Statistical Science
2008, Vol. 23, No. 3, 383–403
DOI: 10.1214/08-STS266

Multiway Spectral Clustering: A Margin-Based Perspective

Zhihua Zhang and Michael I. Jordan

Abstract. Spectral clustering is a broad class of clustering procedures in which an intractable combinatorial optimization formulation of clustering is "relaxed" into a tractable eigenvector problem, and in which the relaxed solution is subsequently "rounded" into an approximate discrete solution to the original problem. In this paper we present a novel margin-based perspective on multiway spectral clustering. We show that the margin-based perspective illuminates both the relaxation and rounding aspects of spectral clustering, providing a unified analysis of existing algorithms and guiding the design of new algorithms. We also present connections between spectral clustering and several other topics in statistics, specifically minimum-variance clustering, Procrustes analysis and Gaussian intrinsic autoregression.

Key words and phrases: Spectral clustering, spectral relaxation, graph partitioning, reproducing kernel Hilbert space, large-margin classification, Gaussian intrinsic autoregression.

1. INTRODUCTION

Spectral clustering is a promising approach to clustering that has recently been undergoing rapid development (Shi and Malik, 2000; Kannan, Vempala and Vetta, 2000; Zha et al., 2002; Ng, Jordan and Weiss, 2002; Shortreed and Meilă, 2005; Ding, He and Simon, 2005; Bach and Jordan, 2006; von Luxburg, 2007). In the spectral framework a clustering problem is posed as a discrete optimization problem (an integer program). This problem is generally intractable computationally, and approximate solutions are obtained by a two-step procedure in which (1) the problem is "relaxed" into a simplified continuous optimization problem that can be solved efficiently, and (2) the resulting continuous solution is "rounded" into an approximate solution to the original discrete problem. The adjective "spectral" refers to the fact that the relaxed problem generally takes the form of an eigenvector problem (the original objective function involves quadratic constraints, which yields a Rayleigh coefficient in the relaxed problem).

Department of Statistics and Department of Electrical Engineering and Computer Science, University of California, Berkeley, USA (e-mail: zzhang@stat.berkeley.edu; jordan@stat.berkeley.edu).

The solutions of the relaxed problem are often referred to as *spectral embeddings* and have applications outside of the clustering context (Belkin and Niyogi, 2002). Our focus here, however, will be on spectral clustering.

Spectral clustering was first developed in the context of graph partitioning problems (Donath and Hofmann, 1973; Fiedler, 1973), where the problem is to partition a weighted graph into disjoint pieces, minimizing the sum of the weights of the edges linking the disjoint pieces. The methodology is applied to data analysis problems by identifying nodes of the graph with data points and identifying the edge weights with the similarity (or "distance") function used in clustering. The problem then is to choose an appropriate relaxation of the weighted graph partitioning problem and an appropriate rounding procedure. The current literature offers many such choices (see, e.g., von Luxburg, 2007).

Naive formulations of graph cut problems yield uninteresting solutions in which single nodes are separated from the rest of the graph. The spectral formulation becomes interesting (and computationally intractable) when some sort of constraint is imposed so that the partition is balanced. There have been two main approaches to imposing balancing constraints. In the *ratio cut* (RCUT) formulation (Chan, Schlag and

Zien, 1994), the constraints are expressed in terms of cardinalities of subsets of nodes. In the *normalized cut* (NCUT) formulation (Shi and Malik, 2000), the constraints are expressed in terms of the degrees of nodes. In this paper we study a general *penalized cut* (PCUT) formulation that includes RCUT and NCUT as special cases and we emphasize the close relationships between the spectral relaxations resulting from RCUT and NCUT formulations.

A seemingly very different approach to clustering is the classical minimum-variance formulation where one minimizes the trace of the pooled within-class covariance matrix (Webb, 2002). As we show, however, this formulation is closely related to PCUT. In particular, posing the minimum-variance problem in the reproducing kernel Hilbert space (RKHS) defined by a kernel function (Wahba, 1990), we establish a connection between spectral relaxation and minimum-variance clustering by treating the Laplacian matrix in the PCUT formulation as the Moore–Penrose inverse of the kernel matrix in the minimum-variance formulation.

Other forms of clustering procedures have been usefully analyzed in terms of their relationships to discrimination or classification procedures (Webb, 2002), and in the current paper we aim to develop connections of this kind in the case of spectral clustering. In this regard, it is important to note that our focus is on the *multiway* clustering problem, in which a data set is directly partitioned into c sets where $c > 2$. This differs from the classical graph-partitioning literature, where the focus has been on algorithms that partition a graph into two pieces ("binary cuts"), with the problem of partitioning a graph into multiple pieces ("multiway cuts") often approached by the recursive invocation of a binary cut algorithm.

In the case of binary cuts, an interesting connection to classification has been established by Rahimi and Recht (2004), who have noted that NCUT-based spectral clustering can be interpreted as finding a hyperplane in an RKHS that falls in a "gap" in the empirical distribution. In the current paper we show that this idea can be extended to general multiway PCUT spectral relaxation, where the intuitive idea of a "gap" can be expressed precisely using ideas from the classification literature, specifically the idea of a multiclass margin.

Turning to the rounding problem, we note first that for binary cuts the rounding problem is a relatively simple problem, generally involving the choice of a threshold for the elements of an eigenvector (Juhász and Mályusz, 1977; Weiss, 1999). The problem is significantly more complex in the multiway case, however, where it essentially involves an auxiliary clustering problem based on the spectral embedding. For example, Yu and Shi (2003) proposed a rounding scheme that works with an alternative iteration between singular value decomposition (SVD) and nonmaximum suppression, whereas Bach and Jordan (2006) devised K-means and weighted K-means algorithms for rounding. In the current paper we show that rounding can be usefully approached within the framework of Procrustes analysis (Gower and Dijksterhuis, 2004). Moreover, we show that this approach again reveals links between spectral methods and multiway classification; in particular, we show that the auxiliary Procrustes problem that we must solve can be analyzed using the tools of margin-based classification.

Extant multiway spectral algorithms, including those of Bach and Jordan (2006) and Yu and Shi (2003), as well as many others (Ng, Jordan and Weiss, 2002; Zha et al., 2002; Ding, He and Simon, 2005; Shortreed and Meilă, 2005), are based on the representation of spectral embeddings as c-dimensional vectors. The redundancy inherent in using c-dimensional vectors is inconvenient, however, preventing the flow of results from the binary case to the multiway case (Shi and Malik, 2000). The margin-based perspective that we pursue here shows the value of working with a nonredundant, $(c-1)$-dimensional representation of the spectral embedding.

Our overall approach to spectral clustering is as follows. We first construct a nonredundant, margin-based representation of multiway spectral relaxation problems. Such a margin-based spectral relaxation is a tractable constrained eigenvalue problem. We then carry out a rounding scheme by solving an auxiliary Procrustes problem, which is again associated with a margin-based classification method. We refer to the resulting clustering framework—margin-based spectral relaxation with margin-based rounding—as *margin-based spectral clustering*.

The margin-based approach not only provides substantial insight into the relationships among spectral clustering procedures, but it also yields probabilistic interpretations of these procedures. Specifically, we show that the spectral relaxation obtained from the PCUT framework can be interpreted as a form of Gaussian intrinsic autoregression (Besag and Kooperberg, 1995). These are limiting forms of Gaussian conditional autoregressions (Besag, 1974; Mardia, 1988) that retain the Markov property (two vertices in a graph

are not connected if and only if their corresponding embeddings in the intrinsic autoregression are conditionally independent).

In summary, the current paper develops a mathematical perspective on spectral clustering that unifies the various algorithms that have been studied and emphasizes connections to other areas of statistics. Specifically we discuss connections to multiway classification, reproducing kernel Hilbert space methods, Procrustes analysis and Gaussian intrinsic autoregression.

The remainder of the paper is organized as follows. Sections 2 and 4 describe multiway spectral relaxation problems based on the general PCUT formulation and the minimum variance formulation, respectively. The relationship between these two formulations is also discussed in Section 4. In Section 3 we present two rounding schemes, one based on Procrustean transformation and the other based on K-means. We present a geometric perspective on spectral clustering using margin-based principles in Section 5, and we discuss the connection to Gaussian intrinsic autoregression models in Section 6. Experimental comparisons are given in Section 7 and we present our conclusions in Section 8. Note that several proofs are deferred to the Appendix.

We use the following notation in this paper. $\mathbf{I}_m$ denotes the $m \times m$ identity matrix, $\mathbf{1}_m$ the $m \times 1$ of ones, $\mathbf{0}$ the zero vector or matrix zero of appropriate size and $\mathbf{H}_m = \mathbf{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m'$ the $m \times m$ centering matrix. For an $n \times 1$ vector $\mathbf{a} = (a_1, \ldots, a_n)'$, diag($\mathbf{a}$) represents the $n \times n$ diagonal matrix with $a_1, \ldots, a_n$ as its diagonal entries and $\|\mathbf{a}\|$ is the Euclidean norm of $\mathbf{a}$. For an $m \times m$ matrix $\mathbf{A} = [a_{ij}]$, we let dg($\mathbf{A}$) be the diagonal matrix with $a_{11}, \ldots, a_{mm}$ as its diagonal entries, $\mathbf{A}^+$ be the Moore–Penrose inverse of $\mathbf{A}$, tr($\mathbf{A}$) be the trace of $\mathbf{A}$, rk($\mathbf{A}$) be the rank of $\mathbf{A}$ and $\|\mathbf{A}\|_F$ be the Frobenius norm of $\mathbf{A}$.

2. SPECTRAL RELAXATION FOR PENALIZED CUTS

Given a set of n d-dimensional data points, $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, our goal is to cluster the $\mathbf{x}_i$ into c disjoint classes such that each $\mathbf{x}_i$ belongs to one and only one class. We consider a graphical representation of this problem. Let $V = \{1, 2, \ldots, n\}$ denote the index set of the data points and consider an undirected graph $\mathcal{G} = (V, \mathcal{E})$ where V is the set of nodes in the graph and $\mathcal{E}$ is the set of edges. Associated with the graph is a symmetric $n \times n$ *affinity matrix* (also referred to as a *similarity matrix*), $\mathbf{W} = [w_{ij}]$, defined on pairs of indices such that $w_{ij} \geq 0$ for $(i, j) \in \mathcal{E}$ and $w_{ij} = 0$ otherwise. The values w_{ij} are often obtained via a function evaluated on the corresponding pairs of data vectors; that is, $w_{ij} = \psi(\mathbf{x}_i, \mathbf{x}_j)$ for some (symmetric) function ψ. A variety of different ways to map a data set into a graph $\mathcal{G}$ and an affinity matrix $\mathbf{W}$ have been explored in the literature; for a review see von Luxburg (2007).

The problem is thus to partition V into c subsets V_j; that is, $V_i \cap V_j = \varnothing$ for $i \neq j$ and $\bigcup_{j=1}^{c} V_j = V$, where the cardinality of V_j is n_j so that $\sum_{j=1}^{c} n_j = n$. This problem is typically formulated as a combinatorial optimization problem. Let $W(A, B) = \sum_{i \in A, j \in B} w_{ij}$ for two (possibly overlapping) subsets A and B of V and consider the following *multiway penalized cut* criterion:

$$\text{PCUT} = \sum_{j=1}^{c} \frac{W(V_j, V) - W(V_j, V_j)}{\sum_{i \in V_j} \pi_i}, \tag{2.1}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)'$ is a user-defined vector of weights (examples are provided below) with $\pi_i > 0$ for all i. The numerator of each of the terms in this expression is equal to the sum of the affinities on edges leaving the subset V_j. Thus the minimization of PCUT with respect to the partition $\{V_1, \ldots, V_c\}$ aims at finding a partition in which edges with large affinities tend to stay within the individual subsets V_j. The denominator weights $\sum_{i \in V_j} \pi_i$ encode a notion of "size" of the subsets V_j and act to balance the partition.

The PCUT criterion can also be written in matrix notation as follows. Define $\mathbf{D} = \text{diag}(\mathbf{W}\mathbf{1}_n)$ and let $\mathbf{L} = \mathbf{D} - \mathbf{W}$ denote the *Laplacian matrix* of the graph. (An $n \times n$ matrix $\mathbf{L} = [l_{ij}]$ is a *Laplacian matrix* if $l_{ii} > 0$ for $i = 1, \ldots, n$; $l_{ij} = l_{ji} \leq 0$ for $i \neq j$; $\sum_{j=1}^{n} l_{ij} = 0$ for $i = 1, \ldots, n$. Note that Laplacian matrices are positive semidefinite (Mohar, 1991).) Let $\boldsymbol{\Pi} = \text{diag}(\pi_1, \ldots, \pi_n)$ be a diagonal matrix of weights. Let $t_i \in \{1, \ldots, c\}$ denote the assignment of $\mathbf{x}_i$ to a cell in the partition and define the indicator matrix $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_n]'$, where $\mathbf{e}_i \in \{0, 1\}^{c \times 1}$ is a binary vector whose t_ith entry is one and all other entries are zero. It can now be readily verified that PCUT takes the following form:

$$\text{PCUT} = \text{tr}\big(\mathbf{E}'\mathbf{L}\mathbf{E}(\mathbf{E}'\boldsymbol{\Pi}\mathbf{E})^{-1}\big), \tag{2.2}$$

where it is helpful to note that $(\mathbf{E}'\boldsymbol{\Pi}\mathbf{E})^{-1}$ is a diagonal matrix, implying that PCUT is simply a scaled quadratic form. We wish to optimize this scaled quadratic form with respect to $\mathbf{E}$.

Two well-known examples of the PCUT problem are the *ratio cut* (RCUT) problem (Chan, Schlag and

Zien, 1994), in which $\mathbf{\Pi} = \mathbf{I}_n$, and the *normalized cut* (NCUT) problem (Shi and Malik, 2000), in which $\mathbf{\Pi} = \mathbf{D}$. In the RCUT problem the notion of "size" of a subset V_j is simply the number of nodes in the subset, whereas in the NCUT problem "size" is captured by the total degree of the nodes in the subset.

The spectral clustering approach to minimizing PCUT involves two stages: (1) we *relax* the problem into a tractable spectral analysis problem in which continuous variables replace the indicators $\mathbf{E}$, and (2) we then employ a *rounding* scheme to obtain a partition $\{V_1, \ldots, V_n\}$ from the continuous relaxation. In the remainder of this section, we focus on the first step (the relaxation) and we return to the rounding problem in Section 3.

The standard presentation of spectral relaxation proceeds somewhat differently in the case of a binary partition and a multiway partition (von Luxburg, 2007). In both cases, spectral relaxation is motivated by the observation that the PCUT criterion in (2.2) has the form of a Rayleigh coefficient, and that replacing the indicator matrix $\mathbf{E}$ with a real-valued matrix yields a classical generalized eigenvector problem. In the binary case, the indicator matrix $\mathbf{E}$ has two columns, which yields two generalized eigenvectors in the relaxed problem. However, in the subsequent rounding procedure, the problem is to discriminate between two classes, for which a single vector direction suffices. To deal with this redundancy it is standard to place a (linear) constraint upon the relaxation, such that it is the second generalized eigenvector that is used for rounding (von Luxburg, 2007). In the multiway case, on the other hand, no such constraint is imposed; the redundancy inherent in having c generalized eigenvectors to discriminate among c classes is generally not addressed. (It is resolved implicitly at the rounding stage.)

We find this distinction between the binary case and the multiway case to be inconvenient, and thus in the approach to be described in the following section we adopt an idea from the literature on multiway classification (e.g., Zou, Zhu and Hastie, 2006; Shen and Wang, 2007) where nonredundant, $(c-1)$-dimensional vectors are used to discriminate among c classes. These vectors are referred to as *margin vectors*. We refer the reader to the classification literature for the geometric rationale behind the terminology of "margin" (although we note that a geometric interpretation of margin vectors will also appear in the current paper in Section 5.1).

2.1 Spectral Relaxation

To formulate a spectral relaxation of (2.2), we replace the indicator matrix $\mathbf{E}$ with a real $n \times (c-1)$ matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]'$. The following proposition, which is based on a result of Bach and Jordan (2006), shows that we can express the PCUT criterion in terms of real-valued matrices $\mathbf{Y}$ satisfying certain conditions.

PROPOSITION 1. *Let* $\mathbf{Y}$ *be an* $n \times (c-1)$ *real matrix such that*: (a) *the columns of* $\mathbf{Y}$ *are piecewise constant with respect to the partition* $\mathbf{E}$, (b) $\mathbf{Y}'\mathbf{\Pi}\mathbf{Y} = \mathbf{I}_{c-1}$ *and* (c) $\mathbf{Y}'\mathbf{\Pi}\mathbf{1}_n = \mathbf{0}$. *Then* PCUT *is equal to* $\operatorname{tr}(\mathbf{Y}'\mathbf{L}\mathbf{Y})$.

The proof of Proposition 1 is given in Appendix A.1.

For this proposition to be useful it is necessary to show that matrices satisfying the three conditions in Proposition 1 exist. Condition (a) for $\mathbf{Y}$ is equivalent to the statement that $\mathbf{Y}$ can be expressed as $\mathbf{Y} = \mathbf{E}\mathbf{\Psi}$ where $\mathbf{\Psi}$ is some $c \times (c-1)$ matrix. Thus, the question becomes whether there exists a $\mathbf{\Psi}$ such that $\mathbf{Y}$ satisfies conditions (b)–(c). In Appendix A.2 we provide a general procedure for constructing such a $\mathbf{\Psi}$. This establishes the following proposition.

PROPOSITION 2. *Matrices* $\mathbf{Y}$ *satisfying the three conditions in Proposition* 1 *exist*.

We now obtain a spectral relaxation by dropping condition (a). This yields the following optimization problem:

$$\begin{aligned} &\min_{\mathbf{Y}\in\mathbb{R}^{n\times(c-1)}} \operatorname{tr}(\mathbf{Y}'\mathbf{L}\mathbf{Y}) \\ &\text{s.t. } \mathbf{Y}'\mathbf{\Pi}\mathbf{Y} = \mathbf{I}_{c-1} \text{ and } \mathbf{Y}'\mathbf{\Pi}\mathbf{1}_n = \mathbf{0}, \end{aligned} \tag{2.3}$$

which is a constrained generalized eigenvalue problem.

2.2 Solving the Spectral Relaxation

Letting $\mathbf{Y}_0 = \mathbf{\Pi}^{1/2}\mathbf{Y}$, we can transform (2.3) into the following problem:

$$\begin{aligned} &\min_{\mathbf{Y}_0\in\mathbb{R}^{n\times(c-1)}} \operatorname{tr}(\mathbf{Y}_0'\mathbf{\Pi}^{-1/2}\mathbf{L}\mathbf{\Pi}^{-1/2}\mathbf{Y}_0), \\ &\text{s.t. } \mathbf{Y}_0'\mathbf{Y}_0 = \mathbf{I}_{c-1} \text{ and } \mathbf{Y}_0'\mathbf{\Pi}^{1/2}\mathbf{1}_n = \mathbf{0}. \end{aligned} \tag{2.4}$$

The solution to this constrained eigenvalue problem is given in the following theorem.

THEOREM 1. *Suppose that* $\mathbf{L}$ *is a real symmetric matrix such that* $\mathbf{L}\mathbf{1}_n = \mathbf{0}$ *and suppose that the diagonal entries of* $\mathbf{\Pi}$ *are all positive. Let* $\boldsymbol{\mu}_1 = \alpha\mathbf{\Pi}^{1/2}\mathbf{1}_n$ *be the eigenvector associated with the eigenvalue* $\gamma_1 = 0$ *of* $\mathbf{\Pi}^{-1/2}\mathbf{L}\mathbf{\Pi}^{-1/2}$, *where* $\alpha^2 = 1/(\mathbf{1}_n'\mathbf{\Pi}\mathbf{1}_n)$. *Let the remaining eigenvalues of* $\mathbf{\Pi}^{-1/2}\mathbf{L}\mathbf{\Pi}^{-1/2}$ *be*

arranged so that $\gamma_2 \le \cdots \le \gamma_n$, *and let the corresponding orthonormal eigenvectors be denoted by* $\boldsymbol{\mu}_i$, $i = 2, \ldots, n$. *Then the solution of problem* (2.4) *is* $\bar{\mathbf{Y}}_0 = \mathbf{UQ}$ *where* $\mathbf{U} = [\boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_c]$ *and* $\mathbf{Q}$ *is an arbitrary* $(c-1) \times (c-1)$ *orthonormal matrix, with* $\min\{\operatorname{tr}(\mathbf{Y}_0'\boldsymbol{\Pi}^{-1/2}\mathbf{L}\boldsymbol{\Pi}^{-1/2}\mathbf{Y}_0)\} = \sum_{i=2}^{c}\gamma_i$. *Furthermore, if* $\gamma_c < \gamma_{c+1}$, *then* $\bar{\mathbf{Y}}_0$ *is a strict local minimum of* $\operatorname{tr}(\mathbf{Y}_0'\boldsymbol{\Pi}^{-1/2}\mathbf{L}\boldsymbol{\Pi}^{-1/2}\mathbf{Y}_0)$.

It follows from the theorem that the solution of problem (2.3) is $\mathbf{Y} = \boldsymbol{\Pi}^{-1/2}\mathbf{UQ}$. The proof of Theorem 1 is given in Appendix A.3. It is important to note for our later work that this theorem does not require $\mathbf{L}$ to be Laplacian or even positive semidefinite.

The condition $\gamma_c < \gamma_{c+1}$ implies a nonzero eigengap (Chung, 1997). In practice, the eigengap is often used as a criterion to determine the number of classes in clustering scenarios. An idealized situation is that the multiplicity of the eigenvalue zero is c.

3. ROUNDING SCHEMES

We now consider the problem of rounding—transforming the real-valued solution of a spectral relaxation problem into a discrete set of values that can be interpreted as a clustering. In this section we present two different solutions to the rounding problem, one based on Procrustes analysis and the other based on the K-means algorithm.

3.1 Procrustean Transformation for Rounding

In Theorem 1 we have shown that the solution of the spectral relaxation problem is a matrix $\mathbf{Y} = \boldsymbol{\Pi}^{-1/2}\mathbf{UQ}$, where $\mathbf{Q}$ is an arbitrary orthogonal matrix. We have also seen, in Proposition 1, that a matrix $\mathbf{Y}$ in which the columns of $\mathbf{Y}$ are piecewise constant with respect to a partition $\mathbf{E}$ provides a representation of the objective function value PCUT. If we had such a matrix $\mathbf{Y}$ in hand we could straightforwardly find the partition $\mathbf{E}$: Letting $t_i = \arg\max_j\{y_{ij}\}$, allocate $\mathbf{x}_i$ to the t_ith class if $y_{it_i} > 0$ and to the cth class otherwise. On the other hand, if we had the partition we could attempt to find an orthogonal matrix $\mathbf{Q}$ such that $\mathbf{Y} = \boldsymbol{\Pi}^{-1/2}\mathbf{UQ}$ is as close as possible to the partition. This latter problem can be treated as a problem in Procrustes analysis (Gower and Dijksterhuis, 2004).

Specifically, given an indicator matrix $\mathbf{E}$ we pose the following Procrustes problem:

$$\arg\min_{\mathbf{Q}} L(\mathbf{Q}) = \operatorname{tr}(\mathbf{EG} - \mathbf{UQ})(\mathbf{EG} - \mathbf{UQ})', \tag{3.1}$$

where $\mathbf{G} = [\mathbf{I}_{c-1} - \frac{1}{c}\mathbf{1}_{c-1}\mathbf{1}_{c-1}', -\frac{1}{c}\mathbf{1}_{c-1}]'$. This problem has an analytical solution: Denote the singular value decomposition of $\mathbf{U}'\mathbf{EG}$ as $\mathbf{U}'\mathbf{EG} = \boldsymbol{\Theta}\boldsymbol{\Lambda}\mathbf{V}'$. Then the minimizing value of $\mathbf{Q}$ in L is given by $\mathbf{Q} = \boldsymbol{\Theta}\mathbf{V}'$ (see, e.g., Mardia, Kent and Bibby, 1979, page 416).

Algorithm 1. *Spectral Clustering with Procrustean Rounding*

1: *Input*: An affinity matrix $\mathbf{W}$ and a diagonal matrix $\boldsymbol{\Pi}$
2: *Relaxation*: Obtain $\mathbf{Y} = \boldsymbol{\Pi}^{-1/2}\mathbf{UQ}$ from problem (2.3)
3: *Initialize*: Choose the initial partition $\mathbf{E}$
4: *Rounding*: Repeat the following procedure until convergence:
   (a) Recompute $\mathbf{EG}$, implement the SVD of $\mathbf{U}'\mathbf{EG}$ as $\mathbf{U}'\mathbf{EG} = \boldsymbol{\Theta}\boldsymbol{\Lambda}\mathbf{V}'$ and let $\mathbf{Q} = \boldsymbol{\Theta}\mathbf{V}'$
   (b) Recompute $\mathbf{Y} = [y_{ij}] = \boldsymbol{\Pi}^{-1/2}\mathbf{UQ}$, compute $t_i = \arg\max_j y_{ij}$, and recompute $\mathbf{E}$ by allocating the ith data point to class t_i if $\max_j y_{ij} > 0$ and to class c otherwise
5: Output $\{t_1, \ldots, t_n\}$.

We summarize this Procrustean approach to rounding in algorithmic form in Algorithm 1 in the context of a generic spectral clustering algorithm.

Yu and Shi (2003) have presented a rounding algorithm that is similar to the Procrustean approach we have presented but different in detail. The authors work with an $n \times c$ matrix Z and solve the relaxation $\min \operatorname{tr}(\mathbf{Z}'\mathbf{LZ})$ subject to $\mathbf{Z}'\mathbf{DZ} = \mathbf{I}_c$. Given the solution $\mathbf{Z}$ of this relaxation, the authors then compute $\hat{\mathbf{Z}} = [\hat{z}_{ij}] = \operatorname{dg}(\mathbf{ZZ}')^{-1/2}\mathbf{Z}$. Their rounding scheme is to allocate the ith data point to class t_i if $t_i = \arg\max_j \hat{z}_{ij}$. This method can be viewed as imposing a constraint; in particular, note that the norms of the rows of $\hat{\mathbf{Z}}$ are equal to 1. To motivate this constraint, the authors assume that the solution $\mathbf{Z}$ can be expressed as a rescaling of $\hat{\mathbf{Z}}$: $\mathbf{Z} = \hat{\mathbf{Z}}(\hat{\mathbf{Z}}'\mathbf{D}\hat{\mathbf{Z}})^{-1/2}$. Inverting this expression yields $\hat{\mathbf{Z}} = \operatorname{dg}(\mathbf{ZZ}')^{-1/2}\mathbf{Z}$. But it is not clear that a solution $\mathbf{Z}$ of the relaxation can be expressed in this form; the constraints on $\hat{\mathbf{Z}}$ are not incorporated into the relaxation. The use of $\hat{\mathbf{Z}}$ defined in this way must be viewed as a heuristic post-processing procedure. The Procrustean approach that we have presented in this section provides a resolution of this difficulty; that approach requires no post-processing of the matrix obtained from the spectral relaxation.

We return to the Procrustean approach in Section 5, where we provide additional justification for Procrustean rounding based on a connection to margin maximization.

3.2 K-means for Rounding

Another approach to removing the "nuisance" orthogonal matrix $\mathbf{Q}$ is to consider rounding methods that are invariant to rotation. The standard K-means algorithm provides an example, and numerous authors have proposed using K-means on the embedding obtained from spectral relaxation as a heuristic rounding procedure (von Luxburg, 2007). Bach and Jordan (2006) have made this approach more formal by showing that (weighted) K-means arises when the rounding problem is formalized in terms of a difference between projection matrices. In this section we review this formulation within our nonredundant representation of spectral relaxation.

Let us rewrite PCUT as

$$\text{PCUT} = \operatorname{tr}(\mathbf{E}'\mathbf{H}_\pi \mathbf{L}\mathbf{H}_\pi' \mathbf{E}(\mathbf{E}'\mathbf{\Pi}\mathbf{E})^{-1}),$$

where we define $\mathbf{H}_\pi = \mathbf{I}_n - \frac{1}{\boldsymbol{\pi}'\mathbf{1}_n}\boldsymbol{\pi}'\mathbf{1}_n$ where we use the fact that $\mathbf{H}_\pi \mathbf{L}\mathbf{H}_\pi' = \mathbf{L}$. Defining $\mathbf{E}_\pi \triangleq \mathbf{H}_\pi' \cdot \mathbf{E}(\mathbf{E}'\mathbf{\Pi}\mathbf{E})^{-1/2}$, we observe that the number of degrees of freedom of both $\mathbf{Y}$ and $\mathbf{E}_\pi$ is $(n-1)(c-1)$. Moreover, given that $\mathbf{E}_\pi'\mathbf{\Pi}\mathbf{E}_\pi = \mathbf{I}_c - (\mathbf{E}'\mathbf{\Pi}\mathbf{E})^{-1/2}\mathbf{E}'\boldsymbol{\pi}\boldsymbol{\pi}' \cdot \mathbf{E}(\mathbf{E}'\mathbf{\Pi}\mathbf{E})^{-1/2}/(\boldsymbol{\pi}'\mathbf{1}_n)$ and $\boldsymbol{\pi}'\mathbf{E}(\mathbf{E}'\mathbf{\Pi}\mathbf{E})^{-1}\mathbf{E}'\boldsymbol{\pi} = \boldsymbol{\pi}'\mathbf{1}_n$, there exists a $c \times c$ permutation matrix $\mathbf{P}$ such that

$$\mathbf{P}\mathbf{E}_\pi'\mathbf{\Pi}\mathbf{E}_\pi\mathbf{P}' = \begin{bmatrix} \mathbf{I}_{c-1} & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{Y}' \\ \mathbf{0} \end{bmatrix} \mathbf{\Pi}[\mathbf{Y}, \mathbf{0}];$$

this suggests viewing $\mathbf{Y}$ as an approximation to $\mathbf{E}_\pi$ in the metric given by $\mathbf{\Pi}$. We quantify this by defining the following distortion between the projection matrices defined by $\mathbf{Y}$ and $\mathbf{E}_\pi$:

$$\begin{aligned} J_k(\mathbf{E}_\pi, \mathbf{Y}) &= \tfrac{1}{2}\|\mathbf{Y}\mathbf{\Pi}\mathbf{Y}' - \mathbf{E}_\pi\mathbf{\Pi}\mathbf{E}_\pi'\|_F^2 \\ &= c - 1 - \operatorname{tr}(\mathbf{Y}'\mathbf{\Pi}\mathbf{E}(\mathbf{E}'\mathbf{\Pi}\mathbf{E})^{-1}\mathbf{E}'\mathbf{\Pi}\mathbf{Y}). \end{aligned}$$

This objective function can be represented as the solution of a weighted K-means problem, as shown by the following result which is due to Bach and Jordan (2006):

THEOREM 2. *Let $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]'$ be a solution of problem (2.3). For any partition $\{V_1, \ldots, V_c\}$, the criterion $F(\mathbf{m}_1, \ldots, \mathbf{m}_c) = \sum_{j=1}^c \sum_{i \in V_j} \|\mathbf{y}_i - \mathbf{m}_j\|^2$ achieves its minimum $J_k(\mathbf{E}_\pi, \mathbf{Y})$ at $\mathbf{m}_j = \frac{1}{\sum_{i \in V_j}\pi_i} \cdot \sum_{i \in V_j}\pi_i \mathbf{y}_i$.*

Thus by updating the mean vectors $\mathbf{m}_j$ in the weighted K-means algorithm we match the criterion $J_k(\mathbf{E}_\pi, \mathbf{Y})$, and by updating the partition using weighted K-means we go downhill in the criterion.

Algorithm 2. *Spectral Clustering with K-means Rounding*

1: *Input*: An affinity matrix $\mathbf{W}$ and a diagonal matrix $\mathbf{\Pi}$
2: *Relaxation*: Obtain $\mathbf{Y} = \mathbf{\Pi}^{-1/2}\mathbf{U}\mathbf{Q}$ from problem (2.3)
3: *Initialize*: Choose the initial partition $\mathbf{E}$
4: *Rounding*: Repeat the following procedure until convergence:
   (a) Compute $\mathbf{m}_j = \frac{1}{\sum_{i \in V_j}\pi_i}\sum_{i \in V_j}\pi_i \mathbf{y}_i$
   (b) Find $t_i = \arg\min_j \|y_i - \mathbf{m}_j\|$, and recompute $\mathbf{E}$ by allocating the ith data point to class t_i
5: *Output* $\{t_1, \ldots, t_n\}$.

Note that in the special case of the RCUT formulation, we obtain the conventional unweighted K-means algorithm (given that $\pi_i = 1$ in that case).

We summarize the K-means approach to rounding in algorithmic form in Algorithm 2.

4. SPECTRAL CLUSTERING AND MINIMUM-VARIANCE CRITERIA

In this section and the following two sections we present some relationships between spectral clustering and various topics in statistics. Our goal is both to illuminate the spectral approach and to suggest directions for further research.

Minimum-variance clustering is a classical approach to clustering (Webb, 2002). In this section, following Zha et al. (2002) and Dhillon, Guan and Kulis (2007), we present spectral solutions to the minimum-variance clustering problem, and we establish connections between minimum-variance clustering and the PCUT framework.

Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathcal{X} \subset \mathbb{R}^d$ denote the observed data. The pooled within-class covariance matrix $\mathbf{S}_W$ is given by

$$\mathbf{S}_W = \frac{1}{n}\sum_{j=1}^c \sum_{i \in V_j} (\mathbf{x}_i - \mathbf{m}_j)(\mathbf{x}_i - \mathbf{m}_j)',$$

where $\mathbf{m}_j = \frac{1}{n_j}\sum_{i \in V_j}\mathbf{x}_i$. Consider the trace of the within-class covariance matrix:

$$\operatorname{tr}(\mathbf{S}_W) = \frac{1}{n}\sum_{j=1}^c \sum_{i \in V_j} \|\mathbf{x}_i - \mathbf{m}_j\|^2.$$

Clustering algorithms which are based on the minimization of this trace are referred to as minimum-variance methods.

In order to establish a connection with the spectral relaxation presented in Section 2, we define a weighted pooled within-class covariance matrix in an reproducing kernel Hilbert space (RKHS) induced by a reproducing kernel K. In particular, assume that we are given the reproducing kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $K(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\phi}(\mathbf{x}_i)'\boldsymbol{\phi}(\mathbf{x}_j)$ for $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$, where $\boldsymbol{\phi}(\mathbf{x})$ is called a *feature vector* corresponding to a data point $\mathbf{x} \in \mathcal{X}$. In the sequel, we use the tilde notation to denote feature vectors. Thus, the data matrix in the feature space is denoted as $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_n]'$. The centered kernel matrix takes the form $\mathbf{K} = \mathbf{H}_n\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{H}_n$; note that it is positive semidefinite and satisfies $\mathbf{K}\mathbf{1}_n = \mathbf{0}$.

Generalizing slightly, we introduce weighted versions of the sample covariance matrix $\tilde{\mathbf{S}}$, the between-class covariance matrix $\tilde{\mathbf{S}}_B$ and the within-class covariance matrix $\tilde{\mathbf{S}}_W$:

$$\tilde{\mathbf{S}} = \frac{1}{\sum_{i=1}^n \pi_i}\sum_{i=1}^n \pi_i(\tilde{\mathbf{x}}_i - \tilde{\mathbf{m}})(\tilde{\mathbf{x}}_i - \tilde{\mathbf{m}})',$$

$$\tilde{\mathbf{S}}_B = \frac{1}{\sum_{i=1}^n \pi_i}\sum_{j=1}^c \sum_{i\in V_j} \pi_i(\tilde{\mathbf{m}}_j - \tilde{\mathbf{m}})(\tilde{\mathbf{m}}_j - \tilde{\mathbf{m}})',$$

$$\tilde{\mathbf{S}}_W = \frac{1}{\sum_{i=1}^n \pi_i}\sum_{j=1}^c \sum_{i\in V_j} \pi_i(\tilde{\mathbf{x}}_i - \tilde{\mathbf{m}}_j)(\tilde{\mathbf{x}}_i - \tilde{\mathbf{m}}_j)',$$

where the π_i are known positive weights, $\tilde{\mathbf{m}} = \frac{1}{\sum_{i=1}^n \pi_i} \cdot \sum_{i=1}^n \pi_i\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{m}}_j = \frac{1}{\sum_{i\in V_j}\pi_i}\sum_{i\in V_j}\pi_i\tilde{\mathbf{x}}_i$. It is clear that $\tilde{\mathbf{S}}_W = \tilde{\mathbf{S}} - \tilde{\mathbf{S}}_B$.

We now formulate a minimum-variance clustering problem in the RKHS as the minimization of $\operatorname{tr}(\tilde{\mathbf{S}}_W)$, which is given by

$$\operatorname{tr}(\tilde{\mathbf{S}}_W) = \frac{1}{\sum_{i=1}^n \pi_i}\sum_{j=1}^c \sum_{i\in V_j} \pi_i\|\tilde{\mathbf{x}}_i - \tilde{\mathbf{m}}_j\|^2.$$

Like the minimization of PCUT, this minimization is computationally infeasible in general. It is therefore natural to consider minimizing $\operatorname{tr}(\tilde{\mathbf{S}}_W)$ by using the spectral relaxations presented in Section 2.2. We present a way to do this in the following section.

4.1 Spectral Relaxation in the RKHS

Let us rewrite $\tilde{\mathbf{S}}$ and $\tilde{\mathbf{S}}_B$ as

$$\tilde{\mathbf{S}} = \frac{1}{\boldsymbol{\pi}'\mathbf{1}_n}\tilde{\mathbf{X}}'\mathbf{H}_\pi\boldsymbol{\Pi}\mathbf{H}_\pi'\tilde{\mathbf{X}}$$

and

$$\tilde{\mathbf{S}}_B = \frac{1}{\boldsymbol{\pi}'\mathbf{1}_n}\tilde{\mathbf{X}}'\mathbf{H}_\pi\boldsymbol{\Pi}\mathbf{E}(\mathbf{E}'\boldsymbol{\Pi}\mathbf{E})^{-1}\mathbf{E}'\boldsymbol{\Pi}\mathbf{H}_\pi'\tilde{\mathbf{X}},$$

recalling that $\mathbf{H}_\pi = \mathbf{I}_n - \frac{1}{\boldsymbol{\pi}'\mathbf{1}_n}\boldsymbol{\pi}\mathbf{1}_n'$. This yields

$$\tilde{\mathbf{S}}_W = \frac{1}{\boldsymbol{\pi}'\mathbf{1}_n}[\tilde{\mathbf{X}}'\mathbf{H}_\pi\boldsymbol{\Pi}\mathbf{H}_\pi'\tilde{\mathbf{X}} - \tilde{\mathbf{X}}'\mathbf{H}_\pi\boldsymbol{\Pi}\mathbf{E}(\mathbf{E}'\boldsymbol{\Pi}\mathbf{E})^{-1}\mathbf{E}'\boldsymbol{\Pi}\mathbf{H}_\pi'\tilde{\mathbf{X}}].$$

The minimization of $\operatorname{tr}(\tilde{\mathbf{S}}_W)$ is thus equivalent to the maximization of

$$T = \operatorname{tr}(\mathbf{E}'\boldsymbol{\Pi}\mathbf{H}_\pi'\mathbf{K}\mathbf{H}_\pi\boldsymbol{\Pi}\mathbf{E}(\mathbf{E}'\boldsymbol{\Pi}\mathbf{E})^{-1}), \tag{4.1}$$

because $\tilde{\mathbf{X}}'\mathbf{H}_\pi\boldsymbol{\Pi}\mathbf{H}_\pi'\tilde{\mathbf{X}}$ is independent of $\mathbf{E}$ and we have $\mathbf{H}_n\mathbf{H}_\pi = \mathbf{H}_\pi$. Let $\boldsymbol{\Delta} = [\delta_{ij}^2]$, where δ_{ij} is the squared distance between $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_j$, that is,

$$\begin{aligned}\delta_{ij}^2 &= (\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j)'(\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j)' \\ &= K(\mathbf{x}_i, \mathbf{x}_i) + K(\mathbf{x}_j, \mathbf{x}_j) - 2K(\mathbf{x}_i, \mathbf{x}_j).\end{aligned}$$

Given that $-\frac{1}{2}\mathbf{H}_\pi'\boldsymbol{\Delta}\mathbf{H}_\pi = \mathbf{H}_\pi'\mathbf{K}\mathbf{H}_\pi$, the minimization of $\operatorname{tr}(\tilde{\mathbf{S}}_W)$ is thus equivalent to that of $\operatorname{tr}(\mathbf{E}'\boldsymbol{\Pi}\mathbf{H}_\pi'\boldsymbol{\Delta}\mathbf{H}_\pi \cdot \boldsymbol{\Pi}\mathbf{E}(\mathbf{E}'\boldsymbol{\Pi}\mathbf{E})^{-1})$.

Recall that in the proof of Proposition 1, $\mathbf{L}$ is required to satisfy only the conditions $\mathbf{L} = \mathbf{L}'$ and $\mathbf{L}\mathbf{1}_n = \mathbf{0}$. Note that $\boldsymbol{\Pi}\mathbf{H}_\pi'\mathbf{K}\mathbf{H}_\pi\boldsymbol{\Pi}\mathbf{1}_n = \mathbf{0}$. Thus, if $\mathbf{Y}$ is an $n \times (c-1)$ matrix subject to the three conditions in Proposition 1, we have $T = \operatorname{tr}(\mathbf{Y}'\boldsymbol{\Pi}\mathbf{H}_\pi'\mathbf{K}\mathbf{H}_\pi\boldsymbol{\Pi}\mathbf{Y})$. This allows us to relax the maximization of T with respect to $\mathbf{E}$ as follows:

$$\begin{aligned}&\max_{\mathbf{Y}\in\mathbb{R}^{n\times(c-1)}} \operatorname{tr}(\mathbf{Y}'\boldsymbol{\Pi}\mathbf{H}_\pi'\mathbf{K}\mathbf{H}_\pi\boldsymbol{\Pi}\mathbf{Y}) \\ &\quad = \operatorname{tr}(\mathbf{Y}'\boldsymbol{\Pi}\mathbf{K}\boldsymbol{\Pi}\mathbf{Y}) \\ &\quad \text{s.t. } \mathbf{Y}'\boldsymbol{\Pi}\mathbf{Y} = \mathbf{I}_{c-1} \text{ and } \mathbf{Y}'\boldsymbol{\Pi}\mathbf{1}_n = \mathbf{0},\end{aligned} \tag{4.2}$$

where the second equality in the objective is due to the identity $\mathbf{Y}'\boldsymbol{\Pi}\mathbf{H}_\pi' = \mathbf{Y}'\boldsymbol{\Pi}$. Letting $\mathbf{Y}_0 = \boldsymbol{\Pi}^{1/2}\mathbf{Y}$ leads to

$$\begin{aligned}&\max_{\mathbf{Y}_0\in\mathbb{R}^{n\times(c-1)}} \operatorname{tr}(\mathbf{Y}_0'\boldsymbol{\Pi}^{1/2}\mathbf{H}_\pi'\mathbf{K}\mathbf{H}_\pi\boldsymbol{\Pi}^{1/2}\mathbf{Y}_0) \\ &\quad \text{s.t. } \mathbf{Y}_0'\mathbf{Y}_0 = \mathbf{I}_{c-1} \text{ and } \mathbf{Y}_0'\boldsymbol{\Pi}^{1/2}\mathbf{1}_n = \mathbf{0}.\end{aligned} \tag{4.3}$$

This optimization problem is solved in Appendix A.4. In particular, let $\mathbf{U}$ be an $n \times (c-1)$ matrix whose columns are the top $c-1$ eigenvectors of $\boldsymbol{\Pi}^{1/2}\mathbf{H}_\pi'\mathbf{K} \cdot \mathbf{H}_\pi\boldsymbol{\Pi}^{1/2}$. The solution of problem (4.3) is then $\mathbf{Y}_0 = \mathbf{U}\mathbf{Q}$ where $\mathbf{Q}$ is an arbitrary $(c-1) \times (c-1)$ orthonormal matrix. Hence, the solution of problem (4.2) is $\mathbf{Y} = \boldsymbol{\Pi}^{-1/2}\mathbf{U}\mathbf{Q}$.

4.2 Minimum Variance Formulations versus PCUT Formulations

Since the Laplacian matrix $\mathbf{L}$ is symmetric and positive semidefinite, its Moore–Penrose (MP) inverse is also positive semidefinite. Thus we can regard $\mathbf{L}$ as the MP inverse of a kernel matrix $\mathbf{K}$ and investigate the relationship between the spectral relaxations obtained from the minimum variance and the PCUT formulations. In fact, we have the following theorem, whose proof is given in Appendix A.5.

THEOREM 3. *Assume that* $\mathbf{L}^+ = \mathbf{K}$. *If* $\mathrm{rk}(\mathbf{L}) = \mathrm{rk}(\mathbf{K}) = n - 1$, *then* $\mathbf{Y}$ *is the solution of problem* (2.3) *if and only if it is the solution of problem* (4.2).

Thus, an equivalent formulation of spectral clustering based on the PCUT criterion is obtained by considering the minimum variance criterion with $\mathbf{K} = \mathbf{L}^+$. Note that $\mathbf{\Pi}$ consists of the diagonal elements of $\mathbf{K}^+$ in the NCUT setting, so it is not expedient computationally to obtain $\mathbf{\Pi}$ from $\mathbf{K}$—we would need to calculate $\mathbf{K}^+$. We thus suggest defining $\mathbf{\Pi} = \mathbf{I}_n$ in the minimum-variance setting, corresponding to the ratio cut formulation.

It is also possible to start from a minimum-variance formulation (with $\mathbf{\Pi} = \mathbf{I}_n$) and obtain a RCUT problem. However, in the corresponding RCUT problem, the matrix $\mathbf{K}^+$ is not guaranteed to be Laplacian, because the off-diagonal entries of $\mathbf{K}^+$ are possibly positive for an arbitrary kernel matrix $\mathbf{K}$. In this case, we can let $\mathbf{L} = \mathbf{K}^+ + n\beta\mathbf{H}_n$ where $\beta = \min\{\max_{i\neq j}\{[\mathbf{K}^+]_{ij}\}, 0\}$. Such an $\mathbf{L}$ is Laplacian. Moreover, we have $\mathrm{tr}(\mathbf{Y}'(\mathbf{K}^+ + n\beta\mathbf{H}_n)\mathbf{Y}) = \mathrm{tr}(\mathbf{Y}'\mathbf{K}^+\mathbf{Y}) + n(c-1)\beta$ due to $\mathbf{Y}'\mathbf{Y} = \mathbf{I}_{c-1}$ and $\mathbf{Y}' \cdot \mathbf{1}_n = \mathbf{0}$. Since $\min(\mathrm{tr}(\mathbf{Y}'(\mathbf{K}^+ + n\beta\mathbf{H}_n)\mathbf{Y}))$ is equivalent to $\min(\mathrm{tr}(\mathbf{Y}'\mathbf{K}^+\mathbf{Y}))$, it is not necessary to compute the value of β.

It is worth noting that the condition $\mathrm{rk}(\mathbf{L}) = \mathrm{rk}(\mathbf{K}) = n - 1$ is necessary. Without this condition, $\mathbf{\Pi}^{-1/2}\mathbf{L}\mathbf{\Pi}^{-1/2}$ is a generalized inverse of $\mathbf{\Pi}^{1/2}\mathbf{H}_\pi'\mathbf{L}^+\mathbf{H}_\pi\mathbf{\Pi}^{1/2}$, because

$$\mathbf{\Pi}^{1/2}\mathbf{H}_\pi'\mathbf{L}^+\mathbf{H}_\pi\mathbf{\Pi}^{1/2}\mathbf{\Pi}^{-1/2}\mathbf{L}\mathbf{\Pi}^{-1/2}\mathbf{\Pi}^{1/2}\mathbf{H}_\pi'\mathbf{L}^+ \cdot \mathbf{H}_\pi\mathbf{\Pi}^{1/2} = \mathbf{\Pi}^{1/2}\mathbf{H}_\pi'\mathbf{L}^+\mathbf{H}_\pi\mathbf{\Pi}^{1/2},$$

but it is not necessarily the MP inverse. In this case, it is no longer the case that $\mathbf{\Pi}^{-1/2}\mathbf{L}\mathbf{\Pi}^{-1/2}$ and $\mathbf{\Pi}^{1/2}\mathbf{H}_\pi'\mathbf{L}^+\mathbf{H}_\pi\mathbf{\Pi}^{1/2}$ are guaranteed to have the same eigenvectors associated with nonzero eigenvalues. Thus, in this case, even if $\mathbf{K} = \mathbf{L}^+$, the solutions of (4.2) and (2.3) are different. In summary we see that the spectral clustering formulations based on the minimum-variance criteria and PCUT, while closely related, are not fully equivalent.

Dhillon, Guan and Kulis (2007) pursue a slightly different connection between minimum-variance criteria and spectral relaxation. They formulate the minimum-variance criterion via the maximization of

$$T' = \mathrm{tr}(\mathbf{E}'\mathbf{\Pi}\mathbf{K}\mathbf{\Pi}\mathbf{E}(\mathbf{E}'\mathbf{\Pi}\mathbf{E})^{-1}), \tag{4.4}$$

which is readily shown to be equal to $T + \boldsymbol{\pi}'\mathbf{K} \cdot \boldsymbol{\pi}/(\boldsymbol{\pi}'\mathbf{1}_n)$, where T is defined by (4.1). Thus the maximization of T' is equivalent to the maximization of T. Dhillon, Guan and Kulis (2007) then formulate the cut minimization problem as an equivalent maximization problem:

$$\max\bigl(\mathbf{E}'\mathbf{\Pi}(\mathbf{\Pi}^{-1} - \mathbf{\Pi}^{-1}\mathbf{L}\mathbf{\Pi}^{-1})\mathbf{\Pi}\mathbf{E}(\mathbf{E}'\mathbf{\Pi}\mathbf{E})^{-1}\bigr),$$

and treat $\mathbf{\Pi}^{-1} - \mathbf{\Pi}^{-1}\mathbf{L}\mathbf{\Pi}^{-1}$ as $\mathbf{K}$ in T'. However, $\mathbf{\Pi}^{-1} - \mathbf{\Pi}^{-1}\mathbf{L}\mathbf{\Pi}^{-1}$ is generally indefinite, a difficulty that the authors circumvent by letting $\mathbf{K} = \rho\mathbf{I}_n - \mathbf{L}$ in RCUT and $\mathbf{K} = \rho\mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{W}\mathbf{D}^{-1}$ in NCUT, where ρ is a constant chosen to make $\mathbf{K}$ positive semidefinite.

The idea of considering a kernel matrix that is the MP inverse of a Laplacian matrix will return in later sections, in particular in Section 5.1 where we will see that it allows us to provide a geometrical interpretation for spectral clustering, and in Section 6, where we present a probabilistic interpretation of spectral relaxation.

5. SPECTRAL CLUSTERING: A MARGIN-BASED PERSPECTIVE

In this section we consider a margin-based perspective on spectral clustering. First, we show that the margin-based perspective provides us with insight into the relationship between spectral embedding and rounding. In particular, we show that the problems in (2.3) and (4.2) can be understood in terms of the fitting of hyperplanes in an RKHS. For a data point $\mathbf{x}$, we show that the elements of the embedding $\mathbf{y}$ are proportional to the signed distances of feature vector $\tilde{\mathbf{x}}$ to each of these hyperplanes. This provides support for the Procrustean rounding in which rounding is achieved by nonmaximum suppression of the elements of $\mathbf{y}$. Second, we provide some additional direct justification for the Procrustean approach, showing that the rounding problem can be analyzed in terms of the approximation of a margin-based multiway classification criterion.

5.1 Hyperplanes in the RKHS

Let us consider a multiway classification problem. That is, we consider a problem in which data points are pairs, $(\mathbf{x}_i, t_i)$, where t_i is the label of the ith data point. Using the same notation as in Section 4, the multiway classification problem has the following standard formulation in an RKHS based on a kernel function K:

$$\min_{\boldsymbol{\beta}_0, \mathbf{B}} \operatorname{tr}(\mathbf{B}'\mathbf{K}\mathbf{B}) + \frac{\gamma}{n}\sum_{i=1}^{n} f_{t_i}(\mathbf{B}'\mathbf{k}_i + \boldsymbol{\beta}_0), \tag{5.1}$$

where $f_j(\cdot)$ is a convex surrogate of the 0–1 loss, $\mathbf{k}_i = (K(\mathbf{x}_1, \mathbf{x}_i), \ldots, K(\mathbf{x}_n, \mathbf{x}_i))'$ is the ith column of the kernel matrix $\mathbf{K}$, $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_{c-1}]$ is an $n \times (c-1)$ matrix of regression vectors, $\boldsymbol{\beta}_0$ is a $(c-1) \times 1$ vector of intercepts and $\gamma > 0$ is a regularization parameter. We can use this optimization problem as the basis of a clustering formulation by simply omitting the term $\frac{\gamma}{n}\sum_{i=1}^{n} f_{t_i}(\cdot)$, reflecting the fact that we have no labeled data in the clustering setting. We obtain

$$\begin{aligned} &\min_{\mathbf{B}} \operatorname{tr}(\mathbf{B}'\mathbf{K}\mathbf{B}) \\ &\text{s.t. } \mathbf{B}'\mathbf{K}\boldsymbol{\Pi}\mathbf{1}_n = \mathbf{0} \text{ and } \mathbf{B}'\mathbf{K}\boldsymbol{\Pi}\mathbf{K}\mathbf{B} = \mathbf{I}_{c-1}. \end{aligned} \tag{5.2}$$

We now consider problem (5.2) from two points of view. From the first point of view, we let $\mathbf{Y} = \mathbf{K}\mathbf{B}$ and transform (5.2) into

$$\begin{aligned} &\min_{\mathbf{Y}} \operatorname{tr}(\mathbf{Y}'\mathbf{K}^{+}\mathbf{Y}) \\ &\text{s.t. } \mathbf{Y}'\boldsymbol{\Pi}\mathbf{1}_n = \mathbf{0} \text{ and } \mathbf{Y}'\boldsymbol{\Pi}\mathbf{Y} = \mathbf{I}_{c-1}, \end{aligned} \tag{5.3}$$

where we have used the identity $\mathbf{K} = \mathbf{K}\mathbf{K}^{+}\mathbf{K}$. It is readily seen that (5.3), and hence (5.2), is identical with the spectral relaxation in (2.3) by taking $\mathbf{K}^{+} = \mathbf{L}$. We also obtain a relationship between (5.3) and (4.2) from Section 4.2; in particular, in the special case in which $\operatorname{rk}(\mathbf{K}) = n - 1$, it follows from Theorem 3 that (5.3) and (4.2) are equivalent.

From a second point of view, we let $\mathbf{S} = \tilde{\mathbf{X}}'\mathbf{B}$ (recall that $\tilde{\mathbf{X}}$ is the data matrix in the feature space). The problem (5.2) is then transformed into

$$\begin{aligned} &\min_{\mathbf{S}} \operatorname{tr}(\mathbf{S}'\mathbf{S}) \\ &\text{s.t. } \mathbf{S}'\tilde{\mathbf{X}}'\boldsymbol{\Pi}\mathbf{1}_n = \mathbf{0} \text{ and } \mathbf{S}'\tilde{\mathbf{X}}'\boldsymbol{\Pi}\tilde{\mathbf{X}}\mathbf{S} = \mathbf{I}_{c-1}. \end{aligned} \tag{5.4}$$

Letting $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_{c-1}]$ denote the solution of (5.4), the equations $\mathbf{s}_j'\tilde{\mathbf{x}} = 0$, $j = 1, \ldots, c-1$, define hyperplanes that pass through the weighted centroid $\sum_{i=1}^{n} \pi_i \tilde{\mathbf{x}}_i$ of the feature vectors $\tilde{\mathbf{x}}_i$. Moreover, the signed distance between feature vector $\tilde{\mathbf{x}}_i$ and the hyperplane $\mathbf{s}_j'\tilde{\mathbf{x}} = 0$ is $\mathbf{s}_j'\tilde{\mathbf{x}}_i$. Recall that $\mathbf{Y} = [y_{ij}] = \mathbf{K}\mathbf{B} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{B} = \tilde{\mathbf{X}}\mathbf{S}$. We thus have $y_{ij} = \mathbf{s}_j'\tilde{\mathbf{x}}_i$. That is, y_{ij} is the signed distance of $\tilde{\mathbf{x}}_i$ to the jth hyperplane. We can therefore interpret the spectral relaxation in (2.3) and (4.2) as yielding vectors whose elements are—using the language of multiway classification—margin vectors. Given this interpretation, it is reasonable to allocate labels by finding the maximum element of $(y_{i1}, \ldots, y_{i,c-1}, 0)$. This motivates the Procrustean approach to rounding, which can be viewed as identifying boundaries between clusters by projecting feature vectors onto hyperplanes in an RKHS. A graphical interpretation of this result is provided in Figure 1.

5.2 Margin-Based Rounding Scheme

We can also provide a direct connection between classification and rounding. Let us return to the objective function in (5.1), which we rewrite as

$$\min_{\mathbf{Y}} \operatorname{tr}(\mathbf{Y}'\mathbf{K}^{+}\mathbf{Y}) + \frac{\gamma}{n}\sum_{i=1}^{n} f_{t_i}(\mathbf{y}_i)$$

by letting $\mathbf{Y} = \mathbf{K}\mathbf{B}$ and setting $\boldsymbol{\beta}_0 = 0$. Assume that we have obtained a matrix $\mathbf{Y}$ from spectral relaxation and recall that $\mathbf{Y}$ depends on an arbitrary orthogonal matrix $\mathbf{Q}$. From the classification perspective we can view the subsequent rounding problem as the problem of minimizing the classification loss $\frac{1}{n}\sum_{i=1}^{n} f_{t_i}(\mathbf{y}_i)$ under the constraint $\mathbf{Q}\mathbf{Q}' = \mathbf{I}_{c-1}$. In this section we explore some of the consequences of this perspective.

In the multiway classification problem, we define *class-conditional probabilities* $P_j(\mathbf{x})$ for the c classes $j = 1, \ldots, c$. Using this notation, we define the expected error at $\mathbf{x}$ as follows:

$$R(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{c} \mathbb{I}_{[t \neq j]} P_j(\mathbf{x}), \tag{5.5}$$

where $t = \arg\max_j y_j$ or $t = c$ if $\max\{y_j\} < 0$ and where $\mathbb{I}_{[\#]}$ defines the 0–1 loss: it is 1 if # is true and 0 otherwise. Since $\mathbb{I}_{[\cdot]}$ is a non-convex objective function that leads to an intractable optimization problem, the standard practice in the classification literature is to replace $\mathbb{I}_{[\cdot]}$ with a "surrogate loss function" $f_j(\mathbf{y})$ that is an upper bound on the 0–1 loss (Bartlett, Jordan and McAuliffe, 2006; Shen and Wang, 2007).

The surrogate loss function that we consider in the current paper is the following exponential loss:

$$f_j(\mathbf{y}) = \sum_{l \neq j} \exp(y_l - y_j), \tag{5.6}$$

where for convenience we extend $\mathbf{y}$ to a c-dimensional vector in which $y_c = 0$. Note that the variables to be optimized are the entries of the matrix $\mathbf{Q}$. Clearly, $f_j(\mathbf{y})$ is

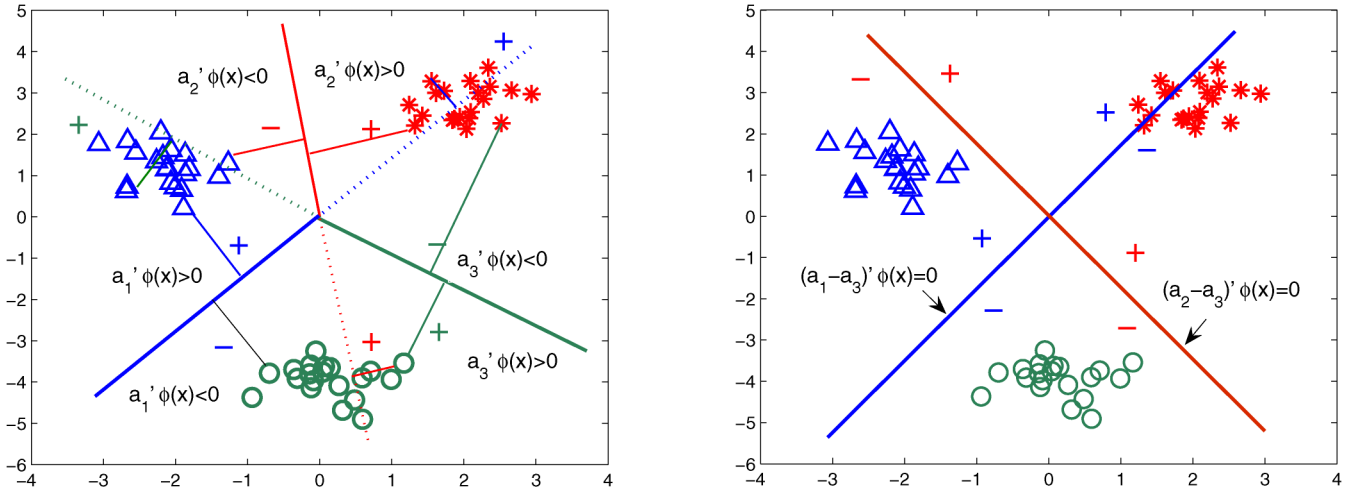

FIG. 1. *Illustrations of spectral clustering in the feature space for a three-class separable example. The clustering is based on the signed distances of the feature vector* $\tilde{\mathbf{x}} = \phi(\mathbf{x})$ *to suitably defined hyperplanes.* (a) *Hyperplanes in the feature space are represented by their normals,* $\mathbf{a}_j$, $j = 1, 2, 3$, *subject to the sum-to-zero constraints. These hyperplanes are computed from the vectors* $\mathbf{s}_1$ *and* $\mathbf{s}_2$ *obtained from spectral relaxation via* $\mathbf{a}_1 = \mathbf{s}_1 - \frac{1}{3}(\mathbf{s}_1 + \mathbf{s}_2)$, $\mathbf{a}_2 = \mathbf{s}_2 - \frac{1}{3}(\mathbf{s}_1 + \mathbf{s}_2)$ *and* $\mathbf{a}_3 = -\frac{1}{3}(\mathbf{s}_1 + \mathbf{s}_2)$. (b) *The hyperplanes defined by the vectors* $\mathbf{s}_1$ *and* $\mathbf{s}_2$. *Note that* $\mathbf{s}_1 = \mathbf{a}_1 - \mathbf{a}_3$ *and* $\mathbf{s}_2 = \mathbf{a}_2 - \mathbf{a}_3$.

an upper bound of $\mathbb{I}_{[t \neq j]}$, because if $\mathbf{x}$ does not belong to class j, there exists at least one y_l such that $l \neq j$ and $y_l - y_j \geq 0$, and hence $\exp(y_l - y_j) \geq 1$. This surrogate loss function also has an important Fisher consistency property:

PROPOSITION 3. *Assume* $P_j(\mathbf{x}) > 0$ *for* $j = 1, \ldots, c$. *We then have*

$$\hat{y}_j = \arg\max_{\mathbf{y}} \sum_{j=1}^{c} \sum_{l \neq j} \exp(y_l - y_j) P_j(\mathbf{x}) = \frac{1}{2} \log \frac{P_j(\mathbf{x})}{P_c(\mathbf{x})}.$$

The proof of Proposition 3 is a straightforward calculation, so we omit it. This proposition shows that the surrogate loss function that we have chosen is justified from the point of view of classification as yielding a Bayes consistent rule (Bartlett, Jordan and McAuliffe, 2006; Zou, Zhu and Hastie, 2006).

Returning to the rounding problem, we now consider the labels $\{t_i\}$ as temporarily fixed and consider the empirical risk function defined over the set of pairs $(\mathbf{x}_i, t_i)$ given by

$$J(\mathbf{Q}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{l \neq t_i} \exp(y_{il} - y_{it_i}).$$

We wish to optimize this empirical risk with respect to $\mathbf{Q}$. This problem does not have a closed-form solution under the constraint $\mathbf{Q}\mathbf{Q}' = \mathbf{I}_{c-1}$. However, we can consider a Taylor expansion around $y_{ij} = 0$. We have

$$J(\mathbf{Q}) \approx (c-1) - \frac{c}{n} \sum_{i=1}^{n} \mathbf{g}_{t_i}' \mathbf{y}_i + c^2 \sum_{i=1}^{n} \pi_i^{-1},$$

where $\mathbf{g}_j$ is the jth column of $\mathbf{G}' = [\mathbf{I}_{c-1} - \frac{1}{c}\mathbf{1}_{c-1}\mathbf{1}_{c-1}', -\frac{1}{c}\mathbf{1}_{c-1}]$, and where we have used the fact that $\mathbf{y}_i' \mathbf{g}_{t_i} \cdot \mathbf{g}_{t_i}' \mathbf{y}_i = \pi_i^{-1} \boldsymbol{\mu}_i' \mathbf{Q} \mathbf{g}_{t_i} \mathbf{g}_{t_i}' \mathbf{Q}' \boldsymbol{\mu}_i \leq 1/\pi_i$ because $\mathbf{I}_{c-1} - \mathbf{g}_{t_i} \mathbf{g}_{t_i}'$ is positive semidefinite. We thus see that the maximization of the linear term $\sum_{i=1}^{n} \mathbf{g}_{t_i}' \mathbf{y}_i$ with respect to $\mathbf{Q}$ yields an approximate procedure for minimizing $J(\mathbf{Q})$. But this is precisely the Procrustean problem (3.1) discussed in Section 3.

It would also be possible to attempt to optimize $J(Q)$ directly by making use of Newton or conjugate gradient methods on the Stiefel manifold (Edelman, Arias and Smith, 1999).

6. SPECTRAL RELAXATION: THE VIEW FROM GAUSSIAN INTRINSIC AUTOREGRESSION

In this section we show that spectral relaxation can be interpreted as a model-based statistical procedure. In particular, we present a connection between spectral relaxation and Gaussian intrinsic autoregression models.

Our focus is the spectral relaxation problem presented in Section 2, specifically the constrained eigenvalue problem in (2.3).

Recall that the Laplacian matrix $\mathbf{L}$ is a positive semidefinite matrix; moreover, the pseudoinverse $\mathbf{L}^+$ is positive semidefinite and can be viewed as a kernel matrix. We found this perspective useful in our discussion of minimum-variance clustering in Section 4.2; note also that (Saerens et al., 2004) have explored connections between spectral embedding and random walks on graphs using the fact that the elements of $\mathbf{L}^+$ are closely related to the commute-time distances obtained from a random walk on the graph. In this section, we take the interpretation of $\mathbf{L}^+$ in a different direction, using it to make the connection to Gaussian intrinsic autoregressions.

Denote $\mathbf{K} = \mathbf{L}^+$ where $\mathbf{L} = \mathbf{D} - \mathbf{W}$. Let us model the $n \times (c-1)$ matrix $\mathbf{Y}$ as a singular matrix-variate normal distribution $N_{n,c-1}(\mathbf{0}, \sigma^2\mathbf{K} \otimes \mathbf{I}_{c-1})$ where we follow the notation for matrix-variate normal distributions in (Gupta and Nagar, 2000). That is,

$$p(\mathbf{Y}) \propto \exp\left(-\frac{1}{2\sigma^2}\operatorname{tr}(\mathbf{Y}'\mathbf{L}\mathbf{Y})\right).$$

Let us set $\sigma^2 = 1/\operatorname{tr}(\mathbf{\Pi}\mathbf{K})$ so that $\mathsf{E}(\mathbf{Y}'\mathbf{\Pi}\mathbf{Y}) = \sigma^2 \operatorname{tr}(\mathbf{\Pi}\mathbf{K})\mathbf{I}_{c-1} = \mathbf{I}_{c-1}$. Finally, we impose the constraint $\mathbf{Y}'\mathbf{\Pi}\mathbf{1}_n = \mathbf{0}$ in order to remove the redundancy $\mathbf{K}^+\mathbf{1}_n = \mathbf{0}$ in $\mathbf{K}^+$. We thus obtain the following proposition.

PROPOSITION 4. *The relaxation problem in* (2.3) *is equivalent to the maximization of the log likelihood* $p(\mathbf{Y})$ *under the constraints* $\mathbf{Y}'\mathbf{\Pi}\mathbf{Y} = \mathbf{I}_{c-1}$ *and* $\mathbf{Y}'\mathbf{\Pi}\mathbf{1}_n = \mathbf{0}$.

We obtain a statistical interpretation of spectral relaxation from the fact that a multivariate normal distribution can be equivalently expressed as a Gaussian conditional autoregression (CAR) (Besag 1974; Mardia 1988). Indeed, given $\mathbf{Y} \sim N_{n,c-1}(\mathbf{0}, \sigma^2\mathbf{K} \otimes \mathbf{I}_{c-1})$, we have that the $\mathbf{y}_i$ can be characterized as $(c-1)$-dimensional CARs with

$$\begin{aligned}\mathsf{E}(\mathbf{y}_i|\mathbf{y}_j, j \neq i) &= -\sum_{j\neq i}\frac{l_{ij}}{l_{ii}}\mathbf{y}_j = \sum_{j=1}^{n}\frac{w_{ij}}{l_{ii}}\mathbf{y}_j,\\ \mathsf{Var}(\mathbf{y}_i|\mathbf{y}_j, j \neq i) &= \frac{\sigma^2}{l_{ii}}\mathbf{I}_{c-1}.\end{aligned} \tag{6.1}$$

That is, we have $\mathbf{y}_i|\{\mathbf{y}_j : j \neq i\} \sim N_{c-1}(\sum_{j=1}^n \frac{w_{ij}}{l_{ii}}\mathbf{y}_j, \frac{\sigma^2}{l_{ii}}\mathbf{I}_{c-1})$, for $i = 1, \ldots, n$. Since $\mathbf{K}$ is positive semidefinite but not positive definite, Besag and Kooperberg (1995) referred to such conditional autoregressions as *Gaussian intrinsic autoregressions*.

The CAR model implicitly requires $w_{ii} = 0$ and $l_{ii} = \sum_{j=1}^n w_{ij}$. In spectral embedding and clustering (Guattery and Miller, 2000; Belkin and Niyogi, 2002; Ng, Jordan and Weiss, 2002), the w_{ij} are usually used to assert adjacency or similarity relationships between the $\mathbf{y}_i$. We will see shortly that these adjacency or similarity relationships have an interpretation as conditional independencies.

Since $\mathbf{D} - \mathbf{W}$ is positive semidefinite, $\mathbf{D} - \omega\mathbf{W}$ is positive definite for $\omega \in (0, 1)$. This fact has been used to devise CAR models based on $\mathbf{D} - \omega\mathbf{W}$ such that $\mathsf{E}(\mathbf{y}_i|\mathbf{y}_j, j \neq i) = \omega\sum_{j=1}^n \frac{w_{ij}}{l_{ii}}\mathbf{y}_j$ (see, e.g., Carlin and Banerjee, 2003). We now have

$$\mathsf{E}(\mathbf{y}_i\mathbf{y}_j'|\mathbf{y}_l, l \neq i, j) = \frac{\omega l_{ij}}{\omega^2 l_{ij}^2 - l_{ii}l_{jj}}\sigma^2\mathbf{I}_{c-1}.$$

As a result, $l_{ij} = 0$ (or $w_{ij} = 0$) implies that $\mathbf{y}_i \perp\!\!\!\perp \mathbf{y}_j|\{\mathbf{y}_l : l \neq i, j\}$; that is, $\mathbf{y}_i$ is conditionally independent of $\mathbf{y}_j$ given the remaining vectors. This Markov property also holds for Gaussian intrinsic autoregressions (Besag and Kooperberg, 1995).

This perspective sheds light on some of the relationships between the NCUT and RCUT formulations of spectral relaxation. Recall that since $\mathbf{\Pi} = \mathbf{D}$ in the NCUT setting, we impose the constraints $\mathbf{Y}'\mathbf{D}\mathbf{Y} = \mathbf{I}_{c-1}$ and $\mathbf{Y}'\mathbf{D}\mathbf{1}_n = \mathbf{0}$. On the other hand, the RCUT formulation uses the constraints $\mathbf{Y}'\mathbf{Y} = \mathbf{I}_{c-1}$ and $\mathbf{Y}'\mathbf{1}_n = \mathbf{0}$ because $\mathbf{\Pi} = \mathbf{I}_n$. Theorem 1 shows that the solution of the NCUT is based on $\mathbf{\Pi}^{-1/2}\mathbf{L}\mathbf{\Pi}^{-1/2} = \mathbf{I}_n - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$, which is a so-called normalized graph Laplacian. The solution of the RCUT problem is based on the unnormalized graph Laplacian $\mathbf{L}$. Now Proposition 1 reveals a problematic aspect of the NCUT formulation—piecewise constancy of the columns of $\mathbf{Y}$ is accompanied by a lack of orthogonality of these columns. Two natural desiderata of spectral clustering are in conflict in the NCUT formulation. This conflict between orthogonality and piecewise constancy is not present for RCUT. However, the existing empirical results showed that the normalized graph Laplacian tends to outperform the unnormalized graph Laplacian. Moreover, von Luxburg, Belkin and Bousquet (2008) provided theoretical evidence of the superiority of the normalized graph Laplacian.

This seeming paradox can be resolved by using an alternative choice for $\mathbf{L}$ in the RCUT formulation. Let us set $\mathbf{L} = (\mathbf{I}_n - \mathbf{C})'(\mathbf{I}_n - \mathbf{C})$, where $\mathbf{C} = [c_{ij}]$ is an $n \times n$ nonnegative matrix such that $c_{ii} = 0$ for all i and $\mathbf{C}\mathbf{1}_n = \mathbf{1}_n$. Such a $\mathbf{L}$ is positive semidefinite but no longer Laplacian. Since $\mathbf{L}\mathbf{1}_n = \mathbf{0}$, we can still solve the spectral relaxation problem (2.4) using Theorem 1.

Our experimental results in Section 7 show that this novel RCUT formulation is very effective. It is

also worth noting that we can connect this formulation to the simultaneous autoregression (SAR) model of Besag (1974). In particular, the $\mathbf{y}_i$ are now specified by n simultaneous equations:

$$\mathbf{y}_i = \sum_{j=1}^{n} c_{ij}\mathbf{y}_j + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n,$$

where the $\boldsymbol{\varepsilon}_i$ are independent normal vectors from $N_{c-1}(\mathbf{0}, \sigma^2\mathbf{I}_{c-1})$. This equation can be written in matrix form as follows:

$$\mathbf{Y} = \mathbf{C}\mathbf{Y} + \boldsymbol{\Sigma}$$

$$\text{with} \quad \boldsymbol{\Sigma} = [\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n]' \sim N_{n,c-1}(\mathbf{0}, \sigma^2\mathbf{I}_n \otimes \mathbf{I}_{c-1}).$$

We thus have $\mathbf{Y} \sim N_{n,c-1}(\mathbf{0}, \sigma^2\mathbf{K} \otimes \mathbf{I}_{c-1})$ with $\mathbf{K}^+ = (\mathbf{I}_n - \mathbf{C})'(\mathbf{I}_n - \mathbf{C})$. In practice, we are especially concerned with the case in which $\mathbf{C} = \mathbf{D}^{-1}\mathbf{W}$. It is worth noting that $\mathbf{I}_n - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ and $\mathbf{I}_n - \mathbf{D}^{-1}\mathbf{W}$ have the same eigenvalues, while the squared singular values of $\mathbf{I}_n - \mathbf{D}^{-1}\mathbf{W}$ are the eigenvalues of $(\mathbf{I}_n - \mathbf{D}^{-1}\mathbf{W})'(\mathbf{I}_n - \mathbf{D}^{-1}\mathbf{W})$. We thus obtain an interesting new relationship between the NCUT formulation and the RCUT formulation.

7. EXPERIMENTS

Although our principal focus has been to provide a unifying perspective on spectral clustering, our analysis has also provided novel spectral algorithms, and it is of interest to compare the performance of these algorithms to existing algorithms. In this section we report the results of experiments conducted with six publicly available data sets: five data sets from the UCI machine learning repository (the *dermatology* data, the *vowel* data, the *NIST* optical handwritten digit data, the *letter* data and the *image segmentation* data) as well as a set of *gene expression* data analyzed by Yeung et al. (2001).

In the dermatology data, there are 366 patients, 8 of whom are excluded due to missing information, with 34 features. The data are clustered into six classes. We standardized the data to have zero mean and unit variance. The NIST data set contains the handwritten digits 0–9, where each instance consists of a 16×16 pixel and where digits are treated as classes. We selected 1000 digits, with 100 instances per digit, for our experiments. The vowel data set contains the eleven steady-state vowels of British English. The letter data set consists of images of the letters "A" to "Z." In our experiments we selected the first 10 letters with 195, 199, 182, 207, 203, 210, 226, 196, 188 and 172 instances, respectively. The image segmentation data consist of seven types of images: "brickface," "sky," "foliage," "cement," "window," "path" and "grass." The gene data set contains 384 genes with 17 time points over two cell cycles. The data were standardized to have mean zero and unit variance (Yeung et al., 2001). We treated the five phases of the cell cycle as five nominal classes for these data, classifying genes into these classes according to their expression level peaks. Table 1 gives a summary of these data sets.

TABLE 1
Summary of the benchmark data sets

	Gene	Dermatology	Vowel	NIST	Letter	Segmentation
n	384	358	990	1000	1978	2100
d	17	34	10	256	16	19
c	5	6	11	10	10	7

n—the number of samples; d—the number of features; c—the number of classes.

We compared our rounding algorithm based on Procrustean transformation (see Algorithm 1) with those based on the rounding procedures given in Bach and Jordan (2006) and Yu and Shi (2003), conducting comparisons using the NCUT, RCUT and minimum-variance criteria. We refer to the weighted K-means and the K-means algorithms of Bach and Jordan (2006) as *BJ-wkm* and *BJ-km*, respectively. Note that the spectral clustering algorithm based on the NCUT formulation and K-means rounding is equivalent to that presented by Ng, Jordan and Weiss (2002). We initialized the K-means algorithms by the orthogonal initialization method in Ng, Jordan and Weiss (2002). For the rounding scheme of Yu and Shi (2003), we used two initialization methods: the orthogonal initialization method and initialization to the identity matrix. We refer to the corresponding algorithms as *YS-1* and *YS-2*. We also used these two initialization methods in our algorithm (Algorithm 1), referring to the results in these two cases as *Margin-1* and *Margin-2*.

7.1 Setup and Evaluation Criterion

We defined the adjacency matrix $\mathbf{W} = [w_{ij}]$ as $w_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\beta)$ with $\beta > 0$. The kernel matrix is defined as $\mathbf{K} = \mathbf{H}_n\mathbf{W}\mathbf{H}_n$. For the margin-based algorithms, however, we set $w_{ii} = 0$ for $i = 1, \ldots, n$; in this case the kernel matrix is defined as $\mathbf{K} = \mathbf{H}_n(\mathbf{I}_n + \mathbf{W})\mathbf{H}_n$. For simplicity, we do not distinguish between these two cases in our notation in the remainder of this section. In the minimum-variance formulation we always set $\boldsymbol{\Pi} = \mathbf{I}_n$. With these settings, the *BJ-wkm* and

BJ-km algorithms are based on the spectral decomposition of $\mathbf{I}_n - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$. The *YS-1* and *YS-2* algorithms are based on the spectral decomposition of $\mathbf{I}_n - \mathbf{D}^{-1}\mathbf{W}$, and the *Margin-1* and *Margin-2* algorithms are based on the spectral decomposition of $\mathbf{I}_n - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$.

Although $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is one natural choice in the RCUT setting, we instead adopted the suggestion in Section 6 and defined $\mathbf{L}$ as

$$\mathbf{L} = (\mathbf{I}_n - \mathbf{D}^{-1}\mathbf{W})'(\mathbf{I}_n - \mathbf{D}^{-1}\mathbf{W}). \tag{7.1}$$

To simplify the comparison among procedures, we fixed β to specific sets of values for each of the data sets, exploring a range of values to investigate the relative sensitivities to the choice of β for the different clustering algorithms. Our specific choices for both the NCUT and RCUT criteria were $\beta \in \{1, 10\}$ for the gene data, $\beta \in \{1, 10, 100\}$ for the "vowel" data, $\beta \in \{5000, 10000, 20000\}$ for the "image segmentation" data, and $\beta \in \{10, 100, 1000\}$ for the "dermatology," "NIST" and "letter" data sets. Since the minimum-variance criterion directly operates on $\mathbf{K}$, we choose a different set of values when working with this criterion; in particular, we used $\beta \in \{10, 100\}$ for the gene data, $\beta \in \{100, 1000\}$ for the "dermatology" data, $\beta \in \{1, 10, 100\}$ for the "vowel" data, $\beta \in \{500, 1000\}$ for NIST data, $\beta \in \{10, 100, 1000\}$ for the "letter" data, and $\beta \in \{10, 100, 1000\}$ for the "image segmentation" data.

To evaluate the performance of the various clustering algorithms we employed the Rand index (RI) (Rand, 1971). Given a set of n objects $S = \{O_1, \ldots, O_n\}$, suppose that $U = \{U_1, \ldots, U_r\}$ and $V = \{V_1, \ldots, V_s\}$ are two different partitions of the objects in S such that $\bigcup_{i=1}^{r} U_i = S = \bigcup_{j=1}^{s} V_j$ and $U_i \cap U_{i'} = \varnothing = V_j \cap V_{j'}$ for $i \neq i'$ and $j \neq j'$. Let a be the number of pairs of objects that are in the same set in U and in the same set in V, and b the number of pairs of objects that are in different sets in U and in different sets in V. The Rand index is given by $\mathrm{RI} = (a + b)/\binom{n}{2}$. If $\mathrm{RI} = 1$, the two partitions are identical.

Since the ground-truth partitions are available for our six data sets, we directly calculated RI between the true partition and the partition obtained from each clustering algorithm. We conducted 50 replicates of each of the algorithms that require random initialization (this is not necessary for *YS-2* and *Margin-2*, which are initialized to the identity matrix). Note that for the RCUT and minimum-variance criteria, *BJ-wkm* and *BJ-km* become identical because in these cases $\mathbf{\Pi} = \mathbf{I}_n$.

7.2 Performance Analysis

Figure 2 displays the results for all six algorithms using the NCUT criterion. We see that the margin-based algorithms are competitive with the other algorithms. The poorest performer in this setting is *BJ-wkm*, which is highly sensitive to the value of β. In particular, when $\beta = 10$ for the "gene" data set, $\beta \in \{10, 100\}$ for the vowel data, $\beta \in \{1000, 100, 10\}$ for the "letter" data, and $\beta = 1000$ for both the "dermatology" and "NIST" data sets, this algorithm almost failed. A possible interpretation for this result is the conflict between orthogonality and piecewise constancy implied in the NCUT setting (see Proposition 1). Indeed, as can be seen from Figure 2, the situation is more favorable for the *BJ-km* rounding algorithm; in this case $\mathbf{D}^{-12}\mathbf{Y}(\mathbf{Y}'\mathbf{D}^{-1}\mathbf{Y})^{-1/2}$ is used, which diminishes the conflict between orthogonality and piecewise constancy. Similarly, the conflict is diminished for the *YS* rounding algorithms and our margin-based rounding methods (because $\arg\max_j d_j^{-1/2} y_{ij}$ is equivalent to $\arg\max_j y_{ij}$).

Recall that the *YS-1* and *YS-2* algorithms need to use a heuristic post-processing procedure; that is, the algorithms operate on $\hat{\mathbf{Z}} = \mathrm{dg}(\mathbf{Z}\mathbf{Z}')^{-1/2}\mathbf{Z}$. We found that the performance of the algorithms depends strongly on this procedure.

Figures 3 and 4 display the experimental results using the RCUT and minimum-variance criteria, respectively. We see again that the margin-based algorithms are competitive with the other algorithms; indeed for several of the data sets the margin-based algorithms yield better performance than the other algorithms.

We see from Figures 3 and 4 that *BJ-km* is competitive with the other algorithms. This shows that the choice of $\mathbf{L}$ given in (7.1) is an effective choice.

We again found it to be the case that the heuristic post-processing procedure was needed for *YS-1* and *YS-2* to yield good clustering performance.

The performances of *Margin-1* and *Margin-2* were similar across the data sets and criteria, showing the relative insensitivity of the margin-based approach to the initialization. Note in particular the larger degree of variability between the performances of *YS-1* and *YS-2*. Note also that the margin-based approach was in general less sensitive to the value of β than the other algorithms.

Finally, recall that $\mathbf{L}$ in (7.1) for the RCUT setting and $\mathbf{L} = \mathbf{K}^{+}$ obtained from the minimum-variance setting are positive semidefinite but they are not Laplacian matrices, because the off-diagonal elements of the

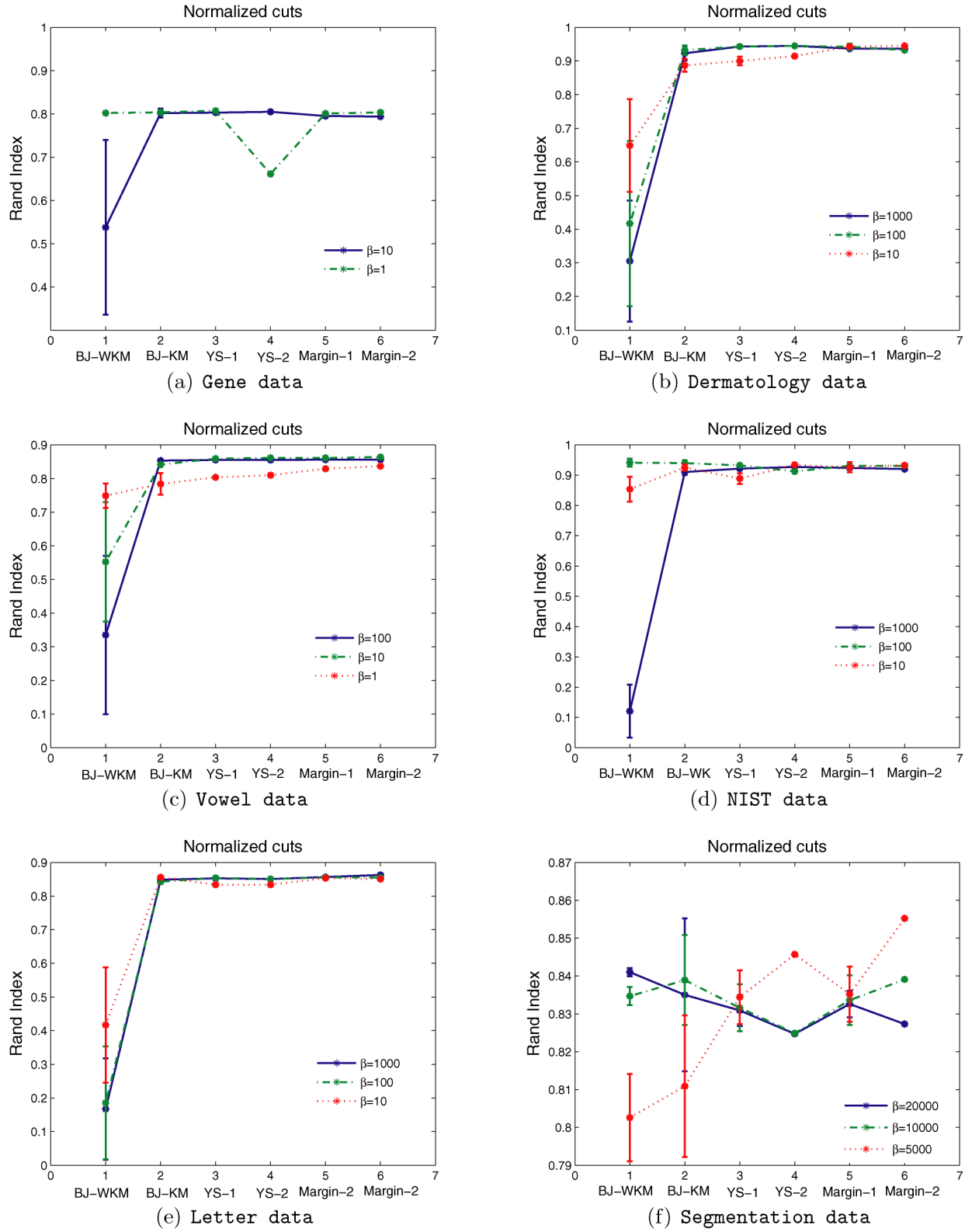

FIG. 2. *Clustering results (Rand index) with normalized cuts. "BJ-WKM": the weighted K-means rounding of Bach and Jordan 2006; "BJ-KM": the K-means rounding of Bach and Jordan 2006; "YS-1": the rounding scheme of Yu and Shi 2003 with the orthogonal initialization method; "YS-2": the rounding scheme of Yu and Shi 2003 with initialization to the identity matrix; "Margin-1": the rounding scheme in Section 3.1 with the orthogonal initialization method; "Margin-2": the rounding scheme in Section 3.1 with initialization to the identity matrix.*

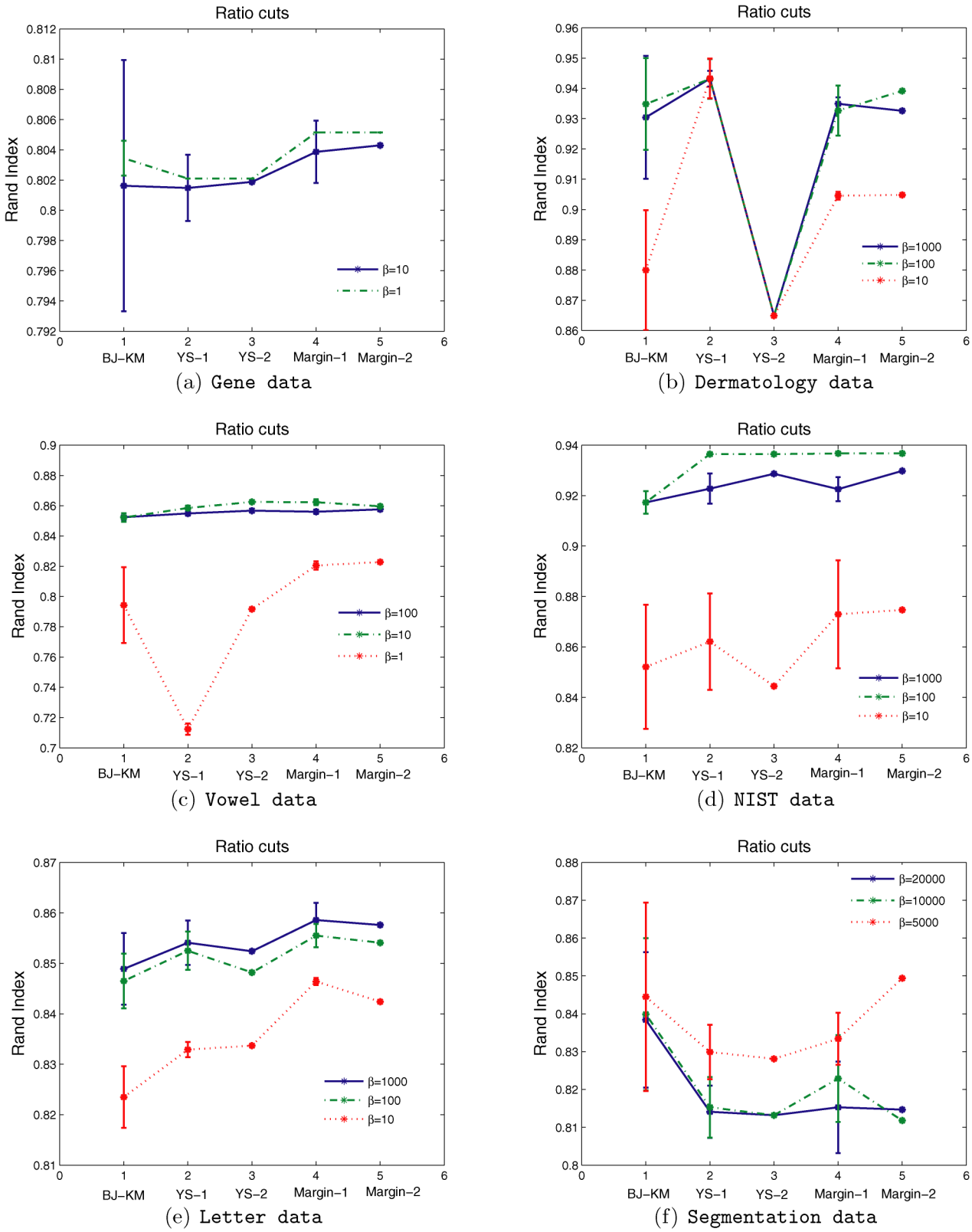

FIG. 3. *Clustering results (Rand index) with ratio cuts. See the caption of Figure 2 for explanation of the acronyms.*

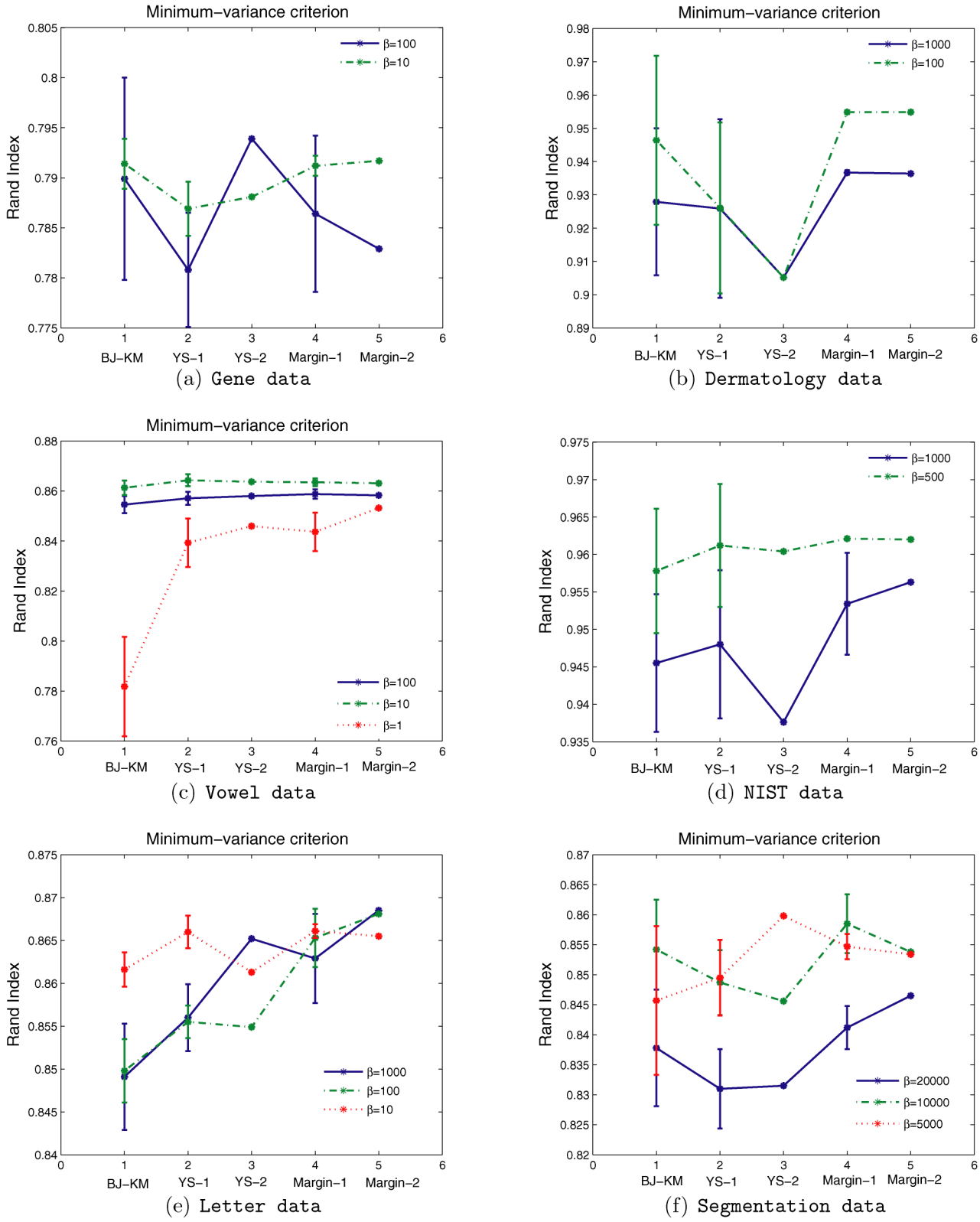

FIG. 4. *Clustering results (Rand index) with the minimum-variance criterion. See the caption of Figure 2 for explanation of the acronyms.*

$\mathbf{W} = \mathbf{L} - \mathbf{D}$ are possibly negative. Nonetheless, our experimental results showed that these two choices are still effective. Thus cuts can be defined through non-Laplacian matrices. Although such cuts lose their original interpretation in terms of the graph partition, as we have shown they do have a clear statistical interpretation in terms of Gaussian intrinsic autoregression models.

8. DISCUSSION

In this paper we have presented a margin-based perspective on multiway spectral clustering. We have shown that both aspects of spectral clustering—relaxation and rounding—can be given an interpretation in terms of margins. The major advantage of this perspective is that it ties spectral clustering to the large literature on margin-based classification. The margin-based perspective has several additional consequences: (1) it permits a deeper understanding of the relationship between the normalized cut and ratio cut formulations of spectral clustering; (2) it strengthens the connections between the minimum-variance criterion and spectral clustering; and (3) it yields a statistical interpretation of spectral clustering in terms of Gaussian intrinsic autoregressions. Also, the preliminary empirical evidence that we presented suggests that the algorithms motivated by the margin-based perspective are competitive with existing spectral clustering algorithms.

One of the most useful consequences of the margin-based perspective is the interpretation that it yields of spectral clustering in terms of projection onto hyperplanes in a reproducing kernel Hilbert space (see Figure 1). This interpretation shows that the performance of the margin-based clustering algorithms depends on the separability of the feature vectors. This suggests that the algorithmic problem of choosing the similarity matrix $\mathbf{W}$ or kernel matrix $\mathbf{K}$ so as to increase separability is an important topic for further research; see Bach and Jordan (2006) and Meilă and Shi (2000) for initial work along these lines.

Although we have focused on undirected graphs in our treatment, it is also worth noting the possibility of considering clustering in a directed graph with the asymmetric weighted matrix $\mathbf{D}^{-1}\mathbf{W}$ (Meilă and Pentney, 2007). This can be related to our discussion in Section 6, where we suggested the use of the matrix $\mathbf{L} = (\mathbf{I}_n - \mathbf{D}^{-1}\mathbf{W})'(\mathbf{I}_n - \mathbf{D}^{-1}\mathbf{W})$ in the RCUT setting. The experimental results in Section 7 showed that such a suggestion is promising. Moreover, although $\mathbf{L}$ is no longer Laplacian, the corresponding spectral relaxation can be interpreted as a simultaneous autoregression model. The relationship between simultaneous autoregression and conditional autoregression (Ripley, 1981) may provide connections between spectral clustering in undirected graphs and directed graphs. We intend to explore this issue in future work.

In delineating a relationship between the PCUT criterion and the kernel minimum-variance criterion, we have proven that the relaxation problems (2.3) and (4.2) have the same solution whenever $\mathrm{rk}(\mathbf{L}) = n - 1$ and $\mathbf{L}^+ = \mathbf{K}$. This leads to the question as to whether the original unrelaxed problems—that is, the minimization of PCUT and the maximization of T with respect to discrete partition matrix $\mathbf{E}$— have the same solution under the conditions $\mathrm{rk}(L) = n - 1$ and $\mathbf{L}^+ = \mathbf{K}$. This is currently an open problem.

APPENDIX

A.1 Proof of Proposition 1

Since the columns of $\mathbf{Y}$ are piecewise constant with respect to the partition $\mathbf{E}$, we can express $\mathbf{Y}$ as $\mathbf{Y} = \mathbf{E}\boldsymbol{\Psi}$ for some $\boldsymbol{\Psi} \in \mathbb{R}^{c\times(c-1)}$. Let $\mathbf{Y}_0 = \boldsymbol{\Pi}^{1/2}\mathbf{Y}$, $\boldsymbol{\Psi}_0 = [\boldsymbol{\Psi}, \alpha\mathbf{1}_c]$, a $c \times c$ matrix, and $\mathbf{Z} = [\mathbf{Y}_0, \alpha\boldsymbol{\Pi}^{1/2}\mathbf{1}_n]$, where $\alpha = 1/\sqrt{\mathbf{1}_n'\boldsymbol{\Pi}\mathbf{1}_n}$. We have $\boldsymbol{\Pi}^{-1/2}\mathbf{Z} = \mathbf{E}\boldsymbol{\Psi}_0$ and $\mathbf{Z}'\mathbf{Z} = [\mathbf{Y}_0, \alpha\boldsymbol{\Pi}^{1/2}\mathbf{1}_n]'[\mathbf{Y}_0, \alpha\boldsymbol{\Pi}^{1/2}\mathbf{1}_n] = \mathbf{I}_c$ due to $\mathbf{E}\mathbf{1}_c = \mathbf{1}_n$, $\mathbf{Y}_0'\mathbf{Y}_0 = \mathbf{Y}'\boldsymbol{\Pi}\mathbf{Y} = \mathbf{I}_{c-1}$ and $\mathbf{Y}_0'\boldsymbol{\Pi}^{1/2}\mathbf{1}_n = \mathbf{Y}'\boldsymbol{\Pi}\mathbf{1}_n = \mathbf{0}$. Furthermore, we have $\boldsymbol{\Psi}_0'\mathbf{E}'\boldsymbol{\Pi}\mathbf{E}\boldsymbol{\Psi}_0 = \mathbf{Z}'\mathbf{Z} = \mathbf{I}_c$. Since $\boldsymbol{\Psi}_0$ and $\mathbf{E}'\boldsymbol{\Pi}\mathbf{E}$ are square, $\boldsymbol{\Psi}_0$ and $\mathbf{E}'\boldsymbol{\Pi}\mathbf{E}$ are invertible. Hence $\boldsymbol{\Psi}_0\boldsymbol{\Psi}_0' = (\mathbf{E}'\boldsymbol{\Pi}\mathbf{E})^{-1}$. We now have

$$\begin{aligned}\mathrm{tr}(\mathbf{Y}'\mathbf{L}\mathbf{Y}) &= \mathrm{tr}(\mathbf{Y}_0'\boldsymbol{\Pi}^{-1/2}\mathbf{L}\boldsymbol{\Pi}^{-1/2}\mathbf{Y}_0)\\ &= \mathrm{tr}(\mathbf{Z}'\boldsymbol{\Pi}^{-1/2}\mathbf{L}\boldsymbol{\Pi}^{-1/2}\mathbf{Z}) = \mathrm{tr}(\boldsymbol{\Psi}_0'\mathbf{E}'\mathbf{L}\mathbf{E}\boldsymbol{\Psi}_0)\\ &= \mathrm{tr}(\mathbf{E}'\mathbf{L}\mathbf{E}\boldsymbol{\Psi}_0\boldsymbol{\Psi}_0') = \mathrm{tr}(\mathbf{E}'\mathbf{L}\mathbf{E}(\mathbf{E}'\boldsymbol{\Pi}\mathbf{E})^{-1}),\end{aligned}$$

completing the proof.

A.2 The Proof of Proposition 2

In this section we provide a constructive proof of Proposition 2 by establishing the existence of $\boldsymbol{\Psi}$. We also provide an example of the construction in the special case of $c = 4$ and $\boldsymbol{\Pi} = \mathbf{I}_n$.

Let $(\mathbf{E}'\boldsymbol{\Pi}\mathbf{E})^{-1} = \mathrm{diag}(1/\beta_1, \ldots, 1/\beta_c)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_c)'$. We then have $\mathbf{1}_n'\boldsymbol{\Pi}\mathbf{1}_n = \boldsymbol{\pi}'\mathbf{1}_n = \boldsymbol{\beta}'\mathbf{1}_c$ and $\mathbf{E}'\boldsymbol{\Pi}\mathbf{1}_n = \boldsymbol{\beta}$. In the proof in Appendix A.1, we obtain $\boldsymbol{\Psi}_0\boldsymbol{\Psi}_0' = (\mathbf{E}'\boldsymbol{\Pi}\mathbf{E})^{-1}$. Thus,

$$\boldsymbol{\Psi}\boldsymbol{\Psi}' = \mathrm{diag}(1/\beta_1, \ldots, 1/\beta_c) - \frac{1}{\boldsymbol{\pi}'\mathbf{1}_n}\mathbf{1}_c\mathbf{1}_c' \quad \text{(denoted } \mathbf{A}\text{)}.$$

In order to make the above equation hold, it is necessary for $\mathbf{A}$ to be positive semidefinite. Given any nonzero $\mathbf{b} = (b_1, \ldots, b_c)' \in \mathbb{R}^c$, we have

$$\mathbf{b}' \operatorname{diag}(\boldsymbol{\beta}) \mathbf{A} \operatorname{diag}(\boldsymbol{\beta}) \mathbf{b} / (\boldsymbol{\pi}' \mathbf{1}_n) = \sum_{j=1}^{c} \frac{\beta_j}{\boldsymbol{\pi}' \mathbf{1}_n} b_j^2 - \left(\sum_{j=1} \frac{\beta_j}{\boldsymbol{\pi}' \mathbf{1}_n} b_j \right)^2 \geq 0,$$

since the function $f(x) = x^2$ is convex. This implies that $\mathbf{A}$ positive semidefinite. Furthermore, it is easy to obtain $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$. Using the SVD of $\mathbf{A}$, we are always able to obtain a $\boldsymbol{\Psi}$ such that $\boldsymbol{\Psi}\boldsymbol{\Psi}' = \mathbf{A}$ and $\boldsymbol{\Psi}'\boldsymbol{\beta} = \mathbf{0}$. Consequently, we have

$$\mathbf{1}_n' \boldsymbol{\Pi} \mathbf{E} \boldsymbol{\Psi} = \boldsymbol{\beta}' \boldsymbol{\Psi} = 0 \quad \text{and} \quad \boldsymbol{\Psi}' \mathbf{E}' \boldsymbol{\Pi} \mathbf{E} \boldsymbol{\Psi} = \mathbf{I}_{c-1}.$$

The latter equality comes from

$$\mathbf{I}_c = \boldsymbol{\Psi}_0' \mathbf{E}' \boldsymbol{\Pi} \mathbf{E} \boldsymbol{\Psi}_0 = \begin{bmatrix} \boldsymbol{\Psi}' \\ \alpha \mathbf{1}_c' \end{bmatrix} \mathbf{E}' \boldsymbol{\Pi} \mathbf{E} [\boldsymbol{\Psi}, \alpha \mathbf{1}_c] = \begin{bmatrix} \boldsymbol{\Psi}' \mathbf{E}' \boldsymbol{\Pi} \mathbf{E} \boldsymbol{\Psi} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}.$$

EXAMPLE 1. Let $\eta = \boldsymbol{\pi}' \mathbf{1}_n$ and $\eta_j = \sum_{i \in V_j} \pi_i$. Assume that $\boldsymbol{\Psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{c-1})'$ where $\boldsymbol{\psi}_1' = (\frac{\sqrt{\eta - \eta_1}}{\sqrt{\eta \eta_1}}, -\frac{\sqrt{\eta_1}}{\sqrt{\eta(\eta - \eta_1)}} \mathbf{1}_{c-1}')$ and

$$\boldsymbol{\psi}_l' = \left(0 * \mathbf{1}_{l-1}', \frac{\sqrt{\sum_{j=l+1}^{c} \eta_j}}{\sqrt{\eta_l \sum_{j=l}^{c} \eta_j}}, \frac{\sqrt{\eta_l}}{\sqrt{\sum_{j=l}^{c} \eta_j \sum_{j=l+1}^{c} \eta_j}} \mathbf{1}_{c-l} \right)$$

for $l = 2, \ldots, c-1$. For instance, if $c = 4$, we have

$$\boldsymbol{\Psi} = \begin{bmatrix} \frac{\sqrt{\eta - \eta_1}}{\sqrt{\eta \eta_1}} & 0 & 0 \\ -\frac{\sqrt{\eta_1}}{\sqrt{\eta(\eta - \eta_1)}} & \frac{\sqrt{\eta_3 + \eta_4}}{\sqrt{\eta_2(\eta - \eta_1)}} & 0 \\ -\frac{\sqrt{\eta_1}}{\sqrt{\eta(\eta - \eta_1)}} & -\frac{\sqrt{\eta_2}}{\sqrt{(\eta_3 + \eta_4)(\eta - \eta_1)}} & \frac{\sqrt{\eta_4}}{\sqrt{(\eta_3 + \eta_4)\eta_3}} \\ -\frac{\sqrt{\eta_1}}{\sqrt{\eta(\eta - \eta_1)}} & -\frac{\sqrt{\eta_2}}{\sqrt{(\eta_3 + \eta_4)(\eta - \eta_1)}} & -\frac{\sqrt{\eta_3}}{\sqrt{(\eta_3 + \eta_4)\eta_4}} \end{bmatrix}.$$

It is easily verified that $\mathbf{Y} = \mathbf{E}\boldsymbol{\Psi}$ satisfies the conditions (a)–(c) listed in Proposition 1. Let $\mathbf{a}_1, \ldots, \mathbf{a}_c$ denote the row vectors of $\boldsymbol{\Psi}$. We note that an arbitrary collection of $c-1$ vectors from the set $\mathbf{a}_1, \ldots, \mathbf{a}_c$ are linearly independent. The convex hull of $\mathbf{a}_1, \ldots, \mathbf{a}_c$ is thus a $(c-1)$-dimensional simplex. (A d-dimensional simplex is the convex hull of an affinely independent point set in $\mathbb{R}^d$. A regular d-dimensional simplex is the convex hull of $d+1$ points with all pairs of points having equal distances.) In addition, we have that the squared distance between $\mathbf{a}_i$ and $\mathbf{a}_j$ is

$$\|\mathbf{a}_i - \mathbf{a}_j\|^2 = \frac{1}{\eta_i} + \frac{1}{\eta_j} \quad \text{for } i \neq j.$$

Note that we have $\eta = n$ and $\eta_j = n_j$ when $\boldsymbol{\Pi} = \mathbf{I}_n$. In particular, if $\boldsymbol{\Pi} = \mathbf{I}_n$ and $n_1 = \cdots = n_c = \frac{n}{c}$, the $\mathbf{a}_i$ constitute the vertices of a $(c-1)$-dimensional regular simplex.

A.3 The Proof of Theorem 1

This theorem is a variation on a standard result in linear algebra; for completeness we present a proof. Let $\mathbf{S} = \boldsymbol{\Pi}^{-1/2} \mathbf{L} \boldsymbol{\Pi}^{-1/2}$ and consider the following Lagrangian:

$$L(\mathbf{Y}_0, \mathbf{A}, \mathbf{b}) = \operatorname{tr}(\mathbf{Y}_0' \mathbf{S} \mathbf{Y}_0) - \operatorname{tr}(\mathbf{A}(\mathbf{Y}_0' \mathbf{Y}_0 - \mathbf{I}_{c-1})) - \mathbf{b}' \mathbf{Y}_0' \boldsymbol{\Pi}^{1/2} \mathbf{1}_n,$$

where $\mathbf{A}$ is a $(c-1) \times (c-1)$ symmetric matrix of Lagrange multipliers and $\mathbf{b}$ is a $(c-1) \times 1$ vector of Lagrange multipliers. We differentiate to obtain

$$\frac{\partial L}{\partial \mathbf{Y}_0} = 2\mathbf{S}\mathbf{Y}_0 - 2\mathbf{Y}_0\mathbf{A} - \boldsymbol{\Pi}^{1/2} \mathbf{1}_n \mathbf{b}'.$$

Letting $\frac{\partial L}{\partial \mathbf{Y}_0} = \mathbf{0}$ leads to

$$2\mathbf{S}\mathbf{Y}_0 - 2\mathbf{Y}_0\mathbf{A} - \boldsymbol{\Pi}^{1/2} \mathbf{1}_n \mathbf{b}' = \mathbf{0},$$

from which we have

$$2\mathbf{1}_n' \boldsymbol{\Pi}^{1/2} \mathbf{S} \mathbf{Y}_0 - 2\mathbf{1}_n' \boldsymbol{\Pi}^{1/2} \mathbf{Y}_0 \mathbf{A} - \mathbf{1}_n' \boldsymbol{\Pi} \mathbf{1}_n \mathbf{b}' = \mathbf{0}.$$

This implies $\mathbf{b} = \mathbf{0}$. Accordingly, we obtain

$$\mathbf{S}\mathbf{Y}_0 = \mathbf{Y}_0\mathbf{A}.$$

We now take the eigendecomposition of $\mathbf{A}$, letting $\mathbf{A} = \mathbf{Q}' \boldsymbol{\Gamma}_1 \mathbf{Q}$ where $\mathbf{Q}$ is a $(c-1) \times (c-1)$ orthonormal matrix and $\boldsymbol{\Gamma}_1$ is a $(c-1) \times (c-1)$ diagonal matrix. We note that the diagonal entries of $\boldsymbol{\Gamma}_1$ and the columns of $\mathbf{Y}_0\mathbf{Q}'$ are the eigenvalues and the associated eigenvectors of $\mathbf{S}$. Clearly, $\boldsymbol{\Pi}^{1/2}\mathbf{1}_n$ is the eigenvector of $\mathbf{S}$ associated with eigenvalue 0. We now let $\boldsymbol{\Gamma}_1 = \operatorname{diag}(\gamma_2, \ldots, \gamma_c)$. We thus have $\bar{\mathbf{Y}}_0 = [\boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_c]\mathbf{Q}$. Obviously, $\bar{\mathbf{Y}}_0$ satisfies $\bar{\mathbf{Y}}_0' \bar{\mathbf{Y}}_0 = \mathbf{I}_{c-1}$ and $\bar{\mathbf{Y}}_0' \boldsymbol{\Pi}^{1/2} \mathbf{1}_n = \mathbf{0}$ due to $\boldsymbol{\mu}_i' \boldsymbol{\Pi}^{1/2} \mathbf{1}_n = 0$ for $i \neq 1$.

To verify that $\bar{\mathbf{Y}}_0$ is the solution of problem (2.4), we consider the Hessian matrix of L with respect to $\mathbf{Y}_0$. Let $\operatorname{vec}(\mathbf{Y}_0') = (y_{11}, \ldots, y_{1,c-1}, y_{21}, \ldots, y_{n,c-1})'$. The Hessian matrix is then given by

$$\mathbf{H}(\mathbf{Y}_0) = \frac{\partial^2 L}{\partial \operatorname{vec}(\mathbf{Y}_0') \, \partial \operatorname{vec}(\mathbf{Y}_0')'} = \mathbf{I}_{c-1} \otimes \mathbf{S} - \mathbf{A} \otimes \mathbf{I}_n.$$

Let $\mathbf{B}$ be an arbitrary nonzero $n \times (c-1)$ matrix such that $\mathbf{B}'[\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_c] = \mathbf{0}$. We can always express $\mathbf{B} = [\boldsymbol{\mu}_{c+1}, \ldots, \boldsymbol{\mu}_n]\boldsymbol{\Phi}$ where $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{c-1}]$ is an $(n-c) \times (c-1)$ matrix. Denoting $\boldsymbol{\Gamma}_2 = \operatorname{diag}(\gamma_{c+1}, \ldots, \gamma_n)$, we have

$$\begin{aligned}
&\operatorname{vec}((\mathbf{BQ})')'\mathbf{H}(\bar{\mathbf{Y}}_0)\operatorname{vec}((\mathbf{BQ})') \\
&\quad = \operatorname{tr}(\mathbf{Q}'\mathbf{B}'\mathbf{SBQ}) - \operatorname{tr}(\mathbf{AQ}'\mathbf{B}'\mathbf{BQ}) \\
&\quad = \operatorname{tr}(\mathbf{B}'\mathbf{SB}) - \operatorname{tr}(\boldsymbol{\Gamma}_1\mathbf{B}'\mathbf{B}) = \operatorname{tr}(\boldsymbol{\Phi}'\boldsymbol{\Gamma}_2\boldsymbol{\Phi}) - \operatorname{tr}(\boldsymbol{\Gamma}_1\boldsymbol{\Phi}'\boldsymbol{\Phi}) \\
&\quad = \sum_{i=1}^{c-1} \boldsymbol{\phi}_i'\boldsymbol{\Gamma}_2\boldsymbol{\phi}_i - \sum_{i=1}^{c-1} \gamma_{i+1}\boldsymbol{\phi}_i'\boldsymbol{\phi}_i \\
&\quad = \sum_{i=1}^{c-1} \boldsymbol{\phi}_i'(\boldsymbol{\Gamma}_2 - \gamma_{i+1}\mathbf{I}_{n-c})\boldsymbol{\phi}_i \geq 0.
\end{aligned}$$

If $\gamma_c > \gamma_{c+1}$, then the matrices $\boldsymbol{\Gamma}_2 - \gamma_{i+1}\mathbf{I}_{n-c}$, $i = 1, \ldots, c-1$, are positive definite. Thus, the above inequality is strict. This shows that $\bar{\mathbf{Y}}_0$ is a strict local minimum of $\operatorname{tr}(\mathbf{Y}_0'\boldsymbol{\Pi}^{-1/2}\mathbf{L}\boldsymbol{\Pi}^{-1/2}\mathbf{Y}_0)$ under the conditions $\mathbf{Y}_0'\mathbf{Y}_0 = \mathbf{I}_{c-1}$ and $\mathbf{Y}_0'\boldsymbol{\Pi}^{1/2}\mathbf{1}_n = \mathbf{0}$.

A.4 The Solution of Problem (4.3)

Let $\mathbf{T} = \boldsymbol{\Pi}^{1/2}\mathbf{H}_\pi'\mathbf{K}\mathbf{H}_\pi\boldsymbol{\Pi}^{1/2}$ and consider the following Lagrangian:

$$\begin{aligned}
&L(\mathbf{Y}_0, \mathbf{A}, \mathbf{b}) \\
&\quad = \operatorname{tr}(\mathbf{Y}_0'\mathbf{T}\mathbf{Y}_0) - \operatorname{tr}\big(\mathbf{A}(\mathbf{Y}_0'\mathbf{Y}_0 - \mathbf{I}_{c-1})\big) - \mathbf{b}'\mathbf{Y}_0'\boldsymbol{\Pi}^{1/2}\mathbf{1}_n,
\end{aligned}$$

where $\mathbf{A}$ is a $(c-1) \times (c-1)$ symmetric matrix of Lagrange multipliers and $\mathbf{b}$ is a $(c-1) \times 1$ vector of Lagrange multipliers. Differentiating, we obtain

$$\frac{\partial L}{\partial \mathbf{Y}_0} = 2\mathbf{T}\mathbf{Y}_0 - 2\mathbf{Y}_0\mathbf{A} - \boldsymbol{\Pi}^{1/2}\mathbf{1}_n\mathbf{b}'.$$

Letting $\frac{\partial L}{\partial \mathbf{Y}_0} = \mathbf{0}$ leads to

$$2\mathbf{T}\mathbf{Y}_0 - 2\mathbf{Y}_0\mathbf{A} - \boldsymbol{\Pi}^{1/2}\mathbf{1}_n\mathbf{b}' = \mathbf{0},$$

from which we have

$$2\mathbf{1}_n'\boldsymbol{\Pi}^{1/2}\mathbf{T}\mathbf{Y}_0 - 2\mathbf{1}_n'\boldsymbol{\Pi}^{1/2}\mathbf{Y}_0\mathbf{A} - \mathbf{1}_n'\boldsymbol{\Pi}\mathbf{1}_n\mathbf{b}' = \mathbf{0}.$$

Since $\mathbf{1}_n'\boldsymbol{\Pi}^{1/2}\mathbf{T} = \mathbf{1}_n'\boldsymbol{\Pi}\mathbf{H}_\pi'\mathbf{K}\mathbf{H}_\pi\boldsymbol{\Pi}^{1/2} = \mathbf{1}_n'\mathbf{H}_\pi\boldsymbol{\Pi}\mathbf{K} \cdot \mathbf{H}_\pi\boldsymbol{\Pi}^{1/2} = \mathbf{0}$, we obtain $\mathbf{b} = \mathbf{0}$. This implies

$$\mathbf{T}\mathbf{Y}_0 = \mathbf{Y}_0\mathbf{A}.$$

Now following the proof in Appendix A.3, we find that the top $c-1$ eigenvectors of $\mathbf{T}$ provide the solution for $\mathbf{Y}_0$ in problem (4.3).

A.5 The Proof of Theorem 3

Our proof is based on the following lemma.

LEMMA 1. *Assume that* $\mathbf{A}$ *is an* $n \times n$ *symmetric matrix with* $\operatorname{rk}(\mathbf{A}) = n-1$ *and* $\mathbf{A}\mathbf{1}_n = \mathbf{0}$. *Let* $\mathbf{A}^+$ *be the MP inverse of* $\mathbf{A}$. *Then* $\boldsymbol{\Pi}^{1/2}\mathbf{H}_\pi'\mathbf{A}^+\mathbf{H}_\pi\boldsymbol{\Pi}^{1/2}$ *is the MP inverse of* $\boldsymbol{\Pi}^{-1/2}\mathbf{A}\boldsymbol{\Pi}^{-1/2}$.

PROOF. We first prove $\mathbf{A}^+\mathbf{A} = \mathbf{A}\mathbf{A}^+ = \mathbf{H}_n$. Let $\mathbf{N} = \mathbf{A}'\mathbf{A}$. It is clear that $\mathbf{N}\mathbf{H}_n = \mathbf{H}_n\mathbf{N} = \mathbf{N}$. It thus follows from Corollary 4.5.18 in Horn and Johnson (1985) that there exists an $n \times n$ orthonormal matrix $\mathbf{U}$ such that

$$\mathbf{U}'\mathbf{N}\mathbf{U} = \begin{pmatrix} \boldsymbol{\Delta}_{n-1} & \mathbf{0} \\ \mathbf{0} & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{U}'\mathbf{H}_n\mathbf{U} = \begin{pmatrix} \mathbf{I}_{n-1} & \mathbf{0} \\ \mathbf{0} & 0 \end{pmatrix},$$

where $\boldsymbol{\Delta}_{n-1}$ is an $(n-1) \times (n-1)$ diagonal matrix with positive diagonal entries, and $\mathbf{U} = [\mathbf{U}_1, \frac{1}{\sqrt{n}}\mathbf{1}_n]$ with $\mathbf{U}_1'\mathbf{U}_1 = \mathbf{I}_{n-1}$ and $\mathbf{U}_1\mathbf{1}_n = \mathbf{0}$. Here we use the fact that $\mathbf{1}_n$ is the eigenvector of $\mathbf{N}$ and of $\mathbf{H}_n$ with associated eigenvalue 0. Accordingly, we have

$$\mathbf{N} = \mathbf{U}_1\boldsymbol{\Delta}_{n-1}\mathbf{U}_1' \quad \text{and} \quad \mathbf{H}_n = \mathbf{U}_1\mathbf{U}_1',$$

from which it follows that

$$\mathbf{N}^+ = \mathbf{U}_1\boldsymbol{\Delta}_{n-1}^{-1}\mathbf{U}_1'$$

and hence $\mathbf{N}^+\mathbf{N} = \mathbf{U}_1\mathbf{U}_1' = \mathbf{H}_n$. On the other hand, since $\mathbf{A}^+ = (\mathbf{A}'\mathbf{A})^+\mathbf{A}'$, we have $\mathbf{A}^+\mathbf{A} = \mathbf{N}^+\mathbf{N} = \mathbf{H}_n$. Since $\mathbf{A}$ is symmetric, we also have $\mathbf{A}\mathbf{A}^+ = \mathbf{H}_n$.

Using the identity $\mathbf{A}^+\mathbf{A} = \mathbf{A}\mathbf{A}^+ = \mathbf{H}_n$ and $\mathbf{A}\mathbf{H}_\pi' = \mathbf{A} = \mathbf{H}_\pi\mathbf{A}$, we have

$$\begin{aligned}
&\boldsymbol{\Pi}^{-1/2}\mathbf{A}\boldsymbol{\Pi}^{-1/2}\boldsymbol{\Pi}^{1/2}\mathbf{H}_\pi'\mathbf{A}^+\mathbf{H}_\pi\boldsymbol{\Pi}^{1/2} \\
&\quad = \boldsymbol{\Pi}^{-1/2}\mathbf{H}_\pi\boldsymbol{\Pi}^{1/2} = \boldsymbol{\Pi}^{1/2}\mathbf{H}_\pi'\boldsymbol{\Pi}^{-1/2} \\
&\quad = \boldsymbol{\Pi}^{1/2}\mathbf{H}_\pi'\mathbf{A}^+\mathbf{H}_\pi\boldsymbol{\Pi}^{1/2}\boldsymbol{\Pi}^{-1/2}\mathbf{A}\boldsymbol{\Pi}^{-1/2}.
\end{aligned}$$

We further obtain

$$\begin{aligned}
&\boldsymbol{\Pi}^{-1/2}\mathbf{A}\boldsymbol{\Pi}^{-1/2}\boldsymbol{\Pi}^{1/2}\mathbf{H}_\pi'\mathbf{A}^+\mathbf{H}_\pi\boldsymbol{\Pi}^{1/2}\boldsymbol{\Pi}^{-1/2}\mathbf{A}\boldsymbol{\Pi}^{-1/2} \\
&\quad = \boldsymbol{\Pi}^{-1/2}\mathbf{A}\boldsymbol{\Pi}^{-1/2}
\end{aligned}$$

and

$$\begin{aligned}
&\boldsymbol{\Pi}^{1/2}\mathbf{H}_\pi'\mathbf{A}^+\mathbf{H}_\pi\boldsymbol{\Pi}^{1/2}\boldsymbol{\Pi}^{-1/2}\mathbf{A}\boldsymbol{\Pi}^{-1/2}\boldsymbol{\Pi}^{1/2}\mathbf{H}_\pi'\mathbf{A}^+ \\
&\quad \cdot \mathbf{H}_\pi\boldsymbol{\Pi}^{1/2} = \boldsymbol{\Pi}^{1/2}\mathbf{H}_\pi'\mathbf{A}^+\mathbf{H}_\pi\boldsymbol{\Pi}^{1/2}.
\end{aligned}$$

Thus $\boldsymbol{\Pi}^{1/2}\mathbf{H}_\pi'\mathbf{A}^+\mathbf{H}_\pi\boldsymbol{\Pi}^{1/2}$ is the MP inverse of $\boldsymbol{\Pi}^{-1/2} \cdot \mathbf{A}\boldsymbol{\Pi}^{-1/2}$. □

Since $\mathbf{L}^+$ is the MP inverse of $\mathbf{L}$, $\mathbf{L}^+$ is positive semidefinite and it satisfies $\mathbf{L}^+\mathbf{1}_n = \mathbf{0}$ and $\operatorname{rk}(\mathbf{L}^+) = n-1$. It is obvious that $\operatorname{rk}(\boldsymbol{\Pi}^{-1/2}\mathbf{L}\boldsymbol{\Pi}^{-1/2}) = n-1$ and $\operatorname{rk}(\boldsymbol{\Pi}^{1/2}\mathbf{H}_\pi'\mathbf{L}^+\mathbf{H}_\pi\boldsymbol{\Pi}^{1/2}) = n-1$. Moreover, $\boldsymbol{\Pi}^{1/2}\mathbf{1}_n$ is

eigenvector of both $\mathbf{\Pi}^{-1/2}\mathbf{L}\mathbf{\Pi}^{-1/2}$ and $\mathbf{\Pi}^{1/2}\mathbf{H}_\pi'\mathbf{L}^+\mathbf{H}_\pi\mathbf{\Pi}^{1/2}$ with associated eigenvalue 0. In addition, if $\lambda \neq 0$ is eigenvalue of $\mathbf{\Pi}^{-1/2}\mathbf{L}\mathbf{\Pi}^{-1/2}$ with associated eigenvector $\mathbf{u}$, then λ^{-1} is eigenvalue of $\mathbf{\Pi}^{1/2}\mathbf{H}_\pi'\mathbf{L}^+\mathbf{H}_\pi\mathbf{\Pi}^{1/2}$ with associated eigenvector $\mathbf{u}$. It thus follows from Lemma 1 that (4.3) has the same solution as (2.4) whenever $\mathbf{L}^+ = \mathbf{K}$. As a result, (4.2) has the same solution as (2.3).

REFERENCES

BACH, F. R. and JORDAN, M. I. (2006). Learning spectral clustering, with application to speech separation. *J. Mach. Learn. Res.* **7** 1963–2001. MR2274430

BARTLETT, P. L., JORDAN, M. I. and MCAULIFFE, J. D. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.* **101** 138–156. MR2268032

BELKIN, M. and NIYOGI, P. (2002). Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems* **14** 585–592. MIT Press, Cambridge, MA.

BESAG, J. (1974). Spatial interaction and statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **36** 192–236. MR0373208

BESAG, J. and KOOPERBERG, C. (1995). On conditional and intrinsic autoregression. *Biometrika* **82** 733–746. MR1380811

CARLIN, B. P. and BANERJEE, S. (2003). Hierarchical multivariate CAR models for spatio-temporally correlated survival data (with discussion). In *Bayesian Statistics* **7** 45–63. Oxford Univ. Press. MR2003166

CHAN, P. K., SCHLAG, M. D. F. and ZIEN, J. Y. (1994). Spectral K-way ratio-cut partitioning and clustering. *IEEE Trans. Computer-Aided Design Integrated Circuits Syst.* **13** 1088–1096.

CHUNG, F. R. (1997). *Spectral Graph Theory*. Amer. Math. Soc., Providence, RI. MR1421568

DHILLON, I. S., GUAN, Y. and KULIS, B. (2007). Weighted graph cuts without eigenvectors: A multilevel approach. *IEEE Trans. Pattern Anal. Mach. Intell.* **29** 1944–1957.

DING, C., HE, X. and SIMON, H. D. (2005). On the equivalence of nonnegative matrix factorization and spectral clustering. In *SIAM Conference on Data Mining (SDM)*. Newport Beach, CA.

DONATH, W. E. and HOFMANN, A. J. (1973). Lower bounds for the partitioning of graphs. *IBM J. Res. Develop.* **17** 420–425. MR0329965

EDELMAN, A., ARIAS, T. A. and SMITH, S. T. (1999). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20** 303–353. MR1646856

FIEDLER, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Math. J.* **23** 298–305. MR0318007

GOWER, J. C. and DIJKSTERHUIS, G. B. (2004). *Procrustes Problems*. Oxford Univ. Press. MR2051013

GUATTERY, S. and MILLER, G. L. (2000). Graph embeddings and Laplacian eigenvalues. *SIAM J. Matrix Anal. Appl.* **21** 703–723. MR1740868

GUPTA, A. K. and NAGAR, D. K. (2000). *Matrix Variate Distributions*. Chapman and Hall, London. MR1738933

HORN, R. A. and JOHNSON, C. R. (1985). *Matrix Analysis*. Cambridge Univ. Press. MR0832183

JUHÁSZ, F. and MÁLYUSZ, K. (1977). Problems of cluster analysis from the viewpoint of numerical analysis. In *Numerical Methods, Colloquia Mathematica Societatis Janos Bolyai* (P. Rózsa, ed.) **22** 405–415. North-Holland, Amsterdam. MR0590916

KANNAN, R., VEMPALA, S. and VETTA, A. (2000). On clusterings: Good, bad, and spectral. In *Proceedings of the 41st Annual Symposium on the Foundation of Computer Science* 367–380. MR1931834

MARDIA, K. V. (1988). Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. *J. Multivariate Anal.* **24** 265–284. MR0926357

MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press, New York. MR0560319

MEILĂ, M. and PENTNEY, W. (2007). Clustering by weighted cuts in directed graphs. In *SIAM Conference on Data Mining (SDM)*. Minneapolis, MN.

MEILĂ, M. and SHI, J. (2000). Learning segmentation by random walks. In *Advances in Neural Information Processing* **12** 470–477. MIT Press, Cambridge, MA.

MOHAR, B. (1991). *The Laplacian Spectrum of Graphs* 871–898. Wiley, New York. MR1170831

NG, A. Y., JORDAN, M. I. and WEISS, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* **14** 849–856. MIT Press, Cambridge, MA.

RAHIMI, A. and RECHT, B. (2004). Clustering with normalized cuts is clustering with a hyperplane. In *Workshop on Statistical Learning in Computer Vision*. Prague, Czech Republic.

RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* **66** 846–850.

RIPLEY, B. D. (1981). *Spatial Statistics*. Wiley, New York. MR0624436

SAERENS, M., FOUSS, F., YEN, L. and DUPONT, P. (2004). The principal components analysis of a graph, and its relationships to spectral clustering. In *The 15th European Conference on Machine Learning (ECML)* 371–383.

SHEN, X. and WANG, L. (2007). Generalization error for multi-class margin classification. *Electron. J. Statist.* **1** 307–330. MR2336036

SHI, J. and MALIK, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 888–905.

SHORTREED, S. and MEILĂ, M. (2005). Unsupervised spectral learning. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence* 534–541. AUAI Press.

VON LUXBURG, U., BELKIN, M. and BOUSQUET, O. (2008). Consistency of spectral clustering. *Ann. Statist.* **36** 555–586. MR2396807

VON LUXBURG, U. (2007). A tutorial on spectral clustering. *Statist. Comput.* **17** 395–416. MR2409803

WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia. MR1045442

WEBB, A. R. (2002). *Statistical Pattern Recognition*, 2nd ed. Wiley, Hoboken. MR2191640

WEISS, Y. (1999). Segmentation using eigenvectors: A unifying view. In *IEEE International Conference on Computer Vision* 975–982. IEEE Computer Society.

YEUNG, K. Y., FRALEY, C., MURUA, A., RAFTERY, A. E. and RUZZO, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinform.* **17** 977–987.

YU, S. X. and SHI, J. (2003). Multiclass spectral clustering. In *IEEE International Conference on Computer Vision* 313–319. IEEE Computer Society.

ZHA, H., DING, C., GU, M., HE, X. and SIMON, H. (2002). Spectral relaxation for k-means clustering. In *Advances in Neural Information Processing Systems* **14** 1057–1064. MIT Press, Cambridge, MA.

ZOU, H., ZHU, J. and HASTIE, T. (2006). The margin vector, admissible loss and multi-class margin-based classifiers. Technical report, Dept. Statistics, Stanford Univ.