# Multidimensional partitioning and bi-partitioning: analysis and application to gene expression datasets

Gabriela Kalna* J. Keith Vass† Desmond J. Higham‡

November 7, 2006

**Abstract**

Eigenvectors and, more generally, singular vectors, have proved to be useful tools for data mining and dimension reduction. Spectral clustering and reordering algorithms have been designed and implemented in many disciplines, and they can be motivated from several different standpoints. Here we give a general, unified, derivation from an applied linear algebra perspective. We use a variational approach that has the benefit of (a) naturally introducing an appropriate scaling, (b) allowing for a solution in any desired dimension, and (c) dealing with both the clustering and bi-clustering issues in the same framework. The motivation and analysis is then backed up with examples involving two large data sets from modern, high-throughput, experimental cell biology. Here, the objects of interest are genes and tissue samples, and the experimental data represents gene activity. We show that looking beyond the dominant, or Fiedler, direction reveals important information.

**Keywords**: data mining dimension reduction, feature selection, graph Laplacian, Fiedler vector, microarray, singular value decomposition, tumour classification.

*Department of Mathematics, University of Strathclyde, Glasgow G1 1XH, UK. Supported by EPSRC grant GR/S62383/01.

†The Beatson Institute for Cancer Research, Glasgow G61 1BD, UK

‡Department of Mathematics, University of Strathclyde, Glasgow G1 1XH, UK. Supported by EPSRC grant GR/S62383/01.

**AMS Subject Classification:** 65F15, 92C37.

# 1 Background

Modern technology is responsible for a data deluge that has driven the need for computational algorithms in data mining and dimension reduction. Many large scale data sets take the form of a matrix, $W$, with $w_{ij}$ representing some relationship between objects labeled $i$ and $j$. If objects $i$ and $j$ come from the same list, then $W$ will be square. For example, $w_{ij}$ may be a correlation coefficient between stock prices [3]. If objects $i$ and $j$ come from different lists, then $W$ can be rectangular. For example, $w_{ij}$ may represent the number of occurrences of word $i$ in document $j$, [7].

Given its fundamental role in applied matrix analysis, it is not surprising that the singular value decomposition is an extremely useful tool for summarizing important information from such data sets. We refer to [11] for a list of areas where the general concept of spectral clustering has been applied.

Data mining with singular vectors can be motivated from a number of different directions and is closely related to ideas in Principle Component Analysis [20, 21], support vector machines/kernel based methods [16], machine learning [14] and multidimensional scaling [6]. Our main contribution here is to present a simple, unified framework that justifies the approach while automatically

**(a)** introducing an appropriate scaling,

**(b)** allowing for a solution in any desired dimension, and

**(c)** dealing with both the clustering and bi-clustering issues.

In particular, this work extends that in [11] to allow for bi-clustering of non-square data and for projecting to arbitrary dimension.

To illustrate the analysis, and in particular to emphasize that more than just the first, or Fiedler, direction can be important, we also present numerical results on microarray expression data sets.

Throughout this work we use the following notation:

- $\|\cdot\|_2$ denotes the Euclidean norm,
- $a^{[j]}$ denotes the $j$th column of the matrix $A$,

- $\mathbf{1}$ denotes the vector in $\mathbb{R}^N$ with all elements equal to one.
- $I$ denotes an identity matrix whose dimension is clear from the context,
- $0_{r\times s}$ denotes the zero matrix in $\mathbb{R}^{r\times s}$.

# 2 Square Symmetric Case

In this section, we consider the case where $W = W^T \in \mathbb{R}^{N\times N}$ is a square, symmetric matrix with non-negative elements, $w_{ij} \geq 0$, and with all $w_{ii} = 0$. Here, there are $N$ objects of interest and $w_{ij} = w_{ji}$ represents the pairwise similarity of objects $i$ and $j$. We take the view that a large value of $w_{ij}$ means that objects $i$ and $j$ are very similar. (Some references use the opposite convention, taking a large $w_{ij}$ to mean very dissimilar, but, of course, a simple transformation such as $w_{ij} \mapsto \max_{r,s} w_{rs} - w_{ij}$ converts that format into ours.)

When $N$ is large, the pairwise similarity data in $W$ represents a vast amount of information. In order to create a manageable subset of information that can be easily visualized or otherwise processed, it is necessary to summarize the data. Three typical, and closely related, tasks are

**1.** re-order the objects so that objects close together have strong similarity and objects far apart have weak similarity [2, 10],

**2.** map each object to a point in a low dimensional space, $\mathbb{R}^s$, so that objects close in Euclidean distance have strong similarity and objects far apart in Euclidean distance have weak similarity [4],

**3.** split the objects in to two or more clusters so that objects in the same cluster have strong similarity and objects in different clusters have weak similarity [9].

In this work we focus on task 2, while noting that a method for task 1 then follows automatically—map into $\mathbb{R}^1$ and use the resulting $N$ numbers to order the objects. Similarly, having achieved 2, there are straightforward ways to produce a clustering for task 3 [1, 17].

Now, focusing on task 2, for some $s < N$ our aim is to find vectors $\{y^{[1]}, y^{[2]}, \ldots, y^{[N]}\}$ with each $y^{[j]} \in \mathbb{R}^s$ such that the $j$th object is associated with the vector $y^{[j]}$. The idea is that the relative distance $\| y^{[i]} - y^{[j]} \|_2$ reflects

the pairwise similarity weight, $w_{ij}$. We began with $(N^2 - N)/2$ real numbers (that is, the elements of $W$, allowing for symmetry and a zero diagonal) and we hope to reduce this to $Ns$ numbers (in the vectors $\{y^{[j]}\}_{j=1}^N$). Clearly, if $N$ is large and $s \ll N$ then this is a significant compression.

Given that $\| y^{[i]} - y^{[j]} \|_2$ should be small when $w_{ij}$ is large and vice versa, a reasonable starting point is to consider choosing $\{y^{[j]}\}_{j=1}^N$ to minimize $\sum_i \sum_j \| y^{[i]} - y^{[j]} \|_2^2 w_{ij}$. However, since this objective function can generally be decreased simply by rescaling $y^{[k]} \mapsto \epsilon y^{[k]}$, we must incorporate some normalizing constraint. Considering that the $k$th object gets mapped to a vector whose first component is $y_1^{[k]}$, we will normalize the two-norm of the vector making up these components, when scaled by the square root of the corresponding degree; that is, set

$$\left\| \begin{bmatrix} \sqrt{d_1}\, y_1^{[1]} \\ \sqrt{d_2}\, y_1^{[2]} \\ \vdots \\ \vdots \\ \sqrt{d_N}\, y_1^{[N]} \end{bmatrix} \right\|_2 = 1.$$

Here $d_k := \sum_{r=1}^N w_{kr}$ is the degree of object $k$, that is, the total weight associated with node $k$ in the corresponding graph. Scaling by $\sqrt{d_k}$ tends to penalize the 'promiscuous' nodes, forcing them near the origin, and hence away from particular clusters, and stopping them from dominating in the optimization problem. Another concern is to avoid having all $y_1^{[k]}$ equal, so that all objects are given the same first component. This could be dealt with by a constraint such as $\sum_{k=1}^N y_1^{[k]} = 0$. However, we find it more convenient to return to this issue at a later stage; more precisely, when we move from (6) to (7).

Now, when we consider the second component, the same normalization argument leads to

$$\left\| \begin{bmatrix} \sqrt{d_1}\, y_2^{[1]} \\ \sqrt{d_2}\, y_2^{[2]} \\ \vdots \\ \vdots, \\ \sqrt{d_N}\, y_2^{[N]} \end{bmatrix} \right\|_2 = 1.$$

Also, we don't want this vector to 'overlap' with the previous vector; that is,

we want this component to contain only new information that is not already contained in the first component. This means that we need an orthogonality condition

$$\left[\sqrt{d_1}\, y_1^{[1]}, \sqrt{d_2}\, y_1^{[2]}, \cdots\cdots \sqrt{d_N}\, y_1^{[N]}\right] \begin{bmatrix} \sqrt{d_1}\, y_2^{[1]} \\ \sqrt{d_2}\, y_2^{[2]} \\ \vdots \\ \vdots \\ \sqrt{d_N}\, y_2^{[N]} \end{bmatrix} = 0.$$

Continuing these arguments for all components leads to the constraint $YDY^T = I$. Hence our optimization problem to define a suitable choice of $\{y^{[j]}\}_{j=1}^N$ is

$$\min_{y^{[i]}\in\mathbb{R}^s,\quad YDY^T=I} \sum_{i=1}^{N}\sum_{j=1}^{N} \|y^{[i]} - y^{[j]}\|_2^2 w_{ij}. \tag{1}$$

Note that there is a natural redundancy in this problem. Any solution $Y$ of (1) can be changed to $QY$, where $Q \in \mathbb{R}^{s\times s}$ is orthogonal. Such a transformation doesn't change the relative distances, $\|Qy^{[i]} - Qy^{[j]}\|_2^2 = \|Q(y^{[i]} - y^{[j]})\|_2^2 = \|y^{[i]} - y^{[j]}\|_2^2$, and doesn't affect the constraint, $(QY)D(QY)^T = QYDY^TQ^T = QIQ^T = I$.

## 2.1 Rewrite and Solve

In this subsection, we show that (1) is tractable, having a computationally convenient solution.

First, we note that

$$\sum_{i=1}^{s}(YDY^T)_{ii} = \sum_{i=1}^{s}\sum_{k=1}^{N} {y_i^{[k]}}^2 d_k = \sum_{k=1}^{N} \|{y^{[k]}}^2\|_2^2\, d_k.$$

So the constraint $YDY^T = I$ implies

$$\sum_{k=1}^{N} \|{y^{[k]}}^2\|_2^2 d_k = s. \tag{2}$$

Now, since

$$\|y^{[i]} - y^{[j]}\|_2^2 = \left(y^{[i]} - y^{[j]}\right)^T \left(y^{[i]} - y^{[j]}\right) = \|y^{[i]}\|_2^2 + \|y^{[j]}\|_2^2 - 2{y^{[i]}}^T y^{[j]},$$

we have

$$\begin{aligned}\sum_{i,j=1}^{N} \|y^{[i]} - y^{[j]}\|_2^2 w_{ij} &= \sum_{i=1}^{N} \|y^{[i]}\|_2^2 \sum_{j=1}^{N} w_{ij} + \sum_{j=1}^{N} \|y^{[j]}\|_2^2 \sum_{i=1}^{N} w_{ij} - 2\sum_{i=1}^{N}\sum_{j=1}^{N} {y^{[i]}}^T y^{[j]} w_{ij} \\ &= 2\sum_{i=1}^{N} \|y^{[i]}\|_2^2 d_i - 2\sum_{i=1}^{N}\sum_{j=1}^{N} {y^{[i]}}^T y^{[j]} w_{ij}.\end{aligned}$$

From (2), the first term on the right-hand side is constant and so the problem (1) is equivalent to

$$\max_{Y\in\mathbb{R}^{s\times N},\ YDY^T=I} \sum_{i,j=1}^{N} {y^{[i]}}^T y^{[j]} w_{ij},$$

which may be rewritten

$$\max_{Y\in\mathbb{R}^{s\times N},\ YDY^T=I} \operatorname{trace}\left(YWY^T\right).$$

Setting $X = YD^{\frac{1}{2}} \in \mathbb{R}^{s\times N}$, this problem becomes

$$\max_{X\in\mathbb{R}^{s\times N},\ XX^T=I} \operatorname{trace}\left(XD^{-\frac{1}{2}}WD^{-\frac{1}{2}}X^T\right). \tag{3}$$

Now, suppose $D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$ has the eigen-decomposition

$$D^{-\frac{1}{2}}WD^{-\frac{1}{2}} = U\Gamma U^T,$$

where $U \in \mathbb{R}^{N\times N}$ is orthogonal and $\Gamma \in \mathbb{R}^{N\times N}$ is diagonal with diagonal elements given by the eigenvalues, ordered $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_N$. Letting $Z := XU \in \mathbb{R}^{s\times N}$, the constraint $XX^T = I$ becomes $ZU^TUZ^T = I$, that is, $ZZ^T = I$, and $XD^{-\frac{1}{2}}WD^{-\frac{1}{2}}X^T = ZU^TD^{-\frac{1}{2}}WD^{-\frac{1}{2}}UZ^T = Z\Gamma Z^T$. Hence the problem (3) becomes

$$\max_{Z\in\mathbb{R}^{s\times N},\ XX^T=I} \operatorname{trace}\left(Z\Gamma Z^T\right),$$

which is equivalent to

$$\max_{Z\in\mathbb{R}^{s\times N},\ ZZ^T=I} \sum_{k=1}^{N} \gamma_k \|z^{[k]}\|_2^2, \tag{4}$$

where we recall our notation that $z^{[k]}$ denotes the $k$th column of $Z$.

Now the constraint $ZZ^T = I$ forces $Z \in \mathbb{R}^{s\times N}$ to have orthonormal rows and hence given any feasible $Z$ we may append rows to create an orthogonal matrix

$$\begin{bmatrix} Z \\ \widehat{Z} \end{bmatrix} \in \mathbb{R}^{N\times N}.$$

It follows that $Z$ must have columns of two-norm bounded above by one. Hence, (4) clearly has a set of solutions given by

$$Z = \begin{bmatrix} L & \vdots & 0_{s\times(N-s)} \end{bmatrix}, \tag{5}$$

where $L \in \mathbb{R}^{s\times s}$ is orthogonal.

Using $Y = XD^{-\frac{1}{2}} = ZU^TD^{-\frac{1}{2}}$, this tells us that

$$Y = [L\vdots 0_{s\times(N-s)}] \begin{bmatrix} u^{[1]^T} & \dots & \dots & \dots \\ u^{[2]^T} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ u^{[N]^T} & \dots & \dots & \dots \end{bmatrix} D^{-\frac{1}{2}} = L \begin{bmatrix} (D^{-\frac{1}{2}}u^{[1]})^T & \dots & \dots & \dots \\ (D^{-\frac{1}{2}}u^{[2]})^T & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ (D^{-\frac{1}{2}}u^{[N]})^T & \dots & \dots & \dots \end{bmatrix}.$$

The arbitrary orthogonal factor $L$ is no surprise; it is consistent with the natural redundancy in the problem that we discussed earlier. Without loss of generality, we can take $L = I$, to obtain

$$Y = \begin{bmatrix} (D^{-\frac{1}{2}}u^{[1]})^T & \dots & \dots & \dots \\ (D^{-\frac{1}{2}}u^{[2]})^T & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ (D^{-\frac{1}{2}}u^{[s]})^T & \dots & \dots & \dots \end{bmatrix}. \tag{6}$$

This result shows that the problem (1) is solved by taking the eigenvectors corresponding to the $s$ most positive eigenvalues of the scaled matrix $D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$, and then scaling these on the left by $D^{-\frac{1}{2}}$. The final step of the analysis is to notice that, by construction, $D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$ has an eigenvector $D^{\frac{1}{2}}\mathbf{1}$, corresponding to the eigenvalue 1. Moreover, it is known that all eigenvalues of $D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$ lie in the range $[-1, 1]$, with 1 being a simple eigenvalue if the graph corresponding to $W$ is connected, [8, 19]. It follows

that we may assume that the first row of $Y$ in (6) is $\mathbf{1}^T$, and following the earlier argument about appropriate constraints, we then ignore this row and take $Y$ to be

$$Y = \begin{bmatrix} (D^{-\frac{1}{2}}u^{[2]})^T & \dots & \dots & \dots \\ (D^{-\frac{1}{2}}u^{[3]})^T & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ (D^{-\frac{1}{2}}u^{[s+1]})^T & \dots & \dots & \dots \end{bmatrix}. \tag{7}$$

**Remarks**

1. Our derivation worked directly with the normalized weight matrix, $D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$. An alternative is to use the *normalized graph Laplacian*, $D^{-\frac{1}{2}}(D-W)D^{-\frac{1}{2}}$, which, of course, has the same eigenvectors with appropriately shifted eigenvalues [11].

2. In this work we are assuming that all weights are non-negative, whence the dominant eigenvector, $D^{-\frac{1}{2}}u^{[1]}$, gives no useful information. However, we point out that this type of spectral analysis carries through to the case where $W$ has both positive and negative entries, and here the dominant eigenvector can reveal important patterns in the data [12].

# 3 Rectangular Case

## 3.1 Data and Problem

We now consider the case where $W \in \mathbb{R}^{M\times N}$, with $M$ different to $N$, in general. As with section 2, we suppose that $w_{ij} \geq 0$ represents similarity between objects, but now we think of two separate lists of objects, so that $w_{ij}$ relates object $i$ from the first list to object $j$ from the second list. In section 4 we deal with the case where $w_{ij}$ represents the expression level of gene $i$ in tissue sample $j$. Following the approach in 2, our aim is to find vectors $\{p^{[i]}\}_{i=1}^M$ and $\{q^{[j]}\}_{j=1}^N$ with each $p^{[i]}$ and $q^{[j]}$ in $\mathbb{R}^s$ and $s < \min(M,N)$, such that the $i$th object in the first list is associated with $p^{[i]}$ and the $j$th object in the second list is associated with $q^{[j]}$. Then the arguments that led to (1) can be used to arrive at

$$\min_{p^{[i]},q^{i]}\in\mathbb{R}^s,\ PD_{\text{out}}P^T=QD_{\text{in}}Q^T=I} \sum_{i=1}^{M}\sum_{j=1}^{N} \|p^{[i]}-q^{[j]}\|_2^2 w_{ij}, \tag{8}$$

where $D_{\text{out}} \in \mathbb{R}^{M\times M}$ is the diagonal out-degree matrix, so that $(D_{\text{out}})_{ii} = \sum_{j=1}^{N} w_{ij} =: (d_{\text{out}})_i$, and $D_{\text{in}} \in \mathbb{R}^{N\times N}$ is the diagonal in-degree matrix, so that $(D_{\text{in}})_{jj} = \sum_{i=1}^{M} w_{ij} =: (d_{\text{in}})_j$.

## 3.2 Rewrite and Solve

To solve (8), we first note that

$$\sum_{i=1}^{M}\sum_{j=1}^{N} \|p^{[i]}-q^{[j]}\|_2^2 w_{ij} = \sum_{i=1}^{M} \|p^{[i]}\|_2^2 (d_{\text{out}})_i + \sum_{j=1}^{N} \|q^{[j]}\|_2^2 (d_{\text{in}})_j - 2\sum_{i=1}^{M}\sum_{j=1}^{N} {p^{[i]}}^T q^{[j]} w_{ij}.$$

Applying the analogues of (2), we see that the first two terms are constant, and so the problem (8) is equivalent to

$$\max_{P\in\mathbb{R}^{s\times M},\ Q\in\mathbb{R}^{s\times N},\ PD_{\text{out}}P^T = QD_{\text{in}}Q^T = I} \sum_{i=1}^{M}\sum_{j=1}^{N} {p^{[i]}}^T q^{[j]} w_{ij},$$

which may be rewritten

$$\max_{P\in\mathbb{R}^{s\times M},\ Q\in\mathbb{R}^{s\times N},\ PD_{\text{out}}P^T = QD_{\text{in}}Q^T = I} \operatorname{trace}\left(PWQ^T\right). \tag{9}$$

Setting $A = PD_{\text{out}}^{\frac{1}{2}} \in \mathbb{R}^{s\times M}$ and $B = QD_{\text{in}}^{\frac{1}{2}} \in \mathbb{R}^{s\times N}$, the problem (9) becomes

$$\max_{A\in\mathbb{R}^{s\times M},\ B\in\mathbb{R}^{s\times N},\ AA^T = BB^T = I} \operatorname{trace}\left(AD_{\text{out}}^{-\frac{1}{2}} W D_{\text{in}}^{-\frac{1}{2}} B^T\right). \tag{10}$$

Now, suppose $D_{\text{out}}^{-\frac{1}{2}} W D_{\text{in}}^{-\frac{1}{2}}$ has the singular value decomposition (SVD)

$$D_{\text{out}}^{-\frac{1}{2}} W D_{\text{in}}^{-\frac{1}{2}} = U\Sigma V^T,$$

where $U \in \mathbb{R}^{M\times M}$ and $V \in \mathbb{R}^{N\times N}$ are orthogonal and $\Sigma \in \mathbb{R}^{M\times N}$ is diagonal with diagonal elements $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$. Letting $R := AU \in \mathbb{R}^{s\times M}$ and $S := BV \in \mathbb{R}^{s\times N}$, the constraint $AA^T = I$ becomes $RU^TUR^T = I$, that is, $RR^T = I$, and the constraint $BB^T = I$ becomes $SS^T = I$. Also,

$AD_{\text{out}}^{-\frac{1}{2}} W D_{\text{in}}^{-\frac{1}{2}} B^T = RU^T D_{\text{out}}^{-\frac{1}{2}} W D^{-\frac{1}{2}} V S^T = R\Sigma S^T$. Hence the problem (10) becomes

$$\max_{R\in\mathbb{R}^{s\times N},\ S\in\mathbb{R}^{s\times N},\ RR^T=SS^T=I} \operatorname{trace}\left(R\Sigma S^T\right),$$

which is equivalent to

$$\max_{R\in\mathbb{R}^{s\times M},\ S\in\mathbb{R}^{s\times N},\ RR^T=SS^T=I} \sum_{k=1}^{\min(M,N)} \sigma_k r^{[k]^T} s^{[k]}. \tag{11}$$

Now, repeating the arguments used to obtain (5), and also invoking the Cauchy-Schwarz inequality, we find that (11) has a set of solutions given by

$$R = S = \left[\begin{array}{c:c} L & 0_{s\times(M-s)} \end{array}\right],$$

where $L \in \mathbb{R}^{s\times s}$ is orthogonal.

Using $P = AD_{\text{out}}^{-\frac{1}{2}} = RU^T D_{\text{out}}^{-\frac{1}{2}}$, this tells us that

$$P = \left[\begin{array}{c:c} L & 0_{s\times(M-s)} \end{array}\right] \begin{bmatrix} u^{[1]^T} & \dots & \dots & \dots \\ u^{[2]^T} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ u^{[N]^T} & \dots & \dots & \dots \end{bmatrix} D_{\text{out}}^{-\frac{1}{2}} = L \begin{bmatrix} (D_{\text{out}}^{-\frac{1}{2}} u^{[1]})^T & \dots & \dots & \dots \\ (D_{\text{out}}^{-\frac{1}{2}} u^{[2]})^T & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ (D_{\text{out}}^{-\frac{1}{2}} u^{[N]})^T & \dots & \dots & \dots \end{bmatrix}.$$

Similarly, for the same $L$,

$$Q = L \begin{bmatrix} (D_{\text{in}}^{-\frac{1}{2}} v^{[1]})^T & \dots & \dots & \dots \\ (D_{\text{in}}^{-\frac{1}{2}} v^{[2]})^T & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ (D_{\text{in}}^{-\frac{1}{2}} v^{[N]})^T & \dots & \dots & \dots \end{bmatrix}.$$

Now, as argued in section 2, the orthogonal factor $L$ is arbitrary, and we may take $L = I$, which gives

$$P = \begin{bmatrix} (D_{\text{out}}^{-\frac{1}{2}} u^{[1]})^T & \dots & \dots & \dots \\ (D_{\text{out}}^{-\frac{1}{2}} u^{[2]})^T & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ (D_{\text{out}}^{-\frac{1}{2}} u^{[s]})^T & \dots & \dots & \dots \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} (D_{\text{in}}^{-\frac{1}{2}} v^{[1]})^T & \dots & \dots & \dots \\ (D_{\text{in}}^{-\frac{1}{2}} v^{[2]})^T & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ (D_{\text{in}}^{-\frac{1}{2}} v^{[s]})^T & \dots & \dots & \dots \end{bmatrix}. \tag{12}$$

The result (12) shows that (8) is solved by taking the left and right singular vectors corresponding to the $s$ dominant singular values of $D_{\text{out}}^{-\frac{1}{2}} W D_{\text{in}}^{-\frac{1}{2}}$ and then scaling these on the left by $D_{\text{out}}^{-\frac{1}{2}}$ and $D_{\text{in}}^{-\frac{1}{2}}$, respectively. Just as in section 2.1, the final piece in the analysis is to notice that $D_{\text{out}}^{-\frac{1}{2}} W D_{\text{in}}^{-\frac{1}{2}}$ has a dominant singular value of $\sigma_1 = 1$ and the corresponding first rows, $(D_{\text{out}}^{-\frac{1}{2}} u^{[1]})^T$ and $(D_{\text{in}}^{-\frac{1}{2}} v^{[1]})^T$, of $P$ and $Q$ in (12), have all entries equal to one. Hence, we replace (12) by

$$P = \begin{bmatrix} (D_{\text{out}}^{-\frac{1}{2}} u^{[2]})^T & \dots & \dots & \dots \\ (D_{\text{out}}^{-\frac{1}{2}} u^{[3]})^T & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ (D_{\text{out}}^{-\frac{1}{2}} u^{[s+1]})^T & \dots & \dots & \dots \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} (D_{\text{in}}^{-\frac{1}{2}} v^{[2]})^T & \dots & \dots & \dots \\ (D_{\text{in}}^{-\frac{1}{2}} v^{[3]})^T & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ (D_{\text{in}}^{-\frac{1}{2}} v^{[s+1]})^T & \dots & \dots & \dots \end{bmatrix}. \tag{13}$$

**Remarks**

1. We note that in the non-square matrix, or bi-partite graph, setting of this section there is no commonly used concept of a graph Laplacian.

2. Given a matrix $W \in \mathbb{R}^{M \times N}$, its singular vectors are equivalent to eigenvectors of $W^T W$ and $W W^T$. The matrix $W^T W$ is essentially measuring correlations between the $i$th and $j$th objects in the first list. Similarly, the matrix $W W^T$ is essentially measuring correlations between the $i$th and $j$th objects in the second list. From this viewpoint, the non-square spectral method could be regarded as

   **(a)** converting to a new, square set of data, by correlating over the objects that are not of interest and then

   **(b)** applying the spectral method for square data that was derived in section 2.

   However, this high-level summary would not lead to the same normalization in general, and for this reason we believe that our approach of deriving the solution from first principles is more satisfactory and illuminating.

# 4 Gene Expression Data

We now give some evidence that a combination of more than one singular vector can be required to reveal important information from real data. We also refer to [13] for further examples from a cutting edge clinical investigation.

Here, we have used two Affymetrix microarray data sets from [5]: a colon cancer data set [15] and a prostate cancer data set [18]. In both cases, a tumour sample is always paired with a normal sample from the same patient. Each data set can be regarded as an array $W \in \mathbb{R}^{M\times N}$, where $w_{ij}$ records the activity of the $i$th gene in the $j$th sample. For the colon cancer data set $M = 3697$ and $N = 44$ and for the prostate cancer data set $M = 6593$ and $N = 94$.

This data falls into the rectangular setting of section 3, and we are interested in the unsupervised tumour classification problem—can we indentify the group of tumour samples and the group of normal samples? For this purpose we will use the matrix $Q$ in (13).

Figures 1 and 3 show $N - 1$ singular values (the first one, $\sigma_1 = 1$, is omitted) and scatter plots of pairs of the three dominant singular vectors. In Figure 1, for the colon cancer data, we can see a clear separation of the tumour (stars) and normal (circles) samples by the second, dominant, singular vector. However, the third and fourth singular vectors pick out further subgroups. These three singular vectors correspond to the triplet of values separated from the remaining singular values (top left subfigure). The distinct subclusters may reflect different origins of the samples (laboratory, experiment) or specific features of the patients. Unfortunately, such extra details are not available for these data sets, so this issue cannot be investigated. Figure 2 gives a 3D picture based on the three leading singular values.

Figure 3 shows an example where tumour and normal samples can be distinguished only by combining singular vectors $D_{\text{in}}^{-1/2}v^{[2]}$ and $D_{\text{in}}^{-1/2}v^{[3]}$— neither singular vector alone gives a perfect separation. A 3D plot of this prostate data set in Figure 4 emphasizes the nonlinear shape of the two clusters.

# References

[1] C. J. Alpert and S.-Z. Yao, *Spectral partitioning: The more eigenvectors, the better*, Proceedings of the 32nd Conference on Design Au-

tomation, (1995), pp. 195–200.

[2] S. T. Barnard, A. Pothen, and H. D. Simon, *A spectral algorithm for envelope reduction of sparse matrices*, Numerical Linear Algebra with Applications, 2 (1995), pp. 317–334.

[3] V. Boginski, S. Butenko, and P. M. Pardalos, *On structural properties of the market graph*, in Innovations in Financial and Economic Networks, A. Nagurney, ed., Edward Elgar Publishers, 2003, pp. 29–45.

[4] J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, *Metagenes and molecular pattern discovery using matrix factorization*, Proc. Nat. Ac. Sci., 101 (2004), pp. 4164–4169.

[5] J. Choi, U. Yu, O. Yoo, and S. Kim, *Differential coexpression analysis using microarray data and its application to human cancer*, Bioinformatics, 21 (2005), pp. 4348–4355.

[6] T. F. Cox and M. A. A. Cox, *Multidimensional Scaling*, Chapman and Hall, London, 1994.

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

[8] C. Ding, X. He, and H. Zha, *A spectral method to separate disconnected and nearly-disconnected web graph components*, in The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2001, pp. 275–280.

[9] M. B. Eisen, P. T. Spellman, P.O.Brown, and D.Botstein, *Cluster analysis and display of genome-wide expression patterns*, Genetics, 95 (1998), pp. 14863–14868.

[10] P. Grindrod, *Range-dependent random graphs and their application to modeling large small-world proteome datasets*, Physical Review E, 66 (2002), pp. 066702–1 to 7.

[11] D. J. Higham, G. Kalna, and M. Kibble, *Spectral clustering and its use in bioinformatics*, J. Comp. Appl. Maths, to appear.

[12] D. J. Higham, G. Kalna, and J. K. Vass, *Spectral analysis of two-signed microarray expression data*, Mathematical Medicine and Biology, to appear.

[13] K. D. Hunter, J. K. Thurlow, J. Fleming, P. J. H. Drake, J. K. Vass, G. Kalna, D. J. Higham, P. Herzyk, D. G. MacDonald, E. K. Parkinson, and P. R. Harrison, *Divergent routes to oral cancer*, Cancer Research, to appear (2006).

[14] D. J. C. MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.

[15] D. Notterman, U. Alon, A. Sierk, and A. Levine, *Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays*, Cancer Res., 61 (2001), pp. 3124–3130.

[16] R. Rifkin, S. Mukherjee, P. Tamayo, S. Ramaswamy, C.-H. Yeang, M. Angelo, M. Reich, T. Poggio, E. S. Lander, T. R. Golub, and J. P. Mesirov, *An analytical method for multiclass molecular cancer classification*, SIAM Review, 45 (2003), pp. 706–723.

[17] J. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (2000), pp. 888–905.

[18] D. Singh, P. Febbo, K. Ross, D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. Renshaw, A. D'Amico, J. Richie, E. Lander, M. Loda, P. Kantoff, T. Golub, and W. Sellers, *Gene expression correlates of clinical prostate cancer behavior*, Cancer Cell, 1 (2002), pp. 203–209.

[19] R. Van Driessche and D. Roose, *An improved spectral bisection algorithm and its application to dynamic load balancing*, Parallel Computing, 21 (1995), pp. 29–48.

[20] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, *Singular Value Decomposition and Principal Component Analysis*, in A Practical Approach to Microarray Data Analysis (Eds D.P. Berrar, W. Dubitzky, M. Granzow), Kluwer, LANL LA-UR-02-4001 (2003), pp. 91–109.

[21] K. Y. Yeung and W. L. Ruzzo, *Principal component analysis for clustering gene expression data*, Bioinformatics, 17 (2001), pp. 763–774.

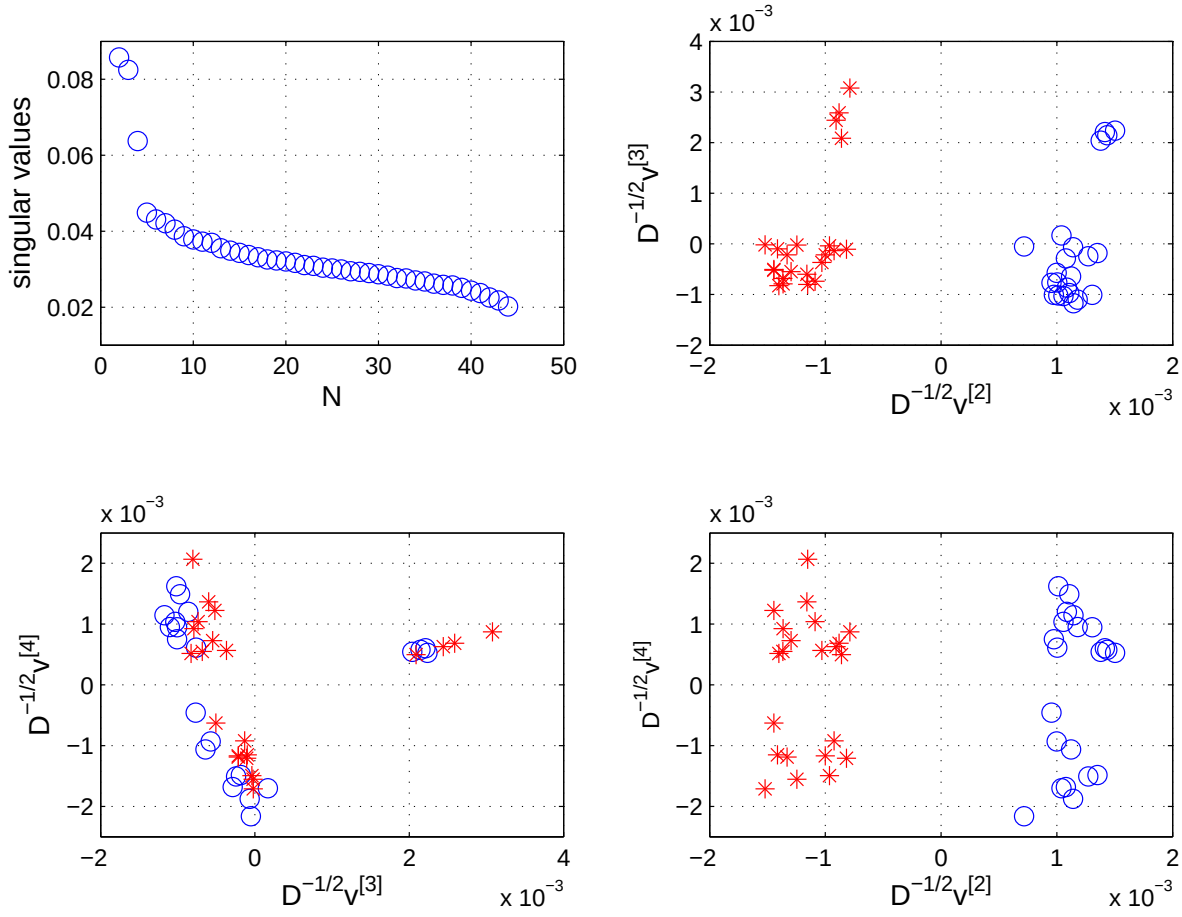


Figure 1: Colon: tumour (stars) and normal (circles) samples. Singular values $\sigma_2, \sigma_3, \ldots, \sigma_N$ (top left) and scatter plots of three pairs of dominant singular vectors.

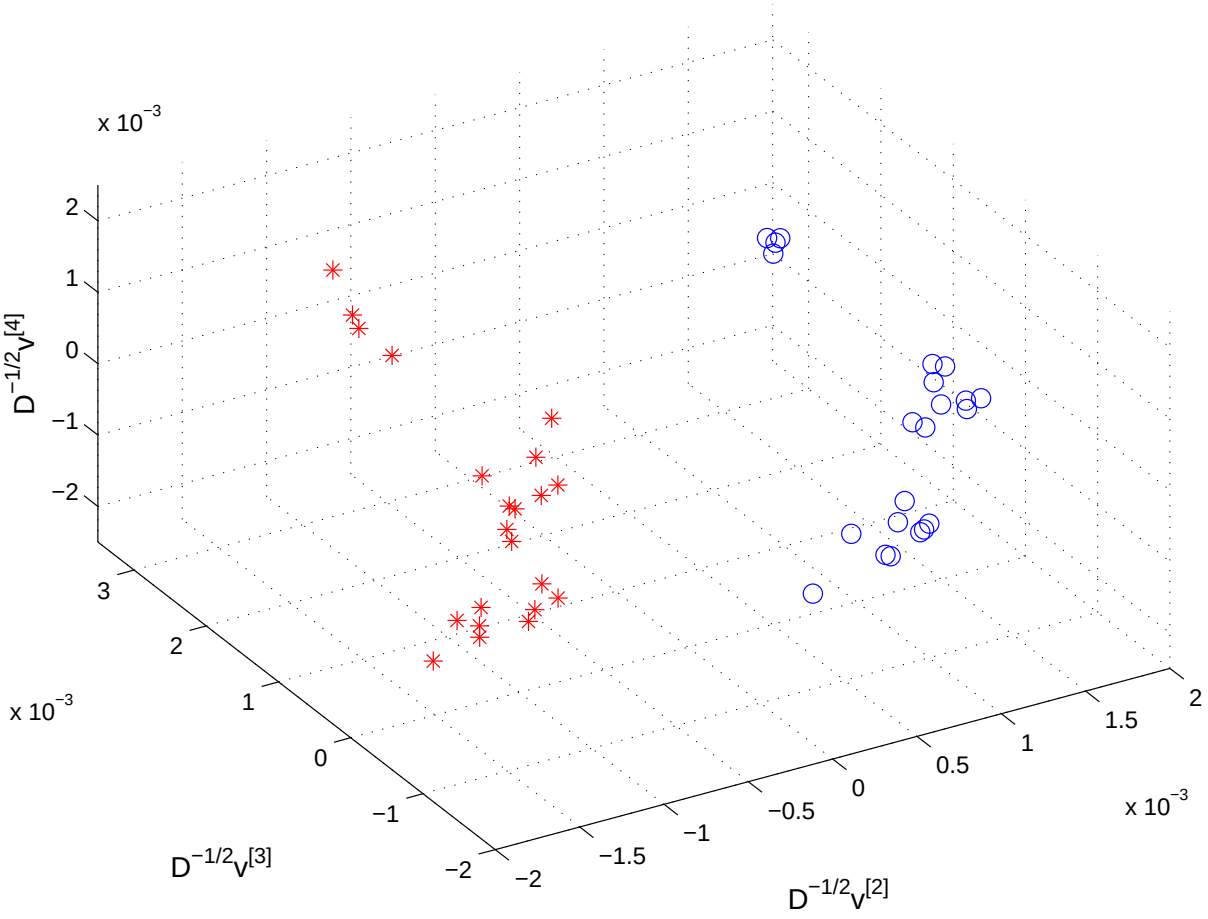


Figure 2: Colon: 3D plot.

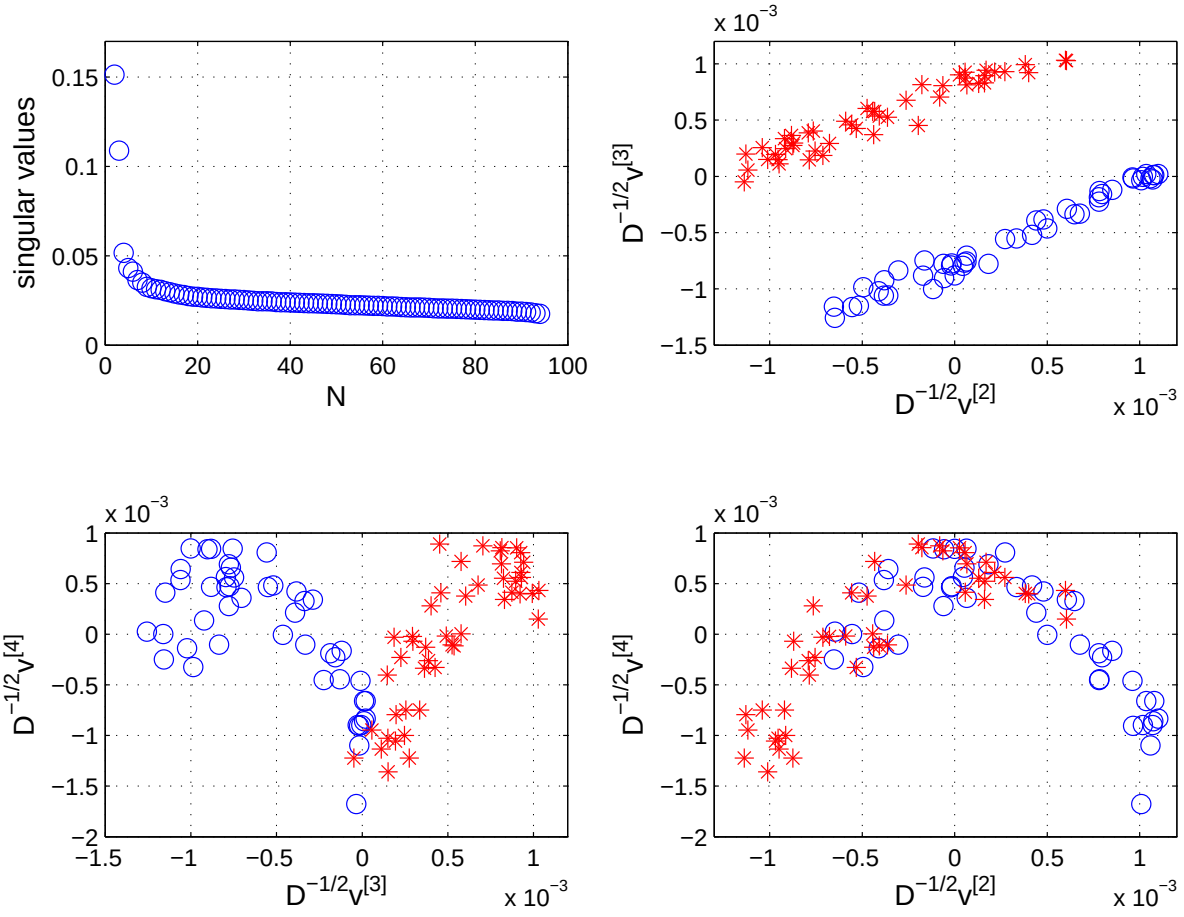


Figure 3: Prostate: tumour (stars) and normal (circles) samples. Singular values $\sigma_2, \sigma_3, \ldots, \sigma_N$ (top left) and scatter plots of three pairs of dominant singular vectors.

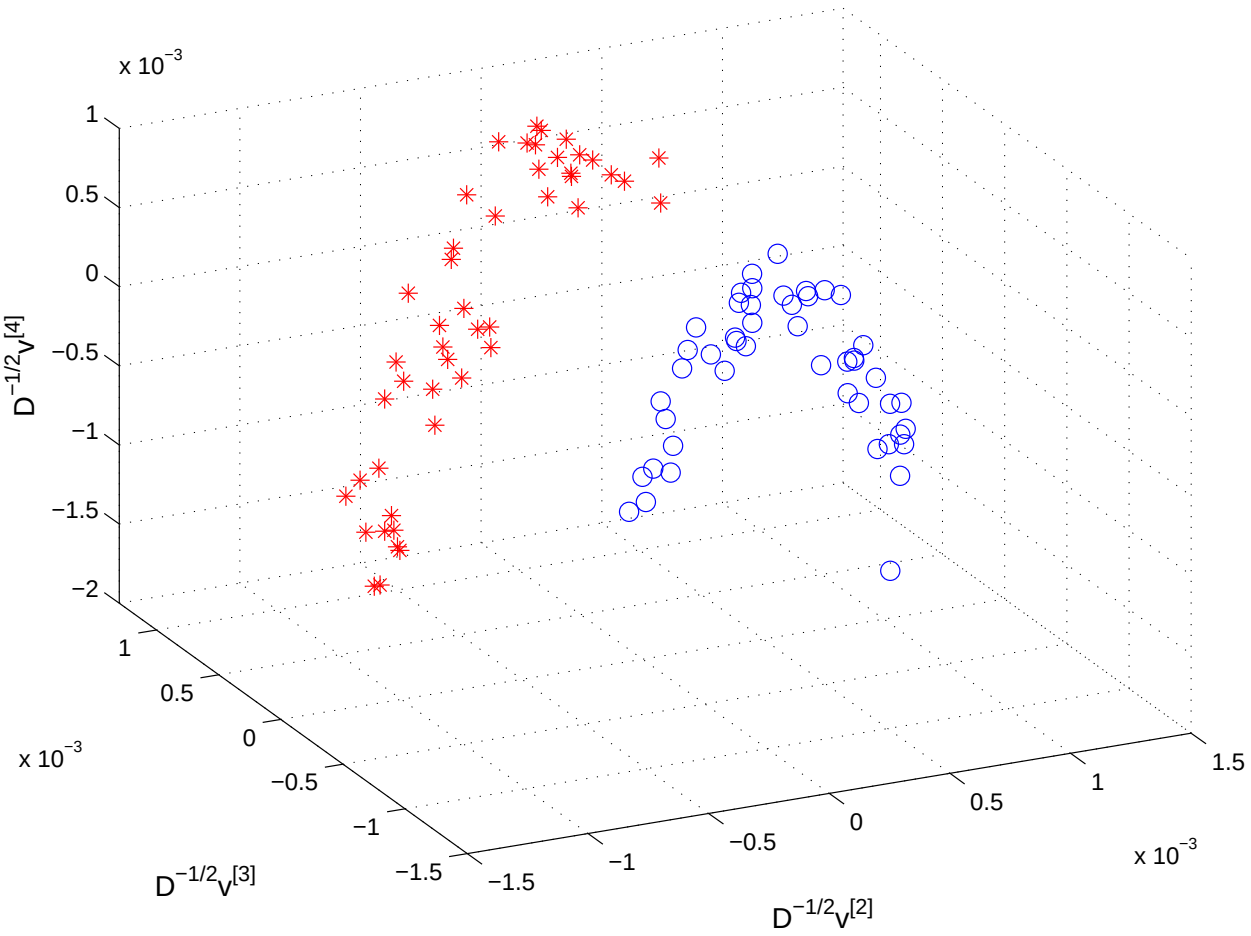


Figure 4: Prostate: 3D plot.