Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

The Hidden Convexity of Spectral Clustering

James Voss
Ohio State University
vossj@cse.ohio-state.edu

Mikhail Belkin
Ohio State University
mbelkin@cse.ohio-state.edu

Luis Rademacher
Ohio State University
lrademac@cse.ohio-state.edu

Abstract

In recent years, spectral clustering has become a standard method for data analysis used in a broad range of applications. In this paper we propose a new class of algorithms for multiway spectral clustering based on optimization of a certain "contrast function" over the unit sphere. These algorithms, partly inspired by certain Indepenent Component Analysis techniques, are simple, easy to implement and efficient.

Geometrically, the proposed algorithms can be interpreted as hidden basis recovery by means of function optimization. We give a complete characterization of the contrast functions admissible for provable basis recovery. We show how these conditions can be interpreted as a "hidden convexity" of our optimization problem on the sphere; interestingly, we use efficient convex maximization rather than the more common convex minimization. We also show encouraging experimental results on real and simulated data.

1 Introduction

Partitioning a dataset into classes based on a similarity between data points, known as cluster analysis, is one of the most basic and practically important problems in data analysis and machine learning. It has a vast array of applications from speech recognition to image analysis to bioinformatics and to data compression. There is an extensive literature on the subject, including a number of different methodologies as well as their various practical and theoretical aspects (Jain and Dubes 1988).

In recent years spectral clustering—a class of methods based on the eigenvectors of a certain matrix, typically the graph Laplacian constructed from data—has become a widely used method for cluster analysis. This is due to the simplicity of the algorithm, a number of desirable properties it exhibits and its amenability to theoretical analysis. In its simplest form, spectral bi-partitioning is an attractively straightforward algorithm based on thresholding the second bottom eigenvector of the Laplacian matrix of a graph. However, the more practically significant problem of multiway spectral clustering is considerably more complex. While hierarchical methods based on a sequence of binary splits have been used, the most common approaches use k-means or weighted k-means clustering in the spectral space or related iterative procedures (Shi and Malik 2000; Ng, Jordan, and Weiss 2002; Bach and Jordan 2006; Yu and Shi 2003). Typical algorithms for multiway spectral clustering follow a two-step process:

1. *Spectral embedding:* A similarity graph for the data is constructed based on the data's feature representation. If one is looking for k clusters, one constructs the embedding using the bottom k eigenvectors of the graph Laplacian (normalized or unnormalized) corresponding to that graph.
2. *Clustering:* In the second step, the embedded data (sometimes rescaled) is clustered, typically using the conventional/spherical k-means algorithms or their variations.

In the first step, the spectral embedding given by the eigenvectors of Laplacian matrices has a number of interpretations. The meaning can be explained by spectral graph theory as relaxations of multiway cut problems (Von Luxburg 2007). In the extreme case of a similarity graph having k connected components, the embedded vectors reside in $\mathbb{R}^k$, and vectors corresponding to the same connected component are mapped to a single point. There are also connections to other areas of machine learning and mathematics, in particular to the geometry of the underlying space from which the data is sampled (Belkin and Niyogi 2003).

In our paper we propose a new class of algorithms for the second step of multiway spectral clustering. The starting point is that when k clusters are perfectly separate, the spectral embedding using the bottom k eigenvectors has a particularly simple geometric form. For the unnormalized (or asymmetric normalized) Laplacian, it is simply a (weighted) orthogonal basis in k-dimensional space, and recovering the basis vectors is sufficient for cluster identification. This view of spectral clustering as basis recovery is related to previous observations that the spectral embedding generates a discrete weighted simplex (see (Weber, Rungsarityotin, and Schliep 2004) and also (Kumar, Narasimhan, and Ravindran 2013) for some applications). For the symmetric normalized Laplacian, the structure is slightly more complex, but is, as it turns out, still suitable for our analysis, and, moreover the algorithms can be used without modification.

The proposed approach relies on an optimization problem resembling certain Independent Component Analysis techniques, such as FastICA (see (Hyvärinen, Karhunen, and Oja 2004) for a broad overview). Specifically, the problem of identifying k clusters reduces to maximizing a certain "admissible" contrast function over a $(k-1)$-sphere. Each local maximum of such a function on the sphere corresponds to exactly one cluster in the data. The main theoretical contribution of our paper is to provide a complete characterization of the admissible contrast functions for geometric basis recovery. We show that such contrast functions have a certain "hidden convexity" property and that this property is necessary and sufficient for guaranteed recovery[1] (Section 2). Rather than the more usual convex minimization, our analysis is based on *convex maximization* over a (hidden) convex domain. Interestingly, while *maximizing* a convex function over a convex domain is generally difficult (even maximizing a positive definite quadratic form over the continuous cube $[0, 1]^n$ is NP-hard[2]), our setting allows for efficient optimization.

Based on this theoretical connection between clusters and local maxima of contrast functions over the sphere, we propose practical algorithms for cluster recovery through function maximization. We discuss the choice of contrast functions and provide running time analysis. We also provide a number of encouraging experimental results on synthetic and real-world datasets.

Finally, we note connections to recent work on geometric recovery. The paper (Anderson, Goyal, and Rademacher 2013) uses the method of moments to recover a continuous simplex given samples from the uniform probability distribution. Like our work, it uses efficient enumeration of local maxima of a function over the sphere. In (Hsu and Kakade 2013), one of the results shows recovery of parameters in a Gaussian Mixture Model using the moments of order three and can be thought of as a case of the basis recovery problem.

2 Summary of the Theoretical Results

In this section we state the main theoretical results of our paper on weighted basis recovery and briefly show how they can be applied to spectral clustering.

A Note on Notation. Before proceeding, we define some notations used throughout the paper. The set $\{1, 2, \ldots, k\}$ is denoted by $[k]$. For a matrix B, b_{ij} indicates the element in its i^{th} row and j^{th} column. The i^{th} row vector of B is denoted $b_{i\cdot}$, and the j^{th} column vector of B is denoted $b_{\cdot j}$. For a vector v, $\|v\|$ denotes its standard Euclidean 2-norm. Given two vectors u and v, $\langle u, v\rangle$ denotes the standard Euclidean inner produce between the vectors. We denote by $\mathbf{1}_{\mathcal{S}}$ the indicator vector for the set $\mathcal{S}$, i.e. the vector which is 1 for indices in $\mathcal{S}$ and 0 otherwise. The null space of a matrix M is denoted $\mathcal{N}(M)$. We denote the unit sphere in $\mathbb{R}^d$ by S^{d-1}. For points $p_1, \ldots, p_m$, $\operatorname{conv}(p_1, \ldots, p_m)$ will denote their convex hull.

Recovering a Weighted Basis. The main technical results of this paper deal with reconstructing a weighted basis by optimizing a certain contrast function over a unit sphere. We show that for certain functions, their maxima over the sphere correspond to the directions of the basis vectors. We give a complete description for the set of such functions, providing necessary and sufficient conditions.

More formally, consider a set $\{Z_1, \ldots, Z_m\}$ of orthonormal vectors in $\mathbb{R}^m$. These vectors form a hidden basis of the space. We define a function $F_g : S^{m-1} \to \mathbb{R}$ in terms of a "contrast function" g and strictly positive weights α_i, β_i as follows:

$$F_g(u) := \sum_{i=1}^{m} \alpha_i g(\beta_i |\langle u, Z_i\rangle|) \ . \tag{1}$$

We will provide a complete description of when the directions $Z_1, \ldots, Z_m$ can be recovered from the local maxima of F_g for arbitrary weights α_i, β_i. This process of finding the local maxima of F_g can be thought of as weighted *basis recovery.*

Here and everywhere else in the paper, we consider contrast functions $g : [0, \infty) \to \mathbb{R}$ that are continuous on $[0, \infty)$ and twice continuously differentiable on $(0, \infty)$. It turns out that the desirable class of functions can be described by the following properties:

P1. Function $g(\sqrt{x})$ is strictly convex.

P2. The (right) derivative at the origin, $\frac{d}{dx}(g(\sqrt{x}))|_{x=0^+}$, is 0 or $-\infty$.

The main idea underlying the proposed framework for weighted basis recovery comes from property P1. In particular, using just this property, it can be shown that the local maxima of F_g are contained in the set $\{\pm Z_i : i \in [m]\}$. The idea is to perform a change of variable to recast maximization of F_g over the unit sphere as a convex maximization problem defined over a (hidden) convex domain. We sketch the proof in order to illustrate this hidden role of convexity in weighted basis recovery.

Proof sketch: Maxima of F_g are contained in $\{\pm Z_i : i \in [m]\}$.

We will need the following fact about convex maximization (Rockafellar 1997, Chapter 32).

For a convex set K, a point $x \in K$ is said to be an *extreme point* if x is not equal to a strict convex combination of two other points of K.

Fact 1. *Let $K \subseteq \mathbb{R}^n$ be a convex set. Let $f : K \to \mathbb{R}$ be a strictly convex function. Then the set of local maxima of f on K is contained in the set of extreme points of K.*

As $Z_1, \ldots, Z_m$ form an orthonormal basis of the space, we may simplify notation by working in the hidden coordinate system in which $Z_1, \ldots, Z_m$ are the canonical vectors $e_1, \ldots, e_m$ respectively. Let $\Delta^{m-1} := \operatorname{conv}(e_1, \ldots, e_m)$ denote a (hidden) simplex. We will make use of the change of variable $\psi : S^{m-1} \to \Delta^{m-1}$ defined by $\psi_i(u) = u_i^2$. In particular, we define a family of functions $h_i : [0, \infty) \to \mathbb{R}$

[1] Interestingly, there are no analogous recovery guarantees in the ICA setting except for the special case of cumulant functions as contrasts. In particular, typical versions of FastICA are known to have spurious maxima (Wei 2015).

[2] This follows from (Gritzmann and Klee 1989) together with Fact 1 below.

for $i \in [m]$ by $h_i(t) = \alpha_i g(\beta_i \sqrt{t})$, and we define a function $H : \Delta^{m-1} \to \mathbb{R}$ as $H(t) = \sum_{i=1}^{m} h_i(t_i)$. Using assumption P1, it can be seen that H is a strictly convex function defined on a convex domain. Further, for any $u \in S^{m-1}$, $(H \circ \psi)(u) = F_g(u)$. Using this equality, we see that u is a maximum of F_g if and only if $\psi(u)$ is a maximum of H.

The extreme points of Δ^{m-1} are $Z_1, \ldots, Z_m$. By Fact 1, the maxima of H are contained in the set $\{Z_i : i \in [m]\}$. Hence, the maxima of F_g are contained in $\psi^{-1}(\{Z_i : i \in [m]\}) = \{\pm Z_i : i \in [m]\}$. □

We have demonstrated that F_g has no local maxima outside of the set $\{\pm Z_i : i \in [m]\}$; however, we have not demonstrated that the directions $\{\pm Z_i : i \in [m]\}$ actually are local maxima of F_g. In general, both P1 and P2 are required to guarantee that $\{\pm Z_i : i \in [m]\}$ is a complete enumeration of the local maxima of F_g. More formally, we have the following main theoretical results:

Theorem 2 (Sufficiency). *Let $\alpha_1, \ldots, \alpha_m$ and $\beta_1, \ldots, \beta_m$ be strictly positive constants. Let $g : [0, \infty) \to \mathbb{R}$ be a continuous function which is twice continuously differentiable on $(0, \infty)$ satisfying properties P1 and P2. If $F_g : S^{m-1} \to \mathbb{R}$ is constructed from g according to equation* (1)*, then all local maxima of F_g are contained in the set $\{\pm Z_i\}_{i=1}^{m}$ of basis vectors. Moreover, each basis vector $\pm Z_i$ is a strict local maximum of F_g.*

Theorem 3 (Necessity). *Let $g : [0, \infty) \to \mathbb{R}$ be a continuous function which is twice continuously differentiable on $(0, \infty)$, and let $F_g : S^{m-1} \to \mathbb{R}$ be constructed from g according to equation* (1)*.*

1. *If P1 does not hold for g, then there exists an integer $m > 1$ and strictly positive values of the parameters α_i, β_i such that F_g has a local maximum not contained in the set $\{\pm Z_i\}_{i=1}^{m}$.*
2. *If P1 holds but P2 does not hold for g, there exist strictly positive values of the parameters α_i, β_i such that at least one of the canonical directions Z_i is not a local maximum for F_g.*

The proofs of Theorems 2 and 3 (along with all other omitted proofs as well as the stability results for our methods) can be found in the long version of this paper.[3]

Spectral Clustering as Basis Recovery. It turns out that geometric basis recovery has direct implications for spectral clustering. In particular, when an n-vertex similarity graph has m connected components, the spectral embedding into $\mathbb{R}^m$ maps each vertex in the j^{th} connected component to a single point $y_j = \beta_j Z_j$ where $\beta_j = \|y_j\|$ and $Z_j = y_j / \|y_j\|$. It happens that the points $Z_1, \ldots, Z_m$ are orthogonal. Thus, letting x_i denote the embedded points and defining $F_g(u) := \frac{1}{n}\sum_{i=1}^{n} g(|\langle u, x_i\rangle|)$, there exist strictly positive weights $\alpha_1, \ldots, \alpha_m$ such that $F_g(u) = \sum_{j=1}^{m} \alpha_j g(\beta_j |\langle u, Z_j\rangle|)$. In particular, α_j is the fraction of vertices contained in the j^{th} component. Recovery of the basis directions $\{\pm Z_j\}_{j=1}^{m}$ corresponds to the recovery of the component clusters.

As the weights α_j and β_j take on a special form in spectral clustering, it happens that property P1 by itself is sufficient to guarantee that the local maxima of F_g are precisely the basis directions $\{\pm Z_j\}_{j=1}^{m}$.

3 Spectral Clustering Problem Statement

Let $G = (V, A)$ denote a similarity graph where V is a set of n vertices and A is an adjacency matrix with non-negative weights. Two vertices $i, j \in V$ are incident if $a_{ij} > 0$, and the value of a_{ij} is interpreted as a measure of the similarity between the vertices. In spectral clustering, the goal is to partition the vertices of a graph into sets $\mathcal{S}_1, \ldots, \mathcal{S}_m$ such that these sets form natural clusters in the graph. In the most basic setting, G consists of m connected components, and the natural clusters should be the components themselves. In this case, if $i' \in \mathcal{S}_i$ and $j' \in \mathcal{S}_j$ then $a_{i'j'} = 0$ whenever $i \neq j$. For convenience, we can consider the vertices of V to be indexed such that all indices in $\mathcal{S}_i$ precede all indices in $\mathcal{S}_j$ when $i < j$. Matrix A takes on the form:

$$A = \begin{pmatrix} A_{\mathcal{S}_1} & 0 & \cdots & 0 \\ 0 & A_{\mathcal{S}_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{\mathcal{S}_m} \end{pmatrix} ,$$

a block diagonal matrix. In this setting, spectral clustering can be viewed as a technique for reorganizing a given similarity matrix A into such a block diagonal matrix.

In practice, G rarely consists of m truly disjoint connected components. Instead, one typically observes a matrix $\tilde{A} = A + E$ where E is some error matrix with (hopefully small) entries e_{ij}. For i and j in different clusters, all that can be said is that $\tilde{a}_{ij}$ should be small. The goal of spectral clustering is to permute the rows and columns of $\tilde{A}$ to form a matrix which is nearly block diagonal and to recover the corresponding clusters.

4 Graph Laplacian's Null Space Structure

Given an n-vertex similarity graph $G = (V, A)$, define the diagonal degree matrix D with non-zero entries $d_{ii} = \sum_{j \in V} a_{ij}$. The unnormalized Graph Laplacian is defined as $L := D - A$. The following well known property of the graph Laplacian (see (Von Luxburg 2007) for a review) helps shed light on its importance: Given $u \in \mathbb{R}^n$,

$$u^T L u = \frac{1}{2} \sum_{i,j \in V} a_{ij}(u_i - u_j)^2 . \qquad (2)$$

The graph Laplacian L is symmetric positive semi-definite as equation (2) cannot be negative. Further, u is a 0-eigenvector of L (or equivalently, $u \in \mathcal{N}(L)$) if and only if $u^T L u = 0$. When G consists of m connected components with indices in the sets $\mathcal{S}_1, \ldots, \mathcal{S}_m$, inspection of equation (2) gives that $u \in \mathcal{N}(L)$ precisely when u is piecewise constant on each $\mathcal{S}_i$. In particular, $\{|\mathcal{S}_1|^{-1/2}\mathbf{1}_{\mathcal{S}_1}, \ldots, |\mathcal{S}_m|^{-1/2}\mathbf{1}_{\mathcal{S}_m}\}$ is an orthonormal basis for $\mathcal{N}(L)$.

In general, letting $X \in \mathbb{R}^{d \times m}$ contain an orthogonal basis of $\mathcal{N}(L)$, it cannot be guaranteed that the rows of X will

[3] See http://arxiv.org/abs/1403.0667 for the long version of this paper.

act as indicators of the various classes, as the columns of X have only been characterized up to a rotation within the subspace $\mathcal{N}(L)$. However, the rows of X are contained in a scaled orthogonal basis of $\mathbb{R}^m$ with the basis directions corresponding to the various classes. We formulate this result as follows (see (Weber, Rungsarityotin, and Schliep 2004), (Verma and Meilă 2003, Proposition 5), and (Ng, Jordan, and Weiss 2002, Proposition 1) for related statements).

Proposition 4. *Let the similarity graph $G = (V, A)$ contain m connected components with indices in the sets $\mathcal{S}_1, \ldots, \mathcal{S}_m$, let $n = |V|$, and let L be the unnormalized graph Laplacian of G. Then, $\mathcal{N}(L)$ has dimensionality m. Let $X = (x_{\cdot 1}, \ldots, x_{\cdot m})$ contain m scaled, orthogonal column vectors forming a basis of $\mathcal{N}(L)$ such that $\|x_{\cdot j}\| = \sqrt{n}$ for each $j \in [m]$. Then, there exist weights $w_1, \ldots, w_m$ with $w_j = \frac{|\mathcal{S}_j|}{n}$ and mutually orthogonal vectors $Z_1, \ldots, Z_m \in \mathbb{R}^m$ such that whenever $i \in \mathcal{S}_j$, the row vector $x_{i\cdot} = \frac{1}{\sqrt{w_j}} Z_j^T$.*

Proposition 4 demonstrates that using the null space of the unnormalized graph Laplacian, the m connected components in G are mapped to m basis vectors in $\mathbb{R}^m$. Of course, under a perturbation of A, the interpretation of Proposition 4 must change. In particular, G will no longer consist of m connected components, and instead of using only vectors in $\mathcal{N}(L)$, X must be constructed using the eigenvectors corresponding to the lowest m eigenvalues of L. With the perturbation of A comes a corresponding perturbation of the eigenvectors in X. When the perturbation is not too large, the resulting rows of X yield m nearly orthogonal clouds of points.

Due to different properties of the resulting spectral embeddings, normalized graph Laplacians are often used in place of L for spectral clustering, in particular $L_{\text{sym}} := D^{-1/2} L D^{-1/2}$ and $L_{\text{rw}} := D^{-1} L$. These normalized Laplacians are often viewed as more stable to perturbations of the graph structure. Further, spectral clustering with L_{sym} has a nice interpretation as a relaxation of the NP-hard multi-way normalized graph cut problem (Yu and Shi 2003), and the use of L_{rw} has connections to the theory of Markov chains (see e.g., (Deuflhard et al. 2000; Meilă and Shi 2001)).

For simplicity, we state all results in this paper in terms of L. However, when G consists of m connected components, $\mathcal{N}(L_{\text{rw}})$ happens to be identical to $\mathcal{N}(L)$, making Proposition 4 and all subsequent results in this paper equally valid for L_{rw}. The algorithms which we will propose for spectral clustering turn out to be equally valid when using any of L, L_{sym}, or L_{rw}, though the structure of $\mathcal{N}(L_{\text{sym}})$ is somewhat different. The description of $\mathcal{N}(L_{\text{sym}})$ and its analysis can be found in the long version of this paper.

5 Basis Recovery for Spectral Clustering

Given a graph G with n vertices and m connected components, let X; $\mathcal{S}_1, \ldots, \mathcal{S}_m$; $w_1, \ldots, w_m$; and $Z_1, \ldots, Z_m$ be constructed from L as in Proposition 4. The basis vectors $Z_1, \ldots, Z_m$ are mutually orthogonal in $\mathbb{R}^m$, and each weight $w_i = \frac{|\mathcal{S}_i|}{n}$ is the fraction of the rows of X indexed as $x_{\ell\cdot}$ coinciding with the point $\frac{1}{\sqrt{w_i}} Z_i^T$. It suffices to recover the basis directions Z_i up to sign in order to cluster the points. That is, each embedded point $x_{j\cdot} \in \mathcal{S}_i$ lies on the line through $\pm Z_i$ and the origin, making these lines correspond to the clusters.

We use an approach based on function optimization over projections of the embedded data. Let $F_g : S^{m-1} \to \mathbb{R}$ be defined on the unit sphere in terms of a "contrast function" $g : [0, \infty) \to \mathbb{R}$ as $F_g(u) := \frac{1}{n} \sum_{i=1}^n g(|\langle u, x_{i\cdot} \rangle|)$. This can equivalently be written as

$$F_g(u) = \sum_{i=1}^m w_i g(\tfrac{1}{\sqrt{w_i}} |\langle u, Z_i \rangle|) \ . \qquad (3)$$

In equation (3), F_g takes on a special form of the basis recovery problem presented in equation (1) with the choices $\alpha_i = w_i$ and $\beta_i = \frac{1}{\sqrt{w_i}}$. Due to the special form of these weights, only property P1 is required in order to recover the directions $\{\pm Z_i : i \in [m]\}$:

Theorem 5. *Let $g : [0, \infty) \to \mathbb{R}$ be a continuous function satisfying property P1. Let $F_g : S^{m-1} \to \mathbb{R}$ be defined from g according to equation* (3). *Then, the set $\{\pm Z_i : i \in [m]\}$ is a complete enumeration of the local maxima of F_g.*

Stability analysis: It can be shown that both the embedding structure (Proposition 4) and the local maxima structure of F_g (Theorem 5) are robust to a perturbation from the setting in which G consists of m connected components. We provide such a stability analysis, demonstrating that our algorithms are robust to such perturbations. The precise statements can be found in the long version of this paper.

6 Spectral Clustering Algorithms

Choosing a contrast function. There are many possible choices of contrast g which are admissible for spectral clustering under Theorem 5 including the following:

$$\begin{array}{ll} g_p(t) = |t|^p \text{ where } p \in (2, \infty) & g_{abs}(t) = -|t| \\ g_{ht}(t) = \log\cosh(t) & g_{gau}(t) = e^{-t^2} \\ g_{sig}(t) = -\dfrac{1}{1 + \exp(-|t|)} & \end{array} .$$

In choosing contrasts, it is instructive to first consider the function $g_2(y) = y^2$ (which fails to satisfy property P1 and is thus not admissible). Noting that $F_{g_2}(u) = \sum_{i=1}^m w_i (\frac{1}{\sqrt{w_i}} \langle u, Z_i \rangle)^2 = 1$, we see that F_{g_2} is constant on the unit sphere. We see that the distinguishing power of a contrast function for spectral clustering comes from property P1. Intuitively, "more convex" contrasts g have better resolving power but are also more sensitive to outliers and perturbations of the data. Indeed, if g grows rapidly, a small number of outliers far from the origin could significantly distort the maxima structure of F_g.

Due to this tradeoff, g_{sig} and g_{abs} could be important practical choices for the contrast function. Both $g_{sig}(\sqrt{x})$ and $g_{abs}(\sqrt{x})$ have a strong convexity structure near the origin. As g_{sig} is a bounded function, it should be very robust to perturbations. In comparison, $g_{abs}(\sqrt{t}) = -|\sqrt{t}|$ maintains a stronger convexity structure over a much larger region of its domain and has only a linear rate of growth as $n \to \infty$. This is a much slower growth rate than is present for instances in g_p with $p > 2$.

Algorithms. We now have all the tools needed to create a new class of algorithms for spectral clustering. Given a similarity graph $G = (V, A)$ containing n vertices, define a graph Laplacian $\tilde{L}$ among L, L_{rw}, and L_{sym} (reader's choice). Viewing G as a perturbation of a graph consisting of m connected components, construct $X \in \mathbb{R}^{n \times m}$ such that $x_{\cdot i}$ gives the eigenvector corresponding to the i^{th} smallest eigenvalue of $\tilde{L}$ with scaling $\|x_{\cdot i}\| = \sqrt{n}$.

With X in hand, choose a contrast function g satisfying P1. From g, the function $F_g(u) = \frac{1}{n}\sum_{i=1}^{n} g(\langle u, x_{i\cdot}\rangle)$ is defined on S^{m-1} using the rows of X. The local maxima of F_g correspond to the desired clusters of the graph vertices. Since F_g is a symmetric function, if F_g has a local maximum at u, F_g also has a local maximum at $-u$. However, the directions u and $-u$ correspond to the same line through the origin in $\mathbb{R}^m$ and form an equivalence class, with each such equivalence class corresponding to a cluster.

Our first goal is to find local maxima of F_g corresponding to distinct equivalence classes. We will use that the desired maxima of F_g should be approximately orthogonal to each other. Once we have obtained local maxima $u_1, \ldots, u_m$ of F_g, we cluster the vertices of G by placing vertex i in the j^{th} cluster using the rule $j = \arg\max_\ell |\langle u_\ell, x_{i\cdot}\rangle|$. We sketch two algorithmic ideas in HBROPT and HBRENUM. Here, HBR stands for hidden basis recovery.

```
 1: function HBROPT($X$, $\eta$)
 2:     $C \leftarrow \{\}$
 3:     for $i \leftarrow 1$ to $m$ do
 4:         Draw $u$ from $S^{m-1} \cap \operatorname{span}(C)^\perp$
            uniformly at random.
 5:         repeat
 6:             $u \leftarrow u + \eta(\nabla F_g(u) - uu^T \nabla F_g(u))$
                    $(= u + \eta P_{u^\perp} \nabla F_g(u))$
 7:             $u \leftarrow P_{\operatorname{span}(C)^\perp} u$
 8:             $u \leftarrow \frac{u}{\|u\|}$
 9:         until Convergence
10:         Let $C \leftarrow C \cup \{u\}$
11:     return $C$
```

HBROPT is a form of projected gradient ascent. The parameter η is the learning rate. Each iteration of the repeat-until loop moves u in the direction of steepest ascent. For gradient ascent in $\mathbb{R}^m$, one would expect step 6 of HBROPT to read $u \leftarrow u + \eta \nabla F_g(u)$. However, gradient ascent is being performed for a function F_g defined on the unit sphere, but the gradient described by ∇F_g is for the function F_g with domain $\mathbb{R}^m$. The more expanded formula $\nabla F_g(u) - uu^T \nabla F_g(u)$ is the projection of ∇F_g onto the tangent plane of S^{m-1} at u. This update keeps u near the sphere.

We may draw u uniformly at random from $S^{m-1} \cap \operatorname{span}(C)^\perp$ by first drawing u from S^{m-1} uniformly at random, projecting u onto $\operatorname{span}(C)^\perp$, and then normalizing u. It is important that u stay near the orthogonal complement of $\operatorname{span}(C)$ in order to converge to a new cluster rather than converging to a previously found optimum of F_g. Step 7 enforces this constraint during the update step.

```
1: function HBRENUM($X$, $\delta$)
2:     $C \leftarrow \{\}$
3:     while $|C| < m$ do
4:         $j \leftarrow \arg\max_i \{F_g(\frac{x_{i\cdot}}{\|x_{i\cdot}\|})$ :
                   $\text{angle}(\frac{x_{i\cdot}}{\|x_{i\cdot}\|}, u) > \delta \; \forall u \in C\}$
5:         $C \leftarrow C \cup \{\frac{x_{j\cdot}}{\|x_{j\cdot}\|}\}$
6:     return $C$
```

In contrast to HBROPT, HBRENUM more directly uses the point separation implied by the orthogonality of the approximate cluster centers. Since each embedded data point should be near to a cluster center, the data points themselves are used as test points. Instead of directly enforcing orthogonality between cluster means, a parameter $\delta > 0$ specifies the minimum allowable angle between found cluster means.

By pre-computing the values of $F_g(x_{i\cdot}/\|x_{i\cdot}\|)$ outside of the while loop, HBRENUM can be run in $O(mn^2)$ time. For large similarity graphs, HBRENUM is likely to be slower than HBROPT which takes $O(m^2nt)$ time where t is the average number of iterations to convergence. The number of clusters m cannot exceed (and is usually much smaller than) the number of graph vertices n.

HBRENUM has a couple of nice features which may make it preferable on smaller data sets. Each center found by HBRENUM will always be within a cluster of data points even when the optimization landscape is distorted under perturbation. In addition, the maxima found by HBRENUM are based on a more global outlook, which may be important in the noisy setting.

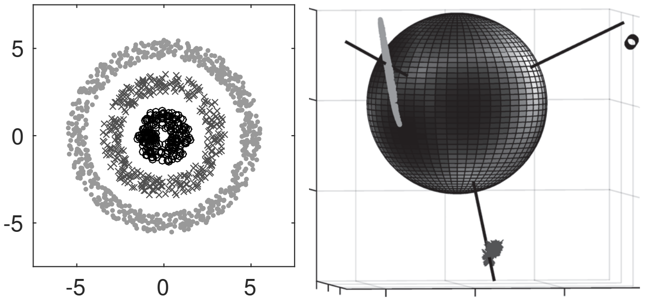
7 Experiments

An Illustrating Example. Figure 1 illustrates our function optimization framework for spectral clustering. In this example, random points p_i were generated from 3 concentric circles: 200 points were drawn uniformly at random from a radius 1 circle, 350 points from a radius 3 circle, and 700 points from a radius 5 circle. The points were then radially perturbed. The generated points are displayed in Figure 1 (a). The similarity matrix A was constructed as $a_{ij} = \exp(-\frac{1}{4}\|p_i - p_j\|^2)$, and the Laplacian embedding was performed using L_{rw}.

Figure 1 (b) depicts the clustering process with the contrast g_{sig} on the resulting embedded points. In this depiction, the embedded data sufficiently encodes the desired basis structure that all local maxima of $F_{g_{sig}}$ correspond to desired clusters. The value of $F_{g_{sig}}$ is displayed by the grayscale heat map on the unit sphere in Figure 1 (b), with lighter shades of gray indicate greater values of $F_{g_{sig}}$. The cluster labels were produced using HBROPT. The rays protruding from the sphere correspond to the basis directions recovered

	oracle-centroids	k-means-cosine	HBRopt					HBRenum				
			g_{abs}	g_{gau}	g_3	g_{ht}	g_{sig}	g_{abs}	g_{gau}	g_3	g_{ht}	g_{sig}
E. coli	79.7	69.0	80.9	81.2	79.3	81.2	80.6	68.7	**81.5**	**81.5**	68.7	**81.5**
Flags	33.2	33.1	**36.8**	34.1	36.6	**36.8**	34.4	34.7	**36.8**	**36.8**	34.7	**36.8**
Glass	49.3	46.8	**47.0**	46.8	**47.0**	**47.0**	46.8	**47.0**	**47.0**	**47.0**	**47.0**	**47.0**
Thyroid Disease	72.4	80.4	**82.4**	81.3	82.2	82.2	81.5	81.8	82.2	82.2	81.8	82.2
Car Evaluation	56.1	36.4	37.0	36.3	36.3	35.2	36.6	49.6	32.3	41.1	**49.9**	41.1
Cell Cycle	74.2	62.7	64.3	64.4	63.8	64.5	64.0	60.1	62.9	**64.8**	61.1	62.7

Table 1: Percentage accuracy of spectral clustering algorithms, with the best performing non-oracle algorithm bolded.



(a) (b)

Figure 1: An illustration of spectral clustering on the concentric circle data. (a) The output of clustering. (b) The embedded data and the contrast function.

by HBRopt, and the recovered labels are indicated by the color and symbol used to display each data point.

Image Segmentation Examples. Spectral clustering was first applied to image segmentation in (Shi and Malik 2000), and it has remained a popular application of spectral clustering. The goal in image segmentation is to divide an image into regions which represent distinct objects or features of the image. Figure 2 illustrates segmentations produced by HBRopt-g_{abs} and spherical k-means on several example images from the BSDS300 test set (Martin et al. 2001). For these images, the similarity matrix was constructed using only color and proximity information.

Stochastic block model with imbalanced clusters. We construct a similarity graph $A = \mathrm{diag}(A_1, A_2, A_3) + E$ where each A_i is a symmetric matrix corresponding to a cluster and E is a small perturbation. We set $A_1 = A_2$ to be 10×10 matrices with entries 0.1. We set A_3 to be a 1000×1000 matrix which is symmetric, approximately 95% sparse with randomly chosen non-zero locations set to 0.001. When performing this experiment 50 times, HBRopt-g_{sig} obtained a mean accuracy of 99.9%. In contrast, spherical k-means with randomly chosen starting points obtained a mean accuracy of only 42.1%. It turns out that splitting the large cluster is in fact optimal in terms of the spherical k-means objective function but leads to poor classification performance. Our method does not suffer from that shortcoming.

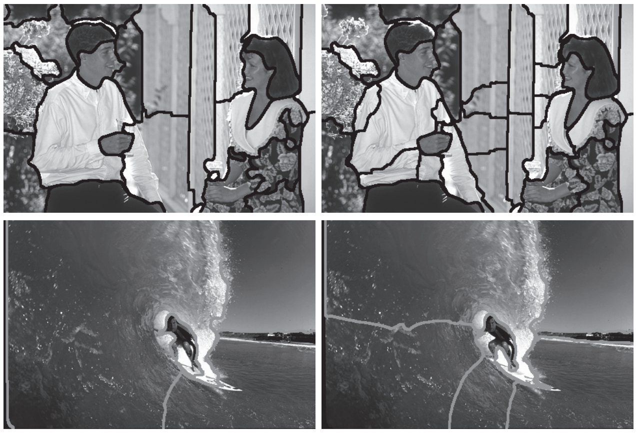


Figure 2: Segmented images. Segmentation using HBRopt-g_{abs} (left panels) compared to k-means (right panels). The borders between segmented regions are marked by black pixels in the top panels and gray pixels in the bottom panels.

Performance Evaluation on UCI datasets. We compare spectral clustering performance on a number of data sets with unbalanced cluster sizes. In particular, the E. coli, Flags, Glass, Thyroid Disease, and Car Evaluation data sets which are part of the UCI machine learning repository (Bache and Lichman 2013) are used. We also use the standardized gene expression data set (Yeung et al. 2001a; 2001b), which is also referred to as Cell Cycle. For the Flags data set, we used religion as the ground truth labels, and for Thyroid Disease, we used the new-thyroid data.

For all data sets, we only used fields for which there were no missing values, we normalized the data such that every field had unit standard deviation, and we constructed the similarity matrix A using a Gaussian kernel $k(y_i, y_j) = \exp(-\alpha \|y_i - y_j\|^2)$. The parameter α was chosen separately for each data set in order to create a good embedding. The choices of α were: 0.25 for E. Coli, 32 for Glass, 32 for Thyroid Disease, 128 for Flags, 0.25 for Car Evaluation, and 0.125 for Cell Cycle.

The spectral embedding was performed using the symmetric normalized Laplacian L_{sym}. Then, the clustering performance of our proposed algorithms HBRopt and HBRenum (implemented with $\delta = 3\pi/8$ radians) were compared with the following baselines:

- oracle-centroids: The ground truth labels are used to set means $\mu_j = \frac{1}{|\mathcal{S}_j|}\sum_{i\in\mathcal{S}_j}\frac{x_{i\cdot}}{\|x_{i\cdot}\|}$ for each $j \in [m]$. Points are assigned to their nearest cluster mean in cosine distance.

- k-means-cosine: Spherical k-means is run with a random initialization of the means, cf. (Ng, Jordan, and Weiss 2002).

We report the clustering accuracy of each algorithm in Table 1. The accuracy is computed using the best matching between the clusters and the true labels. The reported results consist of the mean performance over a set of 25 runs for each algorithm. The number of clusters being searched for was set to the ground truth number of clusters. In most cases, we see improvement in performance over spherical k-means.

8 Acknowledgments

This work was supported by National Science Foundation grants No. IIS 1117707, CCF 1350870 and CCF 1422830.

References

Anderson, J.; Goyal, N.; and Rademacher, L. 2013. Efficient learning of simplices. In Shalev-Shwartz, S., and Steinwart, I., eds., *COLT*, volume 30 of *JMLR Proceedings*, 1020–1045. JMLR.org.

Bach, F. R., and Jordan, M. I. 2006. Learning spectral clustering, with application to speech separation. *Journal of Machine Learning Research* 7:1963–2001.

Bache, K., and Lichman, M. 2013. UCI machine learning repository.

Belkin, M., and Niyogi, P. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* 15(6):1373–1396.

Deuflhard, P.; Huisinga, W.; Fischer, A.; and Schütte, C. 2000. Identification of almost invariant aggregates in reversible nearly uncoupled markov chains. *Linear Algebra and its Applications* 315(1):39–59.

Gritzmann, P., and Klee, V. 1989. On the 0–1-maximization of positive definite quadratic forms. In *Operations Research Proceedings 1988*, 222–227. Springer.

Hsu, D., and Kakade, S. M. 2013. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, 11–20. ACM.

Hyvärinen, A.; Karhunen, J.; and Oja, E. 2004. *Independent component analysis*, volume 46. John Wiley & Sons.

Jain, A. K., and Dubes, R. C. 1988. *Algorithms for clustering data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Kumar, P.; Narasimhan, N.; and Ravindran, B. 2013. Spectral clustering as mapping to a simplex. *2013 ICML workshop on Spectral Learning*.

Martin, D. R.; Fowlkes, C.; Tal, D.; and Malik, J. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 416–425.

Meilă, M., and Shi, J. 2001. A random walks view of spectral segmentation. In *AI and Statistics (AISTATS)*.

Ng, A. Y.; Jordan, M. I.; and Weiss, Y. 2002. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems* 2:849–856.

Rockafellar, R. T. 1997. *Convex analysis*. Princeton Landmarks in Mathematics. Princeton, NJ: Princeton University Press. Reprint of the 1970 original, Princeton Paperbacks.

Shi, J., and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(8):888–905.

Verma, D., and Meilă, M. 2003. A comparison of spectral clustering algorithms. Technical report, University of Washington CSE Department, Seattle, WA 98195-2350. doi=10.1.1.57.6424, Accessed online via CiteSeerx 5 Mar 2014.

Von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and computing* 17(4):395–416.

Weber, M.; Rungsarityotin, W.; and Schliep, A. 2004. *Perron cluster analysis and its connection to graph partitioning for noisy data*. Konrad-Zuse-Zentrum für Informationstechnik Berlin.

Wei, T. 2015. A study of the fixed points and spurious solutions of the deflation-based fastica algorithm. *Neural Computing and Applications* 1–12.

Yeung, K. Y.; Fraley, C.; Murua, A.; Raftery, A. E.; and Ruzzo, W. L. 2001a. Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17(10):977–987.

Yeung, K. Y.; Fraley, C.; Murua, A.; Raftery, A. E.; and Ruzzo, W. L. 2001b. Model-based clustering and data transformations for gene expression data. http://faculty.washington.edu/kayee/model/. Accessed: 20 Jan 2015.

Yu, S. X., and Shi, J. 2003. Multiclass spectral clustering. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 313–319. IEEE.