MERL – A MITSUBISHI ELECTRIC RESEARCH LABORATORY
http://www.merl.com

A unifying theorem for spectral embedding and clustering

Matthew Brand Kun Huang*

TR-2002-42 November 2002

Abstract

Spectral methods use selected eigenvectors of a data affinity matrix to obtain a data representation that can be trivially clustered or embedded in a low-dimensional space. We present a theorem that explains, for broad classes of affinity matrices and eigenbases, why this works: For successively smaller eigenbases (i.e., using fewer and fewer of the affinity matrix's dominant eigenvalues and eigenvectors), the angles between similar vectors in the new representation shrink while the angles between dissimilar vectors grow. Specifically, the sum of the squared cosines of the angles is strictly increasing as the dimensionality of the representation decreases. Thus spectral methods work because the truncated eigenbasis amplifies structure in the data so that any heuristic post-processing is more likely to succeed. We use this result to construct a nonlinear dimensionality reduction (NLDR) algorithm for data sampled from manifolds whose intrinsic coordinate system has linear and cyclic axes, and a novel clustering-by-projections algorithm that requires no post-processing and gives superior performance on challenge problems from the recent literature.

Also presented at NIPS'02 workshop on spectral methods.

*UIUC EECS

Proceedings, 9th International Conference on Artificial Intelligence and Statistics, AISTATS, Key West, Florida

A unifying theorem for spectral embedding and clustering

Matthew Brand and **Kun Huang**
Mitsubishi Electric Research Labs, Cambridge, Massachusetts, USA
Electrical & Computer Engineering, University of Illinois at Urbana-Champagne, USA

Abstract

Spectral methods use selected eigenvectors of a data affinity matrix to obtain a data representation that can be trivially clustered or embedded in a low-dimensional space. We present a theorem that explains, for broad classes of affinity matrices and eigenbases, why this works: For successively smaller eigenbases (i.e., using fewer and fewer of the affinity matrix's dominant eigenvalues and eigenvectors), the angles between "similar" vectors in the new representation shrink while the angles between "dissimilar" vectors grow. Specifically, the sum of the squared cosines of the angles is strictly increasing as the dimensionality of the representation decreases. Thus spectral methods work because the truncated eigenbasis amplifies structure in the data so that any heuristic post-processing is more likely to succeed. We use this result to construct a nonlinear dimensionality reduction (NLDR) algorithm for data sampled from manifolds whose intrinsic coordinate system has linear *and* cyclic axes, and a novel clustering-by-projections algorithm that requires no post-processing and gives superior performance on "challenge problems" from the recent literature.

1 Introduction

Spectral methods for multivariate data analysis are notable both for their practical successes and for their rapidly developing theoretical underpinnings. A spectral algorithm typically begins with an "affinity matrix" of pairwise relationships between the samples or the variates, and derives a more useful representation of the data from its eigenvalue decomposition (EVD), often using just one or a few eigenvectors (a truncated eigenbasis). Many classic dimensionality reduction and nonlinear embedding algorithms have this character: Principal components analysis (PCA) [11] uses the variates' covariance matrix; multidimensional scaling (MDS) [16] uses the samples' pairwise distance matrix; kernel PCA [1, 24] uses a kernel matrix where the kernel function $\kappa(\mathbf{x}_i, \mathbf{x}_j)$ represents the dot product of samples in an unknown "feature space[1]"; and locally linear embedding (LLE) [23] uses a matrix containing correlations of samples' barycentric coordinates.

Spectral methods have been even more successful for data clusterings and graph partitionings: Spectral bipartitioning [9, 6] cuts a graph in two by thresholding the second eigenvector of the graph's normalized Laplacian matrix, and numerous clustering algorithms use selected eigenvectors of dot-product or kernel matrices to re-represent the data for clustering by simpler heuristics such as thresholding or K-means [25, 2, 5, 7, 27, 21, 29, 13, 3, 18, 19, 4, 20, 22].

While the statistical basis and optimality of PCA is well understood, virtually all other spectral methods are motivated by imperfect analogies between data-derived graphs and physical problems (e.g., harmonic analysis[2] and random walks[3]), or as approximations to other problems (e.g., vector quantization [2], min-cut [27], or max-flow [6]).

Underlying all this work is the notion that the truncated eigenvector basis somehow makes the problem simpler for the subsequent analysis. Our theoretical goal is to explain how and why this works.

Embeddings and clusterings imply loss of information, but there has been little effort to bound or even quantify what is lost and characterize what is conserved. This is acutely true for the vast majority of algorithms in which the spectral analysis is just a prelude to further information-lossy data analysis. Promising steps in the right direction include work by Alpert & Yao [2] that equates spectral partitioning with vector quantization (thereby implying an objective function), and analyses by Fiedler, Perona & Freeman, Shi & Malik, Meila & Shi, and Weiss et al. [9, 21, 27, 29, 19, 20] that justify using one or a few eigenvectors as a cluster indicator when the data is already clustered or nearly so. In particular, if the affinity matrix already has a block structure, then some of its eigenvectors will be piece-wise constant, such that if items i and j share cluster membership, elements i and j in these eigenvectors

[1] Feature space is the "linearization space" of Aizerman et al. [1], in which Euclidean relationships between points are consistent with the kernel's similarity measure.

[2] The eigenvectors of a graph's normalized Laplacian matrix are analogous to the modes of vibration the graph would exhibit if shaken [8].

[3] The eigenvectors of a stochastic matrix describe the steady-state properties of an infinite random walk on the graph [6, 19].

will have the same value [29, 19, 20]. For nearly clustered data, the eigenvectors are approximately piecewise.

However, this leaves open many questions, particularly when the data is *not* nearly clustered: What special properties should the affinity matrix have? Stochasticity? Unit diagonal? Positive definiteness? Unit spectral radius? Which and how many eigenvectors should be used? What information is conserved in a truncated eigenbasis? Obviously, the answers to these questions should inform the post-processing of the eigenvectors.

In this paper we develop a unifying view of spectral methods that answers many of these questions and gives guidance for the construction of clustering and nonlinear dimensionality reduction algorithms. For most of our discussion, it will be useful to think of the affinity matrix in terms of a (possibly unknown) kernel: Affinity value $A_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_j)$ is the dot product of two vectors representing the (usually unknown) locations of points i and j in a high-dimensional feature space associated with the kernel. Spectral analysis gives a new data representation derived from the eigenvalues and eigenvectors of the symmetric affinity matrix $\mathbf{A}$. To summarize the main theoretical result:
(1) An eigenvalue-scaled eigenvector representation of the data encodes angles (equivalently, correlations) between points embedded in the surface of a hypersphere.
(2) When the representation is truncated by suppressing the smallest magnitude eigenvalues, the angles (equiv., correlations) between high-affinity points are least distorted, highlighting the manifold structure of the data.
(3) As the representation is further truncated, the angles (equiv., correlations) *decrease* between points having high affinity and *increase* between points having low affinity, highlighting the cluster structure of the data.

In short, nonlinear dimensionality reduction and clustering can be obtained from the same process. The theorem is limited to symmetric non-negative definite affinity matrices, but a corollary establishes relevance to non-positive matrices as well, and to asymmetric matrices (e.g., $\mathbf{B}$) via their Grams ($\mathbf{B}^\top\mathbf{B}$ or $\mathbf{B}\mathbf{B}^\top$).

In the remainder of the paper we leverage this theorem into novel methods for nonlinear dimensionality reduction (NLDR) and clustering. The NLDR algorithm maps the data to a mixed vector and toric space, with the linear or cyclic nature of each axis determined from statistical tests. The clustering algorithm works entirely by projections, whose information loss is easily characterized and minimized or bounded at each step. Experiments show that it produces high-quality clusterings of a wide variety of "challenge problems" exhibited in the recent literature. We also use it to solve an unusually difficult visual segmentation problem.

2 The polarization theorem

Let $\mathbf{A} \in \mathcal{R}^{D\times D}$ be a non-negative definite symmetric matrix having eigenvalue decomposition (EVD) $\mathbf{V}\Lambda\mathbf{V}^\top = \mathbf{A}$ with eigenvalues sorted in descending order on the diagonal of Λ. Define representation $\mathbf{X} \doteq \Lambda^{1/2}\mathbf{V}^\top$ and let truncated representation $\mathbf{X}_{(d)}$ be the top d rows of $\mathbf{X}$—the d principal eigenvectors scaled by the square roots of their associated eigenvalues. A well-known property of such truncated EVDs is that $\mathbf{A}_{(d)} \doteq \mathbf{X}_{(d)}^\top \mathbf{X}_{(d)}$ is the best rank-d approximation to $\mathbf{A}$ with respect to the Frobenius norm; equivalently, the most energy-preserving projection to rank d.

Let $\mathbf{Y}_{(d)}$ be an angle-preserving projection of the column vectors of $\mathbf{X}_{(d)}$ onto the surface of a d-dimensional hypersphere, obtained by scaling each column of $\mathbf{X}_{(d)}$ to unit norm. The angle between two column vectors $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_{(d)}$ (equivalently $\mathbf{y}_i, \mathbf{y}_j \in \mathbf{Y}_{(d)}$) is

$$\theta_{ij} \doteq \angle(\mathbf{x}_i, \mathbf{x}_j) = \arccos \frac{\mathbf{x}_i^\top \mathbf{x}_j}{\|\mathbf{x}_i\| \cdot \|\mathbf{x}_j\|} \tag{1}$$

and the correlation between two vectors $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_{(d)}$ (equivalently $\mathbf{y}_i, \mathbf{y}_j \in \mathbf{Y}_{(d)}$) is

$$\mathrm{corr}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{y}_i^\top \mathbf{y}_j = \cos\theta_{ij}. \tag{2}$$

We may now state the main result:

Theorem (polarization): *As positive (resp., non-negative)* $\mathbf{A}$ *is projected to successively lower ranks* $\mathbf{A}_{(D-1)}, \mathbf{A}_{(D-2)}, \cdots, \mathbf{A}_{(d)}, \cdots, \mathbf{A}_{(2)}, \mathbf{A}_{(1)}$*, the sum of squared angle-cosines* $\sum_{i\neq j}(\cos\theta_{ij})^2$ *(equivalently squared correlations* $\|\mathbf{Y}_{(d)}^\top \mathbf{Y}_{(d)}\|_F^2$*) is strictly increasing (resp., non-decreasing).*

In short, as the dimensionality of the representation is reduced, the distribution of cosines migrates away from 0 toward the two poles ± 1, such that angles migrate from $\theta_{ij} = \pi/2$ to $\theta_{ij} \in \{0, \pi\}$.

The full proof requires a large number of lemmas and runs to several pages; because of page limits it will be published separately. The following proof sketch gives the flavor of the argument: The identity $\mathrm{diag}(\Lambda) = (\mathbf{V} \circ \mathbf{V})\mathrm{diag}(\mathbf{A})$ allows one to derive the distribution of the nonzero eigenvalues $[\gamma_1, \cdots, \gamma_d]$ of the cosine matrix $\mathbf{Y}_{(d)}^\top \mathbf{Y}_{(d)} = \mathrm{diag}(\mathrm{diag}(\mathbf{A}_{(d)}))^{-1/2}\mathbf{A}_{(d)}\mathrm{diag}(\mathrm{diag}(\mathbf{A}_{(d)}))^{-1/2}$. One can then show that as $d \downarrow 1$ the variance of the eigenvalues grows. However, the projection onto the hypersphere keeps the mean root-eigenvalue constant at $\frac{1}{d}\sum_i^d \gamma_i^{1/2} = 1$. Therefore the sum $\sum_i^d \gamma_i = \mathrm{trace}(\mathbf{Y}_{(d)}^\top \mathbf{Y}_{(d)}) = \|\mathbf{Y}_{(d)}^\top \mathbf{Y}_{(d)}\|_F^2 = \sum_{i\neq j}(\cos\theta_{ij})^2 + D$ grows monotonically.

The following two corollaries will be developed into algorithms in the remainder of the paper:

Corollary (embedding): *Suppressing the smallest-magnitude eigenvalues of* $\mathbf{A}$ *gives a* $d < D$ *dimensional embedding in which small angles are least distorted.*

This unsurprising corollary is quite similar to the motivation for PCA. In our case, the mass-preserving embedding spreads the data out on the hypersphere surface, preserving small angles most accurately because their cosines comprise most of the energy in the affinity matrix. This means that local relations between nearby points are well preserved. In the next section we show that this allows one to construct a relatively low-dimensional embedding for affinity data, with the unusual feature that the embedding space may have both linear and cyclic degrees of freedom.

Corollary (clustering): *Truncation of the eigenbasis amplifies any unevenness in the distribution of points on the d-dimensional hypersphere by causing points of high affinity to move toward each other and other to move apart.*

In short, the distribution approaches a clustering for small $d \ll D$. This explains many, if not all, spectral clustering methods: Using a subset of all the eigenvectors emphasizes the data's cluster structure, improving the output of any heuristic clustering procedure. This does *not* mean that the lowest-dimensional embedding is the best one for clustering; there is a tradeoff between amplifying cluster structure and losing information. In section 4, we show that by using a large subset of eigenvectors one can depend entirely on projections and EVDs to do the clustering, removing the need for heuristic post-processing.

Although the theorem is limited to non-negative symmetric affinity matrices, it also has explanatory value for spectral methods that employ selected eigenvectors of non-positive matrixes: Every real symmetric matrix can be written $\mathbf{C} = \mathbf{A} - \mathbf{B}$ where positive semi-definite matrices $\mathbf{A}$ and $\mathbf{B}$ satisfy $\text{rank}(\mathbf{A}) + \text{rank}(\mathbf{B}) = \text{rank}(\mathbf{C})$ and $\mathbf{A}$, constructed from the positive part of $\mathbf{C}$'s spectrum, is the best (least-squares) gram approximation of $\mathbf{C}$ (one offering a real-valued decomposition $\mathbf{A} = \mathbf{X}^\top\mathbf{X}$). The theorem applies to $\mathbf{A}$. For example, Weiss [29] showed that clustering methods related to the min-cut problem (e.g., [21, 27, 18]) are of this nature: Although posed as generalized eigenvalue problems with non-positive Laplacian matrices, these algorithms ultimately consult a single eigenvector from the positive part the spectrum of the normalized Laplacian matrix. It is worth noting that it follows from basic properties of the normalized Laplacian [6] that either (1) this eigenvector is only approximately piecewise constant, presenting some uncertainty for the final clustering, or (2) eigenvalue multiplicity makes the choice of eigenvector ambiguous. The polarization theorem suggests that additional relevant information lies in the remaining eigenvectors; below we construct an algorithm that exploits this information and eliminates the abovementioned ambiguities.

3 Dimensionality reduction

Motivated by the first corollary above, we observe that $\mathbf{Y}_{(d)}$ is a low-dimensional nonlinear embedding of the data onto the surface of a d-dimensional hypersphere, with the arc-length between two points inversely related to their affinity score. Let us choose a dimensionality d that truncates only the lesser eigenvalues of the affinity matrix $\mathbf{A}$. As in PCA, this choice is usually a matter of eyeballing the eigenvalue spectrum, unless one has prior knowledge about the true noise levels in the data and how they affect the kernel.

Since it is difficult to work with spherical embeddings, our goal is to "re-embed" the data in a vector space, where possible. Let $\mathbf{p}$ be the point on the hypersphere surface having smallest arc-length to all points in $\mathbf{Y}$ and let $[\mathbf{u}_1, \cdots, \mathbf{u}_{d-1}]$ be an orthogonal basis of the hyperplane tangent to the surface at $\mathbf{p}$. (Note that $\mathbf{U} \doteq [\mathbf{u}_1, \cdots, \mathbf{u}_{d-1}, \mathbf{p}]$ satisfies $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$.) Each $\mathbf{u}_i$ specifies a direction around the hypersphere which can be visualized as a great circle parallel to $\mathbf{u}_i$ and passing through $\mathbf{p}$. Each $\mathbf{u}_i$ can be interpreted as one of the axes of a projection to $\mathcal{R}^a \times \mathcal{T}^b$—the Cartesian product of an a-dimensional vector space and a b-dimensional toric[4] space—with $a + b = d - 1$. If the data wraps fully around the hypersphere along direction $\mathbf{u}_i$, then axis i is cyclic; if the data wraps only partway around the hypersphere along direction $\mathbf{u}_j$, then axis j is linear.

Much as PCA axes are statistically motivated from multivariate gaussian distribution, the tangent point $\mathbf{p}$ and axes $\mathbf{U}$ can be estimated by fitting a gaussian distribution to the surface of a hypersphere. The complex Bingham distribution [15] is a multivariate gaussian density on $\mathbf{y} \in \mathcal{CS}^{d-1} \subset \mathcal{C}^d$ conditioned on the fact that all vectors are unit-length ($\mathbf{y}^*\mathbf{y} = 1$, where $\mathbf{y}^*$ denotes complex conjugate transpose):

$$p(\mathbf{y}|\Sigma) = C(\Sigma)^{-1} \exp(\mathbf{y}^*\Sigma\mathbf{y}). \tag{3}$$

The complex Bingham is parameterized by hermitian matrix $\Sigma = \mathbf{U}\text{diag}([\kappa_1, \cdots, \kappa_d])\mathbf{U}^*$ whose eigenvalues $\kappa_1 < \kappa_2 < \cdots < \kappa_{d-1} < \kappa_d = 0$ are the concentration parameters of the density. A strongly negative $\kappa_i \ll 0$ indicates that the density has little extent along direction $\mathbf{u}_i$. The last eigenvector $\mathbf{u}_d$ points to the mode of the distribution, thus $\mathbf{p} = \mathbf{u}_d$. The normalizing constant $C(\Sigma)$ is calculated via the matrix confluent hypergeometric function ${}_1F_1$ [14], which in this case has a compact form discovered by Kent [15]:

$$C(\Sigma) = {}_1F_1(\frac{d}{2}, \frac{1}{2}, \text{diag}([\kappa_1, \cdots, \kappa_d])) \tag{4}$$

$$= 2\pi^{d-1} \sum_{j=1}^{d-1} \frac{e^{\kappa_j}}{\prod_{i \neq j}(\kappa_j - \kappa_i)}. \tag{5}$$

Jupp and Mardia [12] showed that the direction vectors in $\mathbf{U}$ and the concentration parameters $[\kappa_1, \cdots, \kappa_{d-1}]$ are related to the scatter of $\mathbf{Y}$ through its EVD $\mathbf{U}\text{diag}([\lambda_1, \cdots, \lambda_d])\mathbf{U}^* = \mathbf{Y}\mathbf{Y}^\top$, with $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_d$

[4] A toric space $\mathcal{T}^n = (\mathcal{S}^1)^n$ has n cyclic axes but, unlike spherical space $\mathcal{S}^n$, every point has a unique set of ordinates modulo 2π; there are no poles presenting singularities. E.g.: when walking, leg and arm phase are two cyclic variates in $\mathcal{T}^2$; but the orientation of a featureless cone in 3-space is described by two variates in $\mathcal{S}^2$ (a.k.a. Euler angles).

satisfying

$$\lambda_j = \frac{\partial \log C(\Sigma)}{\partial \kappa_j}. \tag{6}$$

For large sample sizes with concentrated density, $\kappa_j \approx -c/\lambda_j$ for some constant c. Numerical solution for the concentration parameters in higher dimensions is feasible but nontrivial. Fortunately, for dimensionality reduction, knowing the EVD of the scatter suffices: Its eigenvectors $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_{d-1}, \mathbf{p}]$ are *exactly* the maximum likelihood (ML) estimate of the point of tangency (a.k.a. modal direction) and the axes of the tangent space, while its eigenvalues $[\lambda_1, \cdots, \lambda_d]$ give a rough indication of which axes are cyclic: A small eigenvalue ($\lambda_i \approx 0$) indicates that the data is approximately linear in direction $\mathbf{u}_i$; a large eigenvalue ($\lambda_i \approx \lambda_d$) indicates the axis is cyclic.

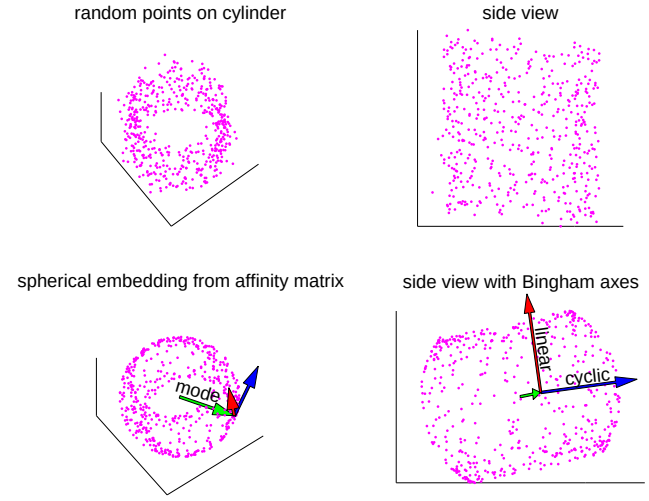

Figure 1: Dimensionality reduction of points distributed on a 3D cylinder (in 10D space) to the 2D space $\mathcal{R}^1 \times \mathcal{T}^1$—one linear axis and one cyclic axis. 500 points are randomly generated on the surface of a 3D cylinder embedded in 10D space, and contaminated with isotropic 10D gaussian noise. The top two images show a 3D and a 2D projection of the points. The bottom two images show an embedding of the affinity matrix on the surface of a 3D sphere. The embedding forms a wide belt around the equator. The arrows show the modal direction and two degrees of freedom of a Bingham distribution fitted to the embedding. Statistical tests for uniformity indicate that the data is cyclic around the equator but linear in the other direction.

To test more precisely whether the data is cyclic in direction $\mathbf{u}_i$ around the surface of the hypersphere, consider the projection of the data onto the great circle parallel to $\mathbf{u}_i$ and passing through the mode $\mathbf{u}_d$. A uniform distribution on this circle ($= \mathcal{S}^1 = \mathcal{T}^1$) implies that the data is cyclic in direction $\mathbf{u}_i$. The projection is $\mathbf{Z} \doteq [\mathbf{z}_1, \cdots, \mathbf{z}_N]$ with $\mathbf{z}_j \propto [\mathbf{u}_d, \mathbf{u}_i]^\top \mathbf{y}_j$, $\|\mathbf{z}_j\| = 1$. With this projection, we can apply a result of Mardia [17] which gives a statistical test

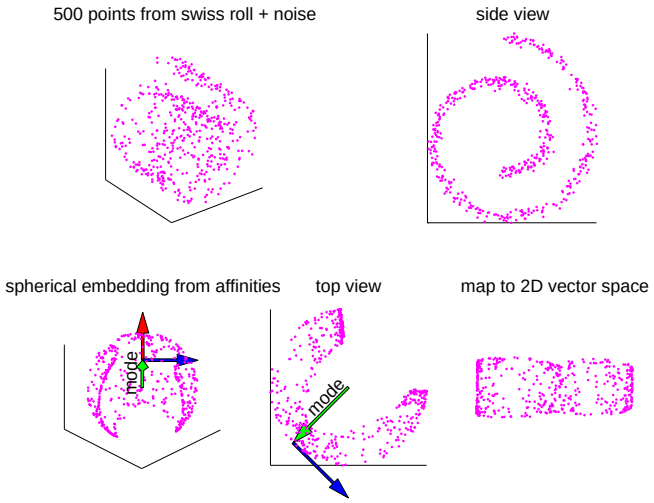

Figure 2: Dimensionality reduction of points distributed on a curled plane in 10D space to the 2D vector space $\mathcal{R}^2$. The data is analyzed as in figure 1, but the spherical embedding does not wrap all the way around the equator. The statistical test indicates that the data is Bingham-distributed in both directions, so the sphere's surface is azimuthally mapped to a 2D vector space, recovering the original planar coordinate system. Since the original data manifold passes close by itself, some points, particularly at the ends, have affinity for nonlocal neighbors, resulting in some distortion in the recovered coordinates.

to assess the hypothesis that a distribution on the (perimeter of the) unit circle is uniform rather than complex Bingham: Let γ_1, γ_2 be the eigenvalues of the normalized scatter $\mathbf{Z}\mathbf{Z}^\top/n$. Then, for large data-sets ($n \gg 0$),

$$3n(\gamma_1 - \gamma_2)^2 \simeq \chi^2_3, \tag{7}$$

where χ^2_3 is a chi-squared distribution with three degrees of freedom. Thus we may reject the hypothesis that axis $\mathbf{u}_i$ is uniform (cyclic) with confidence $\Pr(\chi^2_3 > 3n(\gamma_1 - \gamma_2)^2)$.

Once identified, the non-cyclic axes are isolated by projecting $\mathbf{Y}$ onto the union of the non-cyclic axes and $\mathbf{p}$, then rescaling the resulting vectors to unit norm. After projection onto this reduced hypersphere, the modal point is $[0, \cdots, 0, 1]$ and the Bingham axes are similarly axis-aligned. An azimuthal equidistant mapping at the modal point then takes the points into a vector space. The process is illustrated in figure 1 for 10D points noisily sampled from a 2D nondevelopable[5] manifold having genus 1 (a cylinder) and in figure 2 for 10D points noisily sampled from a 2D developable manifold having genus 0 (a rectangular plane curled into a "swiss roll"). The affinity matrix for all data-sets in this paper is $A_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j) \propto$

[5] A "developable" d-dimensional manifold embedded in $\mathcal{R}^n$ can mapped to $\mathcal{R}^d$ without internal distortions, e.g., a developable surface can be unrolled and flattened without stretching.

$\exp -(\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2)$, where σ is taken to be the average of the distances between each point and its closest neighbor (which we denote σ_s). For simplicity of analysis, we normalize the affinity matrix by projecting it to the nearest doubly stochastic matrix $\mathbf{P} \doteq \mathrm{diag}(\mathbf{d})\mathbf{A}\mathrm{diag}(\mathbf{d})$ using a fast modification of the Sinkhorn procedure [28] to solve for $\mathbf{d}$ satisfying $\mathbf{P1} = \mathbf{1}$ (and $\mathbf{P}^\top\mathbf{1} = \mathbf{1}$ since $\mathbf{P}^\top = \mathbf{P}$). While not crucial to this method, a doubly stochastic matrix has two properties that make it appealing as a model of the data: $\mathbf{P}$ can be interpreted as the transition probabilities of a random walk on the data; and $\mathbf{P}$'s largest eigenvalue $\lambda_{\max} = 1$ has corresponding eigenvector $\mathbf{u}_1 \propto \mathbf{1}$, which implies that the stationary distribution of the random walk is uniform (every point is equally probable). To obtain the embeddings in the figures, we discarded the totally uninformative $\mathbf{u}_1$ and constructed $\mathbf{Y}_{(d)}$ as in section 2 from eigenvectors 2-4 of $\mathbf{P}$, then fitted Bingham densities to the results to determine the appropriate embedding in $\mathcal{R}^a \times \mathcal{T}^b$.

One could also make embeddings in $\mathcal{R}^a \times \mathcal{T}^b \times \mathcal{S}^c$, though testing uniformity hypotheses on $\mathcal{S}^c$ for $c > 1$ requires explicit calculation of the Bingham concentration parameters. In practice, we find that the ML modal estimator for the complex Bingham distribution is rather sensitive to noise; many samples may be necessary to get a good estimate. Fisher or Watson distributions may be better behaved, but currently they are less tractable analytically and computationally. We turn now to the problem of clustering, where we obtain an easily analyzed, highly competitive algorithm.

4 Clustering

Spectral methods have been extensively studied in graph partitioning and clustering problems. Fiedler [9] first showed that the eigenvector of the Laplacian matrix corresponding to the second eigenvalue gives an embedding of the graph in a real line; cutting this embedding at the origin gives a bipartitioning of the graph. This was extended to k-way partitioning where the feature points are mapped into a k-dimensional space with the new coordinates being the normalized row vector of the matrix formed by the first k eigenvectors of the affinity matrix [25, 26]. Similarly, in Ng et al. [20], the normalized row vectors of the matrix formed by the first k weighted eigenvectors are used as the input to a k-means clusterer, and a perturbational analysis was used to show that the results should be stable if the data is already "nearly clustered". In Chan et al. [5], the directional angle between the row vectors of the first k eigenvectors of the Laplacian matrix was used as a new distance measure for partitioning. Alpert & Yao [2] equated partitioning with the problem of clustering these row-vectors, and found that the more eigenvectors used, the better. Spectral bipartitioning methods were adapted for visual clustering problems by Perona and Freeman [21] and Shi & Malik [27]. Analyses by Weiss [29] and Meila & Shi [18, 19] showed that normalizing a nearly block-structured affinity matrix makes its eigenvectors approximately piecewise constant, and therefore easy to interpret as cluster assignments. However, such structure is not guaranteed for real problems, and some post-processing is necessary.

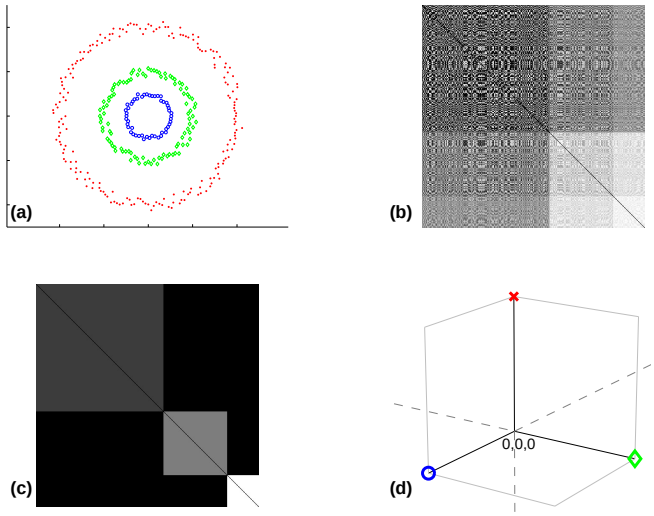

Figure 3: Spectral clustering of data distributed in three rings. (a) Cluster assignments are indicated by different markers. (b) The initial log-affinity matrix, sorted by true clusters (the algorithm is blind to such orderings). (c) The $\mathbf{P}$ matrix at convergence after 4 iterations. (d) In the converged representation, all the points belonging to any one cluster are co-located in a "corner" of a 3D sphere.

Our goal in visiting this already crowded field is to eliminate heuristic post-processing steps. Based on the our theoretical result, we constructed one (of many possible) spectral clustering algorithms in which there is no post-EVD clustering or thresholding; instead, the stochastic eigenvectors form a discrete indicator matrix showing the membership of each point.

Our basic strategy is to cast clustering as two alternating projections: Projection to low-rank, and projection to the set of zero-diagonal doubly stochastic matrices. In both cases it is easy to characterize what is conserved and quantify what is lost. The projection to lower rank $\mathbf{A} \to \mathbf{A}_{(d)}$ (or $\mathbf{P} \to \mathbf{A}_{(d)}$) is exactly the process characterized by the polarization theorem: We polarize the distribution of angles with minimal loss of energy $\|\mathbf{A} - \mathbf{A}_{(d)}\|_F^2$. The projection to a zero-diagonal doubly stochastic matrix $\mathbf{A}_{(d)} \to \mathbf{P} = \mathrm{diag}(\mathbf{d})(\mathbf{A}_{(d)} - \mathrm{diag}(\mathrm{diag}(\mathbf{A}_{(d)})))\mathrm{diag}(\mathbf{d})$ suppresses any differences in the stationary probability of points induced by the projection to low-rank. Disregarding the suppressed diagonal, this projection is simply an angle-preserving rescaling of the embedding vectors, because $\mathrm{diag}(\mathbf{d})\mathbf{A}_{(d)}\mathrm{diag}(\mathbf{d}) = \mathrm{diag}(\mathbf{d})\mathbf{X}_{(d)}^\top\mathbf{X}_{(d)}\mathrm{diag}(\mathbf{d}) = (\mathbf{X}_{(d)}\mathrm{diag}(\mathbf{d}))^\top(\mathbf{X}_{(d)}\mathrm{diag}(\mathbf{d}))$. Suppressing the diagonal induces negative eigenvalues in the spectrum of $\mathbf{P}$ (associated with removing energy placed on the diagonal by the positive eigenvalues); these eigenvalues account for less

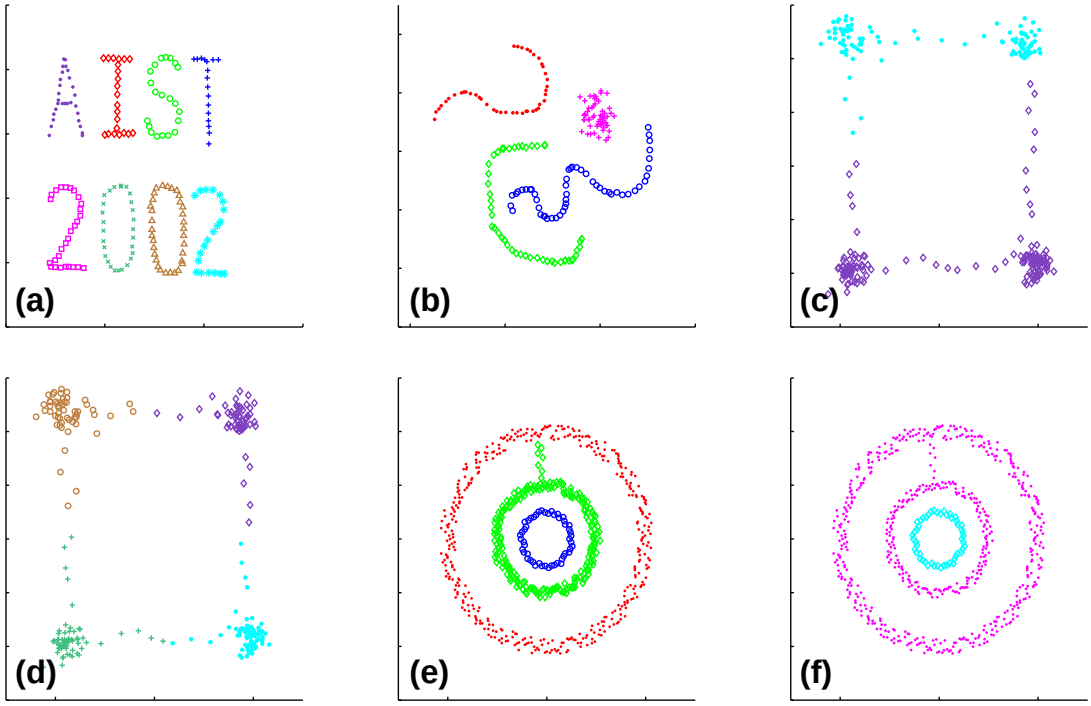

Figure 4: A gallery of "challenge problems" adapted from [20] and successfully clustered by our method. Cluster membership is indicated by marker symbol. As with all radial kernel methods, clustering reflects connectivity and scale similarity; thus the size of the kernel has an effect on the results: The same data-set is differently clustered with $\sigma = 10\sigma_s$ in (c) and $\sigma = 2\sigma_s$ in (d) (setting $\sigma = \sigma_s$ breaks the sides from the corners). A very similar data-set was only bipartitioned after a search over σ in [20]. Similar results can be observed for (e) and (f) where $\sigma = \sigma_s/2$ for (e) and $\sigma = \sigma_s$ for (f).

than half the energy $\|\mathbf{P}\|_F^2$ in $\mathbf{P}$. We project to lower rank by suppressing the negative eigenvalues and the uninformative unit eigenvalue. This gives an automatic determination of d and a bound on the loss of variance. The alternating projections terminate when the resulting $\mathbf{P}$ matrix has two or more stochastic (unit) eigenvalues, implying reducibility. (A reducible matrix $\mathbf{P}$ can be row- and column-permuted into block-diagonal form, e.g., figure 5c).

Analysis: It can be shown that **(1)** If $\mathbf{P}$ has unique positive nonstochastic eigenvalues, alternating the projections $\mathbf{P} \to \mathbf{A}_{(d)}$ and $\mathbf{A}_{(d)} \to \mathbf{P}$ will drive the leading eigenvalues up toward the positive bound $\lambda_i \leq +1$ and all other eigenvalues down toward the negative bound $\lambda_i \geq -1$. Formally, the vector norm of the eigenvalues increases while their sum remains constant. **(2)** Once $\mathbf{P}$ becomes reducible (has multiple stochastic eigenvalues $\lambda_i = +1$), the matrix $\mathbf{X}_{(k)}$ whose k columns are $\mathbf{P}$'s stochastic eigenvectors has exactly k unique rows. **(3)** Let $\mathbf{Z}_{(k)}$ be a matrix formed from these rows. The product $\mathbf{Z}_{(k)}\mathbf{Z}_{(k)}^\top$ is diagonal and the product $\mathbf{X}_{(k)}\mathbf{Z}_{(k)}^\top(\mathbf{Z}_{(k)}\mathbf{Z}_{(k)}^\top)^{-1} = \mathbf{X}_{(k)}\mathbf{Z}_{(k)}^{-1}$ is a binary (0/1) cluster indicator matrix that maps all points in a single cluster to a unique positive axis of $\mathcal{R}^k$ (e.g., figure 3d).

It remains to be shown whether the conditions in (1) are sufficient to guarantee absolute convergence. All our experiments, including those of figures 3–6, have converged quite quickly.

When the affinity matrix is produced by a gaussian kernel, this procedure groups points that all have similar affinity values, essentially creating clusters where the inter-point distances are all on the same scale. Clusters may be non-convex and wrap around each other (e.g., figure 4a, figure 4bef; figure 5ab), but the results are generally in agreement with human judgment; some authors have drawn connections between gaussian kernel clustering and human perceptual gestalts [21, 20].

This procedure automatically produces multi-way partitions without prior knowledge of the number of clusters. However, if the data contains clusters at different scales, the $\mathbf{P}$ matrix may become reducible before all the clusters have been revealed, only giving a partial clustering. Often we find that some remaining non-stochastic eigenvalues are close to 1, indicating the clusters can be further subpartitioned. We partition the $\mathbf{P}$ matrix according to its stochastic eigenvectors and continue the alternating projections on its submatrices, thereby obtaining a hierarchical clustering. Figure 5 treats a well-known problem this way.

4.1 Application to motion segmentation

Spectral clustering has become the preferred method for segmentation problems in computer vision [25, 7, 27, 21, 18, 29, 10, 4, 22]. Motion segmentation (e.g., [7]) takes

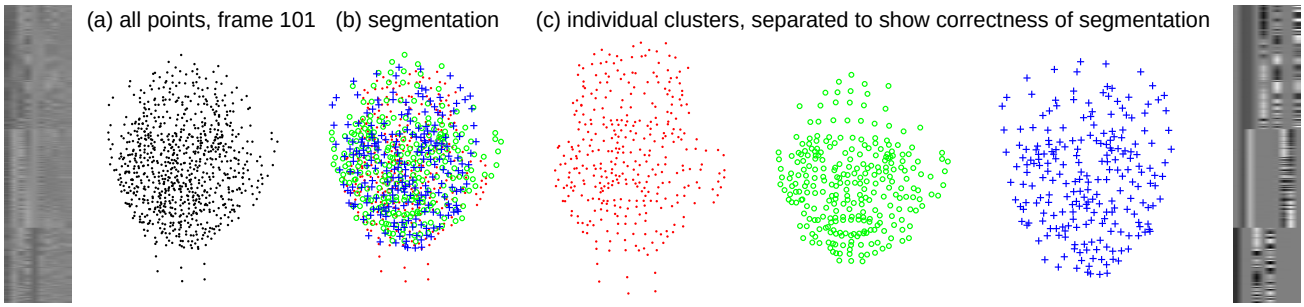

Figure 6: Nonrigid motion segmentation by spectral clustering. 500 frames of 2D tracking data from three people are scaled, centered, and superimposed to remove spatial cues that might aid grouping. The tracking matrix is factored $\mathbf{P} \rightarrow \tilde{\mathbf{M}}\tilde{\mathbf{S}}$ via SVD; spectral clustering of the columns of $\tilde{\mathbf{S}}$ correctly groups the points on the basis of correlated motion though time. The grayscale images depict the top ten eigenvectors of the affinity matrix at initialization (far left) and of the $\mathbf{P}$ matrix at convergence (far right) after 11 iterations. At initialization, there is a noisy hint of the clustering in eigenvectors 2 and 5; at convergence, the clustering is clear in piece-wise constant eigenvectors 1–3, which are associated with unit eigenvalues. Other eigenvalues are near 1, indicating that the faces can be subpartitioned; this results in segmentation of the jaws.

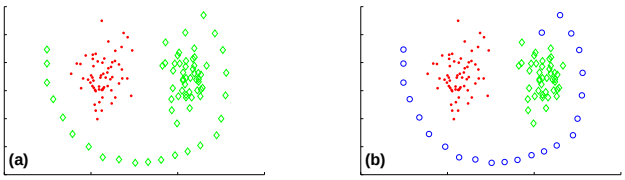

Figure 5: Hierarchical clustering of a problem treated in [18] and [20]. Our algorithm first gives a bipartitioning (a), then recursively analyzes the two subsets, using submatrices of $\mathbf{P}$ from the first partitioning. In the second round one subset is immediately bipartitioned to give (b). One point appears mis-classified but its assignment may be consistent with the kernel, since its distances to other points are more consistent with the scale of inter-point distances on the line ($\circ$'s) than of those in the tight cluster ($\diamond$'s).

2D tracking data for a number of points in the scene, and seeks to group those points into independently moving 3D objects on the basis of correlated motion. The 2D projection of N points on the rigid 3D surface of object j in image I_f is given by $\mathbf{P}_{fj} \in \mathcal{R}^{2\times N} = \mathbf{M}_{fj}\mathbf{S}_j$, where motion matrix $\mathbf{M}_{fj}$ encodes the position of the object relative to the camera, and shape matrix $\mathbf{S}_j \in \mathcal{R}^{3\times N}$ or $\in \mathcal{R}^{4\times N}$ gives the 3D location of the points in object-centered or homogeneous coordinates. With multiple frames and multiple objects,

$$\mathbf{P} = \mathbf{MS} \in \mathcal{R}^{2F\times(N_1+\cdots+N_J)} \quad (8)$$

$$\doteq \begin{bmatrix} \mathbf{M}_{11} & \cdots & \mathbf{M}_{1J} \\ \vdots & \ddots & \vdots \\ \mathbf{M}_{F1} & \cdots & \mathbf{M}_{FJ} \end{bmatrix} \begin{bmatrix} \mathbf{S}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{S}_J \end{bmatrix} \quad (9)$$

Each column of $\mathbf{S}$ describes one point; columns representing points from two different objects are orthogonal (Two columns from the same object may be orthogonal as well, but no column can be orthogonal to all columns from the same object). Unfortunately, $\mathbf{S}$ cannot be recovered directly from $\mathbf{P}$. Instead, a "thin" SVD can be used to factor $\mathbf{P} \rightarrow \tilde{\mathbf{M}}\tilde{\mathbf{S}} \doteq (\mathbf{MG}^{-1})(\mathbf{GS})$, but unknown matrix $\mathbf{G}$ may arbitrarily re-order or mix the columns of $\mathbf{S}$, destroying the orthogonal structure. However, if the motions of the objects are approximately independent (they are never perfectly so), two columns in $\tilde{\mathbf{S}}$ belonging to the same object will be more similar than two columns belonging to different objects, in the sense that their inner product will be more positive because the motions of the corresponding points is highly correlated. Therefore spectrally clustering the columns of $\tilde{\mathbf{S}}$ has been found to to be highly successful for segmenting the set of points into objects (e.g., [4]).

Two objections have been leveled at this approach: (1) It has only been applied to sequences in which the points can be segmented easily using simpler criteria such as spatial grouping. This is understandable; no one has tracking data for overlapping objects. (2) It is not clear that the method would extend to nonrigid objects, where the motions of points on one object are more weakly correlated.

To satisfy these objections, we constructed an unrealistically hard segmentation problem by superimposing dense face-tracking data from three "talking-heads" videos. The motion is highly nonrigid. To remove spatial separation cues, the data for each head was centered on the origin in each frame, so that when the data is combined all three faces overlap and there are no spatial or translational cues for segmentation. After centering, the motion of some points on a face is actually anti-correlated with most of the other points of the face (e.g., lower-lip versus upper-lip points). To accommodate the nonrigid motion, we increased the rank of the SVD by a factor of 5 ($\mathbf{S}_i \in \mathcal{R}^{15\times N}$), which allows 5 modes of deformation per object but yields a harder, higher-dimensional clustering problem. Figure 6 shows that $\tilde{\mathbf{S}}$ is perfectly clustered, yielding the correct motion segmentation. The tracking data, the evolution of the dominant eigenvectors, and the extracted clusters are shown in the accompanying video.

5 Summary

Spectral methods practitioners have long understood that a representation derived from selected eigenvectors of the affinity matrix somehow makes embedding and clustering problems easier for subsequent heuristic algorithms. To date, formal analyses have justified this approach only for problems with very obvious cluster structure and for certain kinds of affinity matrix. The polarization theorem of section 2 provides a unified explanation for virtually all the algorithms and affinity matrices in the cited literature: There exists an eigenvector representation which matches the angles between data-points in feature space; as the dimensionality of this representation is reduced, angles between similar points shrink while angles between dissimilar points grow. This highlights the cluster structure of the data and makes segmentation by heuristic methods significantly more likely to succeed. This theorem invites us to look at the representation as an embedding of the data on the surface of a hypersphere, where the inner product of two vectors gives the cosine of their angle. That insight led us to two algorithms: One finds nonlinear low-dimensional embeddings of data in spaces having a mixture of linear and cyclic axes; the other performs clusterings by repeated projections of the data, eliminating heuristic "post-clustering." The clustering algorithm has the appeal that all steps are well characterized in terms of what information about the distribution is preserved or lost, and the amount of information loss can be bounded and/or minimized. It also performs very well in practice on both synthetic "challenge problems" from the literature and a real-world motion segmentation problem that is considerably harder than those contemplated in the computer vision literature.

We are currently exploring better distributions for spherical data, bounds on convergence rates, and bounds on the rate at which angles change as dimensionality is reduced. In our work, we have benefitted from conversations with Sue Whitesides, Yoav Freund, Josh Tannenbaum, and Paul Viola, to whom we extend thanks.

References

[1] M. Aizerman, E. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[2] C. Alpert and S. Yao. Spectral partitioning: The more eigenvectors, the better. In *Proc. ACM/IEEE Design Automation Conference*, 1995.

[3] Y. Azar, A. Fiat, A. Karlin, F. McSherry, and J. Saia. Spectral analysis of data. In *Proceedings of the 33rd Symposium on Theory of Computing*, pp. 619–626, 2001.

[4] M. Carcassoni and E. Hancock. A hierarchical framework for spectral correspondence. In *Proc. Euro. Conf. Computer Vision*, pp. 266–281, 2002.

[5] P. Chan, D. Schlag, and J. Zien. Spectral K-way ratio-cut partitioning and clustering. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 13(9):1088–1096, 1994.

[6] F. R. Chung. *Spectral graph theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, 1997.

[7] J. Costeira and T. Kanade. A multi-body factorization method for motion analysis. In *Proc. Int. Conf. Computer Vision*, pp. 1071–1076, 1995.

[8] J. Demmel. Lecture notes on graph partitioning, part 2, April 1999. Notes for UC Berkeley CS267.

[9] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematics Journal*, 23:298–305, 1973.

[10] Y. Gdalyahu, D. Weinshall, and M. Werman. Self organization in vision: Stochastic clustering for image segmentation, perceptual grouping, and image database organization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(10):1053–1074, 2001.

[11] H. Hotelling. Analysis of complex statistical variables in principal components. *J. Educational Psychology*, 24:417–441, 498–520, 1933.

[12] P. Jupp and K. Mardia. Maximum likelihood estimators for the matrix von Mises-Fisher and Bingham distributions. *Annals of Statistics*, 7:599–606, 1979.

[13] R. Kannan, S. Vempala, and A. Vetta. On clusterings — good, bad and spectral. In *41st Symposium on the Foundations of Computer Science*, 2000.

[14] J. T. Kent. The Fisher-Bingham distribution on the sphere. *J. Royal Statistical Society, B*, 44:71–80, 1982.

[15] J. T. Kent. The complex Bingham distribution and shape analysis. *J. Royal Statistical Society, B*, 56:285–299, 1994.

[16] J. B. Kruskal and M. Wish. *Multidimensional Scaling*. Sage Publications, Beverly Hills,, 1978.

[17] K. Mardia. Directional statistics and shape analysis. Technical Report 24, U. Leeds Department of Statistics, 1995.

[18] M. Meila and J. Shi. Learning segmentation by random walks. In *Proc. Adv. Neural Info. Processing Systems*, pp. 873–879, 2000.

[19] M. Meila and J. Shi. A random walks view of spectral segmentation. In *AI and STATISTICS (AISTATS) 2001*, 2001.

[20] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Proc. Adv. Neural Info. Processing Systems*, 2002.

[21] P. Perona and W. Freeman. A factorization approach to grouping. In *Proc. Euro. Conf. Computer Vision*, pp. 655–670, 1998.

[22] A. Robles-Kelly and E. Hancock. Pairwise clustering with matrix factorisation and the EM algorithm. In *Proc. Euro. Conf. Computer Vision*, pp. 63–77, 2002.

[23] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, December 22 2000.

[24] B. Schölkopf, A. Smola, and K. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[25] G. Scott and H. Longuet-Higgins. Feature grouping by relocalisation of eigenvectors of the proximity matrix. In *Proc. British Machine Vision Conference*, pp. 103–108, 1990.

[26] G. Scott and H. Longuet-Higgins. An algorithm for associating the features of two patterns. *Proc. Royal Society London*, B244:21–26, 1991.

[27] J. Shi and J. Malik. Motion segmentation and tracking using normalized cuts. In *Proc. Int. Conf. Computer Vision*, pp. 1154–1160, 1998.

[28] R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Annals of Mathematical Statistics*, 35:876–879, 1964.

[29] Y. Weiss. Segmentation using eigenvectors: A unifying view. In *Proc. Int. Conf. Computer Vision*, pp. 975–982, 1999.