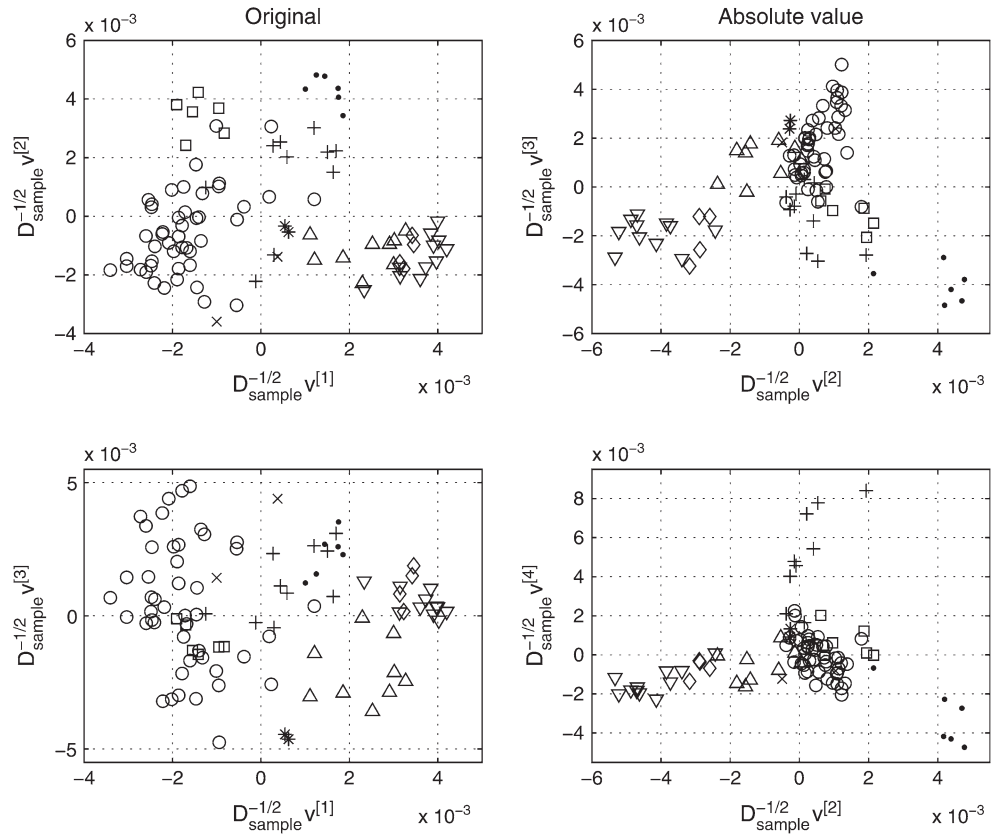

FIG. 4. Lymphoma—scatter plots of the three dominant singular vectors: DLBCL (circles), CLL (triangles down), FL (triangles up), B cells (pluses), T cells (dots), transformed cell lines (squares), resting blood B cells (diamonds), germinal centre B cells (stars) and normal lymph node/tonsil (crosses). Left: original log-ratio data. Right: absolute values of the log-ratio data.

tissues from different patients, and information is known about the existence of different sample types. Hence, we focus here on the ability of the algorithm to identify meaningful sample clusters that are consistent with the biological literature. We begin with a brief description of how the cDNA data are generated.

Microarrays are created by spotting genes from a genome on to a glass microscope slide. Cells are grown under two different conditions: experiment and control. The mRNA is isolated from them and converted to cDNA. Red and green fluorescent dyes are used to distinguish the experiment and control cDNA, which is then hybridized with the microarray. Two computerized images are produced by scanning the green-labelled and red-labelled cDNA. Data are collected as numerical values for each colour and the logarithm of the red/green ratio is calculated. Hence, a log-ratio of zero indicates no change in the expression level between experimental and control samples, a log-ratio greater than zero indicates the increase and a log-ratio less than zero indicates the decrease in the experimental sample. Data are arranged into a matrix with $a_{ij} \in \mathbb{R}$ representing the log-ratio for gene i in sample j. It is quite common that some matrix elements are missing as a result of experimental uncertainties. In our tests, missing log-ratios were set to zero.

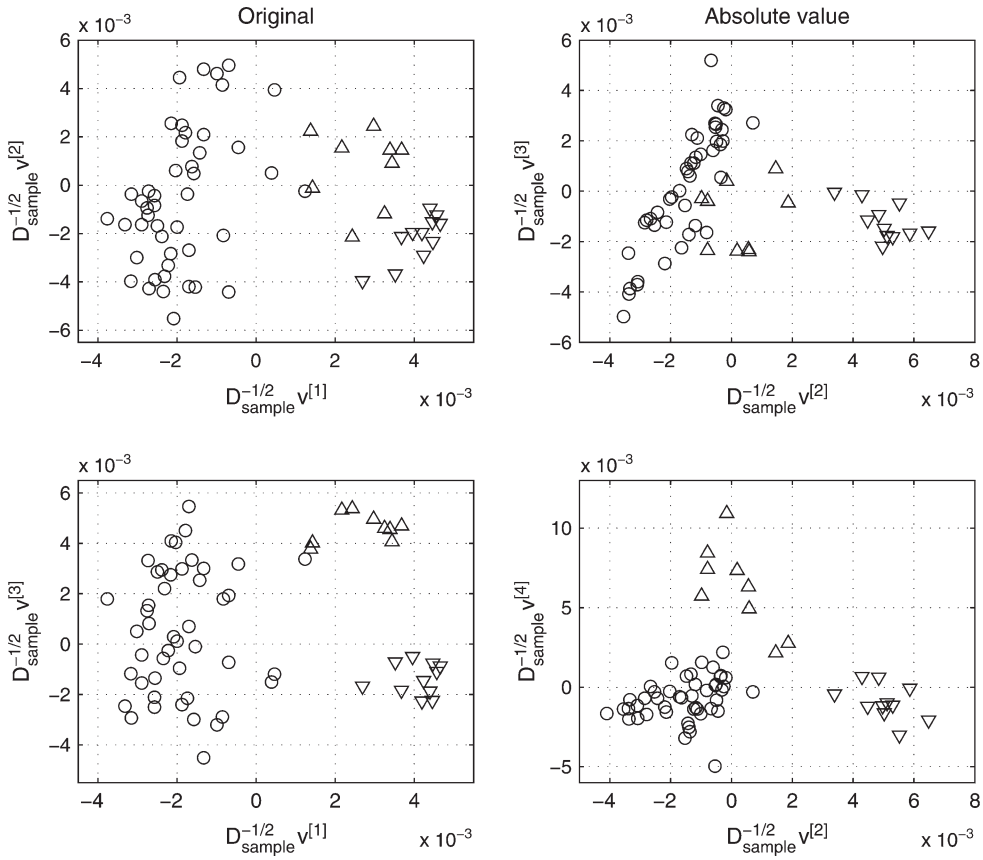

FIG. 5. Lymphoma—scatter plots as in Fig. 4: DLBCL (circles), CLL (triangles down) and FL (triangles up).

In all cases, we use the row/column scaling $D_{\mathrm{L}}^{\mathrm{weight}} = D_{\mathrm{gene}}^{\frac{1}{2}}$ and $D_{\mathrm{R}}^{\mathrm{weight}} = D_{\mathrm{sample}}^{\frac{1}{2}}$, and use the scaled singular vectors $D_{\mathrm{sample}}^{-\frac{1}{2}} v^{[k]}$ to scatter plot the samples.

In addition to using the two-signed data matrix, we also consider the case where the same spectral algorithm is applied to the absolute value matrix $|a_{ij}|$. Using the SVD to cluster nonnegative data is a well-established technique (Dhillon, 2001; Kluger *et al.*, 2003), and in this case the first singular vector, $v^{[1]}$, contains background information that is not relevant to the task. Hence, we compare the information provided by $D_{\mathrm{sample}}^{-\frac{1}{2}} v^{[1]}$, $D_{\mathrm{sample}}^{-\frac{1}{2}} v^{[2]}$ and $D_{\mathrm{sample}}^{-\frac{1}{2}} v^{[3]}$ from the algorithm applied to a_{ij} with that provided by $D_{\mathrm{sample}}^{-\frac{1}{2}} v^{[2]}$, $D_{\mathrm{sample}}^{-\frac{1}{2}} v^{[3]}$ and $D_{\mathrm{sample}}^{-\frac{1}{2}} v^{[4]}$ from the algorithm applied to $|a_{ij}|$.

From a biological perspective,

- using the two-signed data, a_{ij}, we take the view that up-regulated genes are different from down-regulated genes, and we attempt to find groups of samples on which groups of genes are consistently up- or down-regulated, whereas

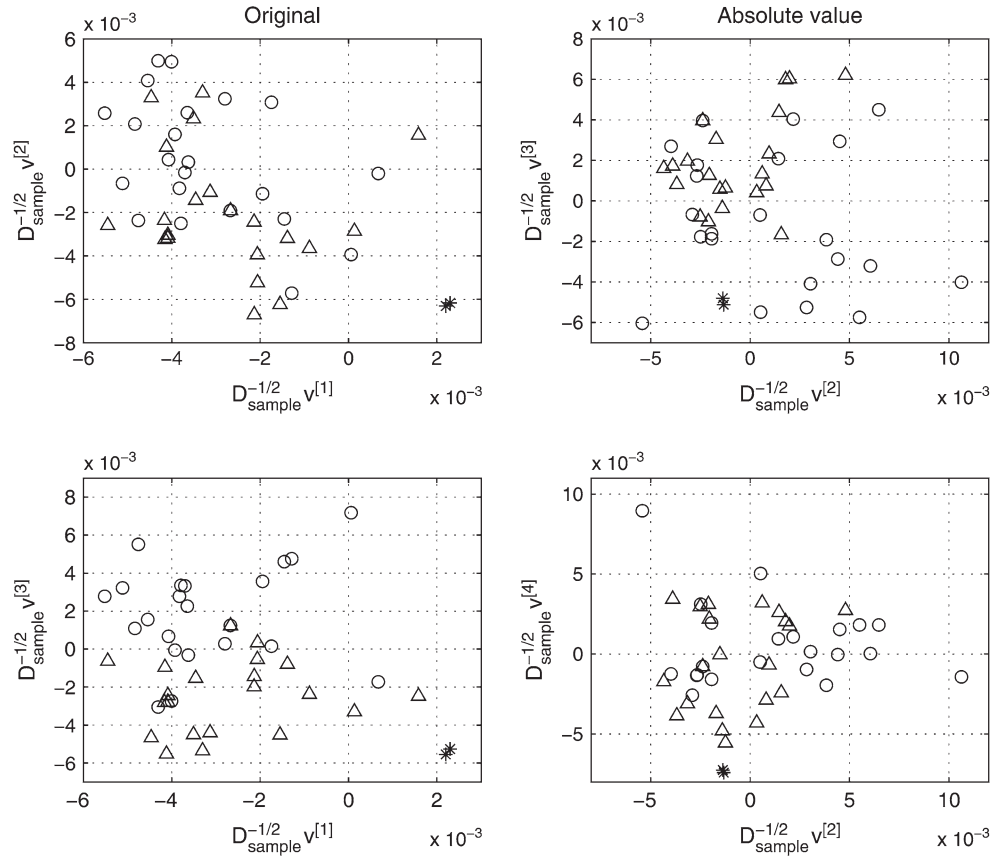

FIG. 6. Lymphoma—scatter plots as in Fig. 4: germinal centre B cells (stars), germinal centre B-like DLBCL (triangles up) and activated B-like DLBCL (circles).

- using $|a_{ij}|$, we take the view that up- or down-regulation is a sign of 'activity' (relative to the normal state of the gene), and we attempt to find groups of samples on which groups of genes are consistently active.

Both viewpoints may lead to useful findings. In this work, we have focussed on deriving and analysing an algorithm for the two-signed case, and we show now that the approach can recover information that is lost when $|a_{ij}|$ is used.

The data set from Alizadeh *et al.* (2000) contains 4026 genes measured across 96 samples: 46 diffuse large B-cell lymphoma (DLBCL), 11 chronic lymphocytic leukaemias (CLLs), nine follicular lymphomas (FLs), 10 activated blood B cells, six transformed cell lines, six T cells, four resting blood B cells, two germinal centre B cells and two normal lymph nodes/tonsils.

We see from Fig. 4 that with the original (two-signed) log-ratio data, the first singular vector distinguishes between DLBCLs and the group of CLLs, FLs and resting blood B cells. In agreement with Alizadeh *et al.* (2000), (a) CLLs and FLs were clustered next to resting B-cell samples, (b) DLBCLs were distinct from CLLs and FLs and (c) some DLBCLs were similar to tonsil. We can also conclude

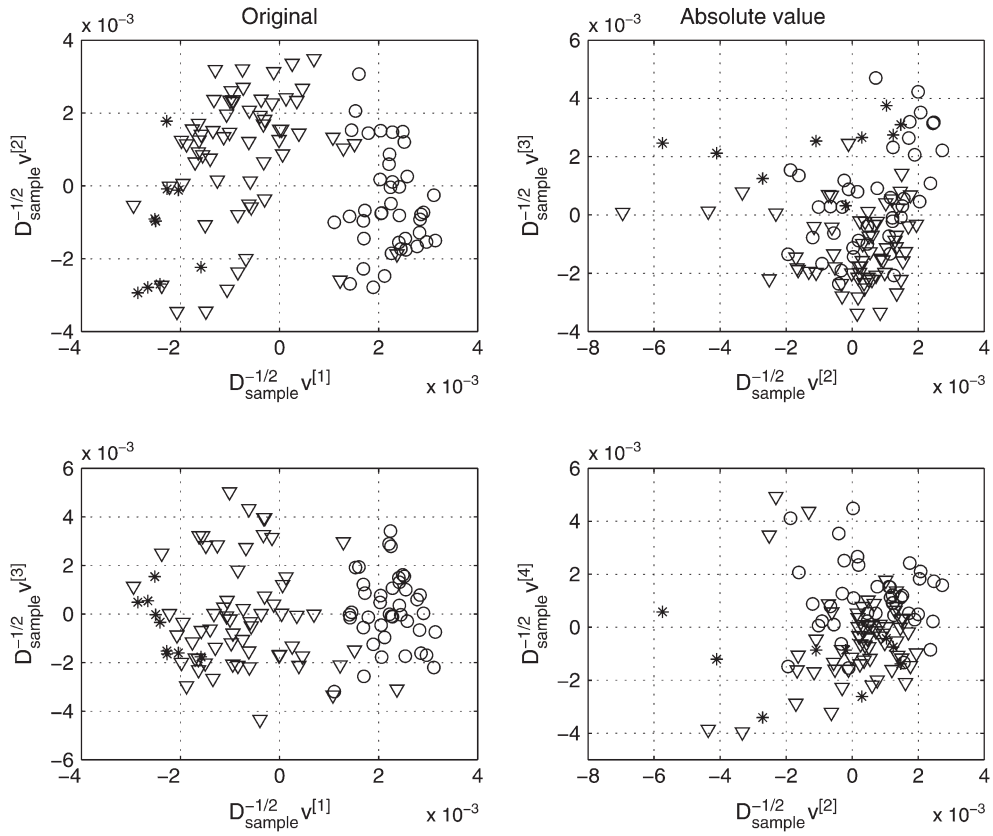

FIG. 7. Prostate cancer—scatter plots as in Fig. 4: normal (circles), tumours (triangles down) and lymph node metastases (stars).

that there was a common expression signature apparent in DLBCLs and transformed cell lines and in B cells and T cells. The second singular vector distinguishes DLBCLs from B and T cells and transformed cell lines. With $|a_{ij}|$, $D_{\text{sample}}^{-\frac{1}{2}} v^{[2]}$ distinguishes between CLLs, FLs and resting blood B cells on the left- (these three groups clustered together as before) and T cells on the right-hand side of the scatter plot.

We now look at the effect of applying the method to a restricted data set, containing fewer sample types. When we restrict the data to samples of the three most prevalent adult lymphoid malignancies, DLBCLs, CLLs and FLs, a clear distinction between the three types can be seen in both the two-signed and the absolute value cases, as shown in Fig. 5. Here, the second and the third dominant singular vectors produced the clearest separation.

In Fig. 6, only the DLBCLs and the germinal centre B cells samples were used in order to rediscover two DLBCL subtypes seen in Alizadeh *et al.* (2000): germinal centre B-like and activated B-like. With the original data, the third singular vector was able to recognize the difference we were looking for, positioning two germinal centre B cells samples into germinal centre B-like subgroup. This example clearly justifies the use of two-signed data: the right-hand side of Fig. 6 shows that the absolute value version did not separate the DLBCL subtypes.

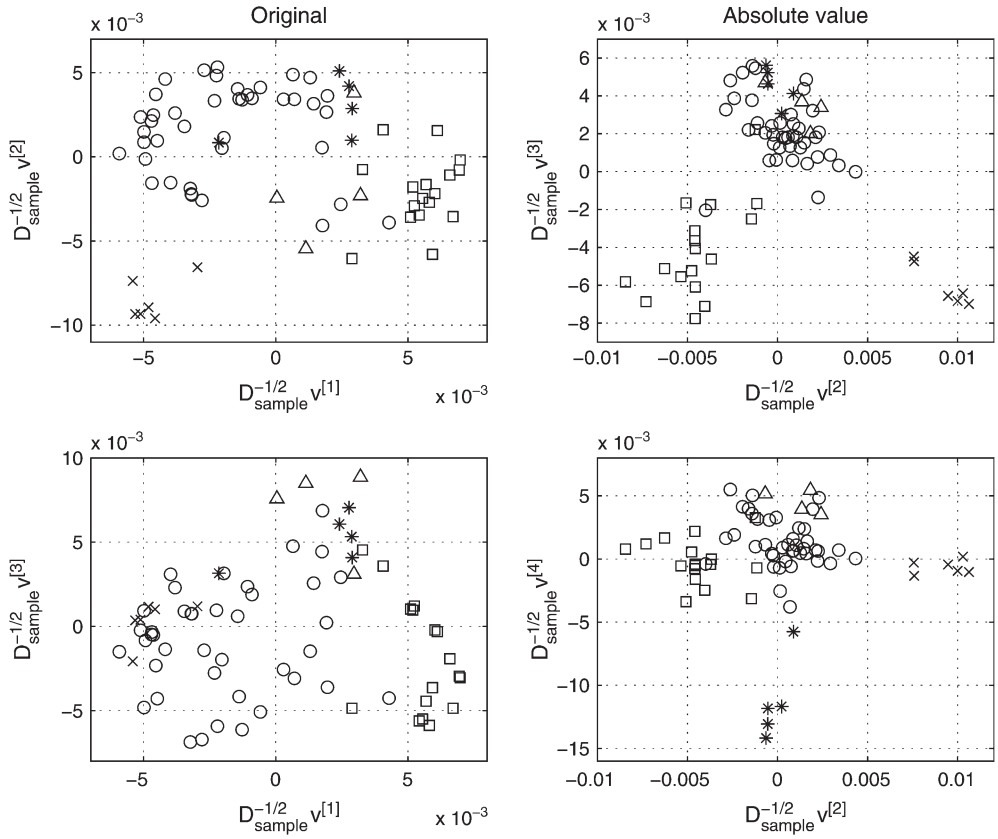

FIG. 8. Lung cancer—scatter plots as in Fig. 4: normal (crosses), adenocarcinomas (circles), squamous (squares), large-cell (triangles up) and small-cell (stars) tumours.

Gene expression in 112 prostate tissues, of which 62 were primary tumours, 41 matched normal prostate tissues and nine unmatched lymph node metastases was studied in Lapointe *et al.* (2004). The 5153 genes whose expression varied most across samples had been selected out of 26 000 genes. In Fig. 7, we see that with the original data, samples were divided into two major clusters by the first singular vector. The left cluster contains tumours, with metastases far left, and the right cluster contains normal samples and five tumours. The absolute value data did not support such a separation.

The gene expression profiles for 67 human lung tumours and six normal tissues were examined in Garber *et al.* (2001) using 918 genes. Lung tumours included adenocarcinomas, squamous, large-cell and small-cell lung tumours. The distinction between small-cell and nonsmall-cell lung cancers seems to be very important due to different medical treatment of patients. Results are shown in Fig. 8. Separation of normal samples and squamous tumours from adenocarcinomas is apparent in both cases. The small-cell lung tumours were recognized in the absolute value case.

6. Summary and conclusions

Our aim in this work was to derive, analyse and test a spectral clustering algorithm for cDNA data containing both positive and negative entries. Using these two-signed data, where positive and negative

values represent up- and down-regulation, respectively, allows a simple biological interpretation of the the reordered data. We gave theoretical support for the approach, based on a biologically justified hypothesis, and showed that the algorithm will exploit 'checkerboard' patterns that are hidden in the data. A tumour classification case study, using three public domain data sets, showed that this approach can find information that is lost when the more traditional 'absolute value' approach is used. Overall, two-signed and absolute value spectral clustering should be viewed as complementary techniques, based on different hypotheses, that therefore summarize the data in different ways.

Acknowledgements

DJH was supported by a Research Fellowship from the Royal Society of Edinburgh/Scottish Executive Education and Lifelong Learning Department and by the Engineering and Physical Sciences Research Council (EPSRC) grant GR/S62383/01. GK was supported by the EPSRC grant GR/S62383/01.

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson Jr., J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., R. L. R, Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. & Staudt, L. M. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Chen, X., Cheung, S. T., So, S., Fan, S. T., Barry, C., Higgins, J., Lai, K. M., Ji, J., Dudoit, S., Ng, I. O., van de Rijn, M., Botstein, D. & Brown, P. O. (2002) Gene expression patterns in human liver cancers. *Mol. Biol. Cell*, **13**, 1929–1939.

Dhillon, I. S. (2001) Co-clustering documents and words using bipartite spectral graph partitioning. *Proceedings of the Seventh ACM SIGKDD Conference*.

Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G. D., Perou, C. M., Whyte, R. I., Altman, R. B., Brown, P. O., Botstein, D. & Petersen, I. (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl. Acad. Sci. USA*, **98**, 13784–13789.

Grindrod, P. (2002) Range-dependent random graphs and their application to modeling large small-world Proteome datasets. *Phys. Rev. E*, **66**, 1–7.

Grindrod, P. & Kibble, M. (2004) Review of uses of network and graph theory concepts within proteomics. *Expert Rev. Proteomics*, **1**, 229–238.

Handl, J., Knowles, J. & Kell, D. B. (2005) Computational cluster validation in post-genomic data analysis. *Bioinformatics*, **21**, 3201–3212.

Higham, D. J. (2003) Unravelling small world networks. *J. Comput. Appl. Math.*, **158**, 61–74.

Horn, R. A. & Johnson, C. R. (1985) *Matrix Analysis*. Cambridge: Cambridge University Press.

Kluger, Y., Basri, R., Chang, J. & Gerstein, M. (2003) Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Res.*, **13**.

Lapointe, J., Li, C., Higgins, J. P., van de Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., Ekman, P., DeMarzo, A. M., Tibshirani, R., Botstein, D., Brown, P. O., Brooks, J. D. & Pollack, J. R. (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl. Acad. Sci. USA*, **101**, 811–816.

Morrison, J. L., Breitling, R., Higham, D. J. & Gilbert, D. R. (2006) A lock-and-key model for protein-protein interactions. *Bioinformatics* (to appear). Advanced access doi:10.1093/bioinformatics/btl338.

PRZULJ, N., CORNEIL, D. G. & JURISICA, I. (2004) Modeling interactome: scale-free or geometric? *Bioinformatics*, **20**, 3508–3515.

THOMAS, A., CANNINGS, R., MONK, N. A. M. & CANNINGS, C. (2003) On the structure of protein-protein interaction networks. *Biochem. Soc. Trans.*, **31**, 1491–1496.

WALL, M. E., RECHTSTEINER, A. & ROCHA, L. M. (2003) Singular value decomposition and principal component analysis. *A Practical Approach to Microarray Data Analysis* (D. P. Berrar, W. Dubitzky & M. Granzow). LANL LA-UR-02-4001. Kluwer, pp. 91–109.

XING, E. P. & KARP, R. P. (2001) Cliff: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics*, **1**, 1–9 (discovery note).

YEUNG, K. Y. & RUZZO, W. L. (2001) Principal component analysis for clustering gene expression data. *Bioinformatics*, **17**, 763–774.