THE HIDDEN CONVEXITY OF SPECTRAL CLUSTERING

Luis Rademacher, Ohio State University,

Computer Science and Engineering.

Joint work with Mikhail Belkin and James Voss

This talk

- A new approach to multi-way spectral clustering.
- An algorithmic primitive, “hidden basis recovery”, that encompasses algorithmic approaches to problems such as Independent Component Analysis and orthogonal tensor decompositions.

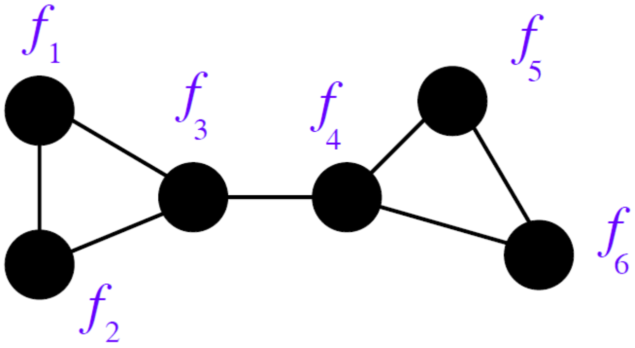
What is spectral clustering?

1. Take a graph.

2. Construct graph Laplacian matrix:
   $L = \mathrm{diag}(\mathrm{degree}) - \mathrm{Adjacency}.$

3. Do something with its bottom eigenvectors to get clusters.

Spectral bi-clustering of a graph



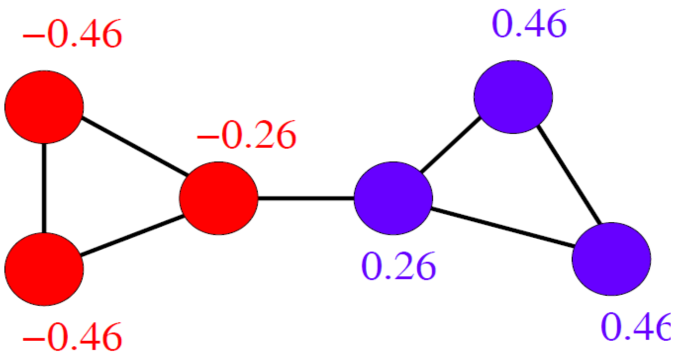
$$\mathbf{L} = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{pmatrix} = D - A$$

$$\underset{S}{\operatorname{argmin}} \sum_{i\in S,\ j\in V-S} w_{ij} = \underset{f_i\in\{-1,1\}}{\operatorname{argmin}} \sum_{i\sim j} (f_i - f_j)^2 = \frac{1}{8} \underset{f_i\in\{-1,1\}}{\operatorname{argmin}} \mathbf{f}^t\,\mathbf{L}\,\mathbf{f}$$

Relaxation of integrality constraint + optimality give **eigenvectors:**

$$Lf = \lambda f$$

Spectral bi-clustering of a graph



$$\mathbf{L} = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{pmatrix}$$

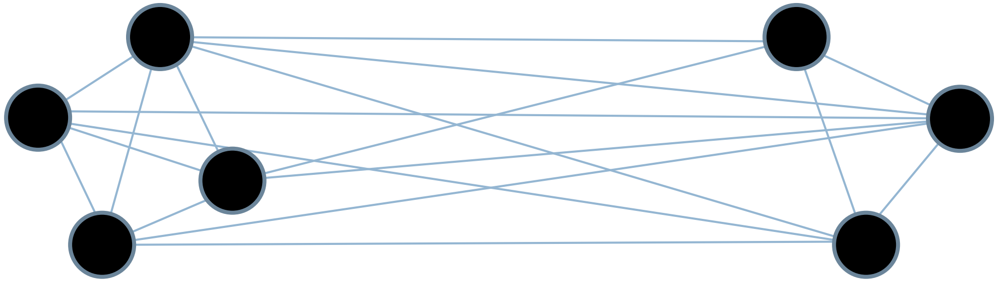
Unnormalized clustering:

$$L\mathbf{e_1} = \lambda_1 \mathbf{e_1} \qquad \mathbf{e_1} = [-0.46, -0.46, -0.26, 0.26, \ 0.46, \ 0.46]$$

(See paper for normalized Laplacian and variations)

Spectral (bi)clustering of data

- Construct a weighted graph, e.g.: $w_{ij} = \exp(-\frac{||x_i - x_j||^2}{t})$



- Second bottom eigenvector of graph Laplacian

$$Le_2 = \lambda_2 e_2$$

- Clusters: $(e_2)_i < 0$; $(e_2)_i \geq 0$

Spectral (bi)clustering of data

- Works well
- Clean and simple
- Some theoretical guarantees
- However, “bi”-clustering is limited.

Multi-way clustering

□ Use several eigenvectors $e_1, e_2, \ldots e_k$

□ Map (Laplacian embedding)

$$\text{Data} \to \mathbb{R}^k$$

$$x_i \to ((e_1)_i, (e_2)_i, \ldots, (e_k)_i)$$

□ Many interesting properties.

For example, eigenvectors of data graph Laplacian (Gaussian weights) approximate eigenfunctions of manifold Laplacian, for manifold data (Belkin, Niyogi 03). Interpretation as diffusion distance (Lafon, Coifman, 05), etc.

Multi-way clustering with k-means

Graph → $\mathbb{R}^k$

$$\phi\colon\ x_i \rightarrow ((e_1)_i, (e_2)_i, \ldots, (e_k)_i)$$

□ Apply k-means in the embedding space.

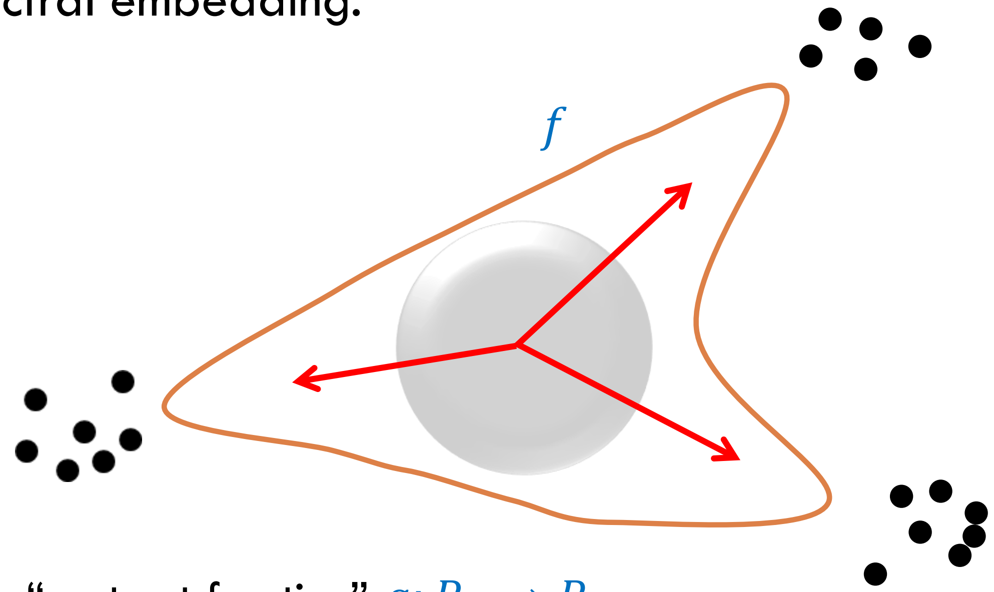
(Shi, Malik 00, Ng, et al, 01, Yu, Shi 03, Bach, Jordan, 06…)

Can be justified as a relaxation of a partition problem.

However initialization dependent. Hard to give guarantees for the algorithm.

Our method

Data after spectral embedding.



Choose allowable "contrast function" $g: R_+ \to R$.

Define $f: \mathcal{S}^{k-1} \to \mathbb{R}$ by $f(v) = \sum_{i=1}^{n} g(|\langle v, \phi(x_i) \rangle|)$

(a sort of "generalized moment")

Claim: all local maxima of f "point" at the clusters.

Allowable contrast functions

Conditions:

- $g(\sqrt{x})$ is strictly convex on $[0, \infty)$.
- $\frac{d}{dx}\left(g(\sqrt{x})\right)\Big|_{0+}$ is 0 or $+\infty$

Some examples:

- $-|x|$
- $|x^p|, p > 2$
- $\exp(-x^2)$
- $\log(\cosh x)$ [from Independent Component Analysis]

Algorithms

Input: $x_1, \dots x_n, k$

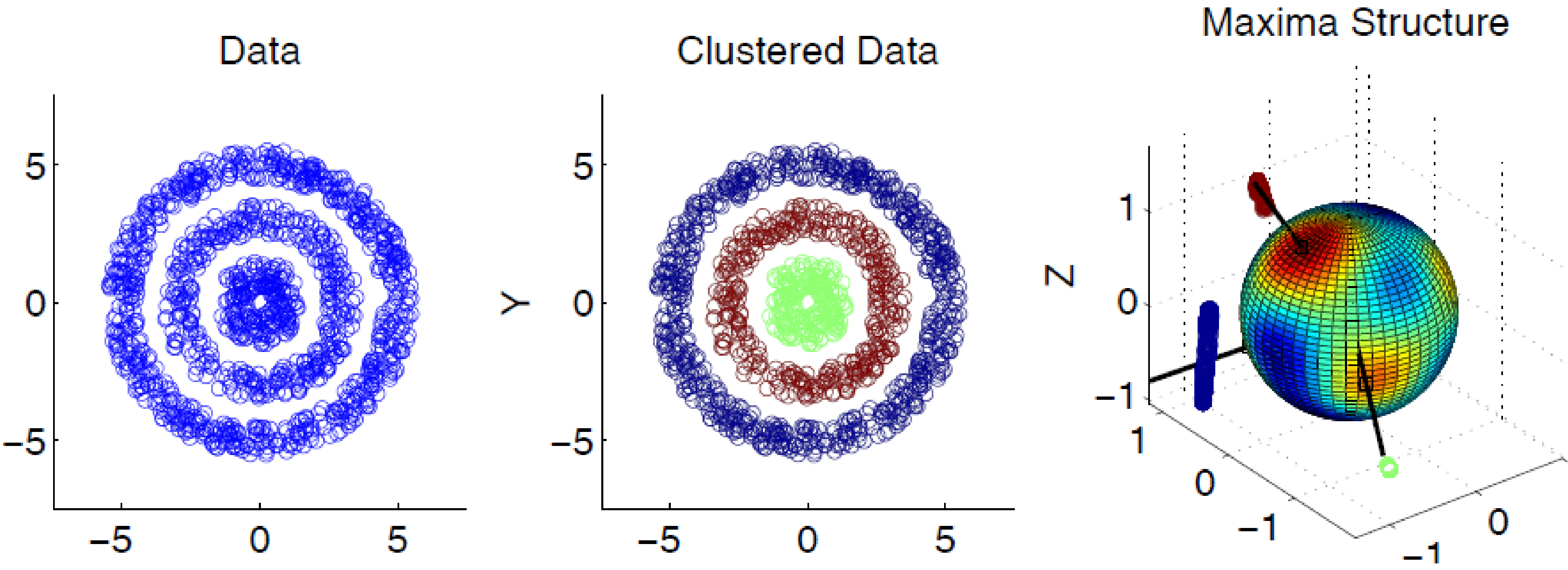
I. Construct graph Laplacian $L = D - A$ and spectral embedding $\phi.$

II. Take $f(v) = \frac{1}{n}\sum_{i=1}^{n} g(\langle v, \phi(x_i)\rangle)$

Algo 1: Gradient ascent for f over a sphere.
Complexity $k^2 n \times \#iterations$

Algo 2: Maximize f over the data points.
Complexity kn^2.

Example

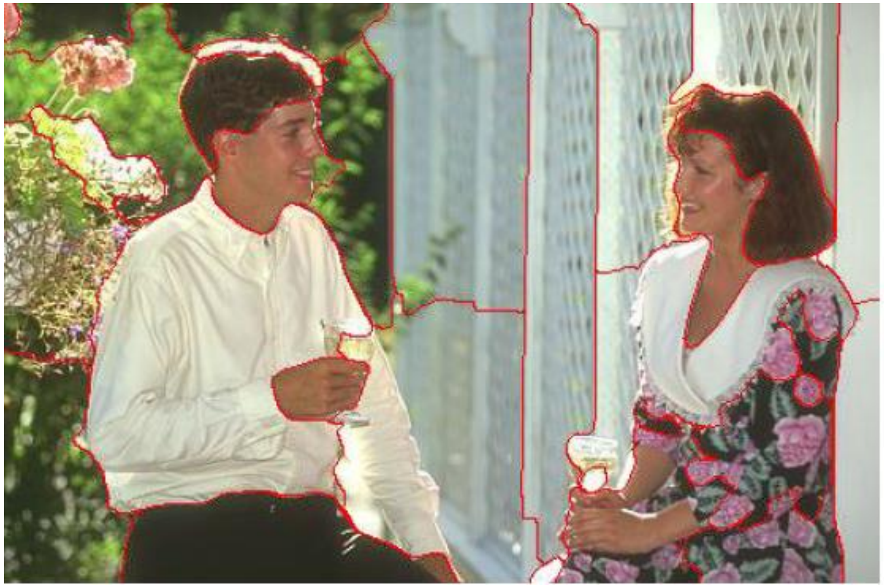


Data

Clustered Data

Maxima Structure

Image segmentation



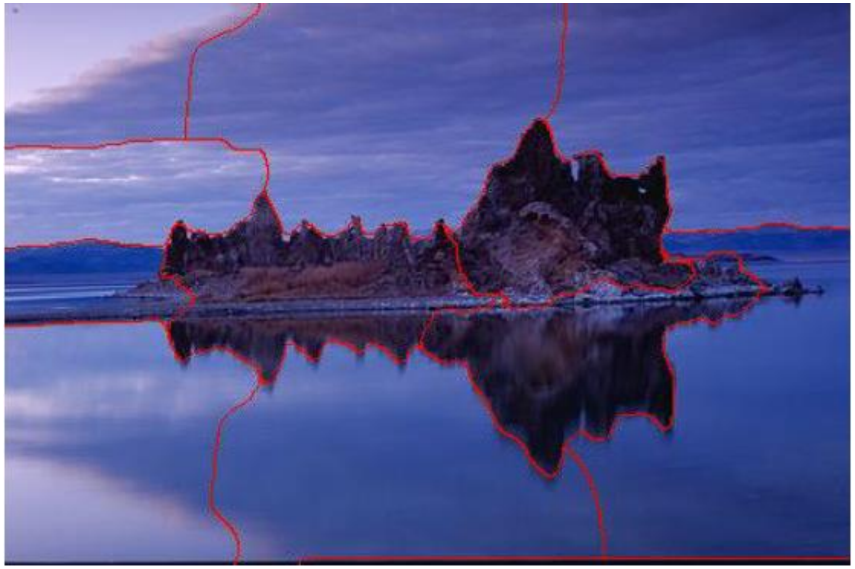


Image segmentation based on the graph of pixel adjacency, weighted by proximity and color similarity. Contrast function $-|x|$.

(Cf. Shi, Malik 97)

Spectral embedding into an orthogonal basis

Claim: k perfect clusters (connected components) means that

$\phi \colon x_i \to ((e_1)_i, (e_2)_i, \ldots, (e_k)_i)$

maps to k orthogonal vectors (hidden basis).

Note that eigenvectors are not uniquely defined (but can be assumed to be orthonormal)

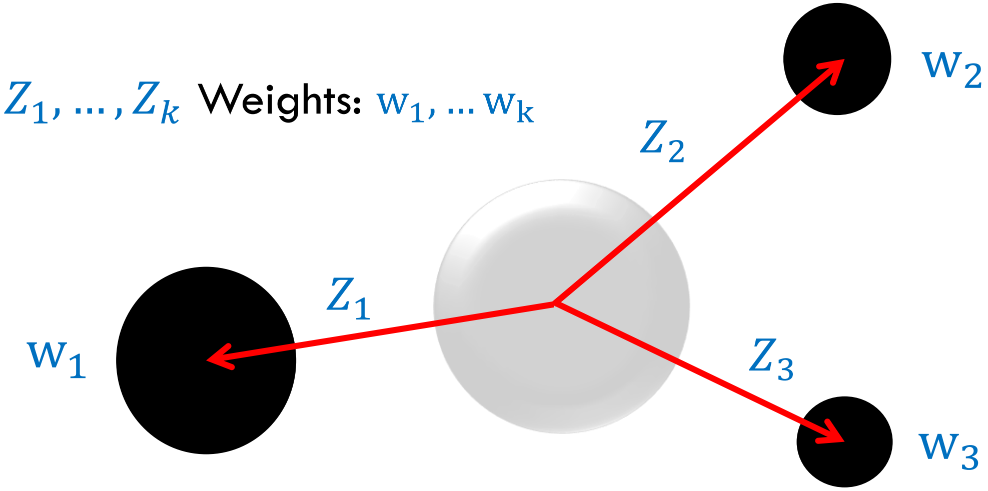
Recovering hidden basis = recovering clusters.

First observed in (Weber, Rungsarityotin, Schliep 04). Also proposed an optimization procedure for hidden basis recovery.

Hidden orthogonal basis structure

Weighted basis vectors.

Basis vectors: $Z_1, \dots, Z_k$ Weights: $\mathrm{w}_1, \dots \mathrm{w}_k$



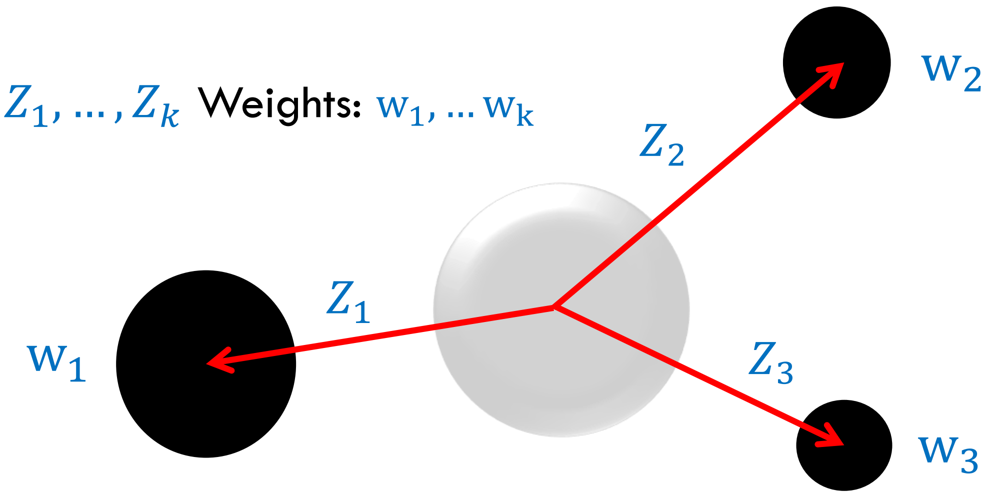
Claim:

1. $\langle Z_i, Z_j \rangle = 0, \; i \neq j$
2. $w_i = \frac{n_i}{n}$
3. $\langle Z_i, Z_i \rangle = \frac{n}{n_i}$

Hidden orthogonal basis structure

Weighted basis vectors.

Basis vectors: $Z_1, \dots, Z_k$ Weights: $\mathrm{w}_1, \dots \mathrm{w}_\mathrm{k}$



Key identity:

$$f(v) = \frac{1}{n}\sum_{i=1}^{n} g(|\langle \mathrm{v}, \phi(x_i)\rangle|) = \sum_{i=1}^{k} w_i g(|\langle \mathrm{v}, Z_i\rangle|)$$

Geometric recovery

Let $f(v) = \sum_{i=1}^{k} w_i g(|\langle \mathrm{v}, Z_i \rangle|)$. $w_i, Z_i, i = 1..k$ orthogonal, weighted hidden basis.

Question: Conditions on g so that set of local maxima of f is $\pm \frac{Z_i}{\|Z_i\|}$, the hidden basis?

Conditions:

P1. $g(\sqrt{x})$ is strictly convex on $[0, \infty)$.

P2. $\frac{d}{dx}\left(g(\sqrt{x})\right)\Big|_{0+}$ is 0 or $+\infty$.

Theorem 1. [Sufficiency]

If both P1 and P2 are satisfied, $\pm \frac{Z_i}{\|Z_i\|}$ is the complete enumeration of the local maxima of f.

Theorem 2. [Necessity]

a. If P1 does not hold for g, then for some (w_i, Z_i) there is a local maximum different from $\pm \frac{Z_i}{\|Z_i\|}$

b. if P1 holds but P2 does not hold, then for some (w_i, Z_i) one of $\pm \frac{Z_i}{\|Z_i\|}$ is not a local maximum.

Geometric recovery

Recall for spectral clustering:

1. $\langle Z_i, Z_j \rangle = 0, \;\; i \neq j$
2. $\langle Z_i, Z_i \rangle = \frac{n}{n_i}$
3. $w_i = \frac{n_i}{n}$

Theorem 3. If P1 and, additionally, properties 2-3 hold, then

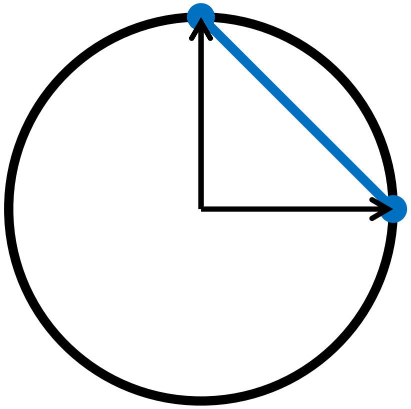
$\pm \frac{Z_i}{\|Z_i\|}$ is a complete enumeration of the local maxima.

Hidden convexity

Analysis via change of variable:

$\tau\colon\ (x_1, \dots, x_k) \to \sqrt{x_1}, \dots, \sqrt{x_k}$

simplex → sphere



Write f over standard simplex in the basis corresponding to $\hat{z}_i = z_i / \|z_i\|$:

[P1, $g(\sqrt{x})$ is strictly convex on $[0, \infty)$] implies

$f\big(\tau(\mathrm{v})\big) = \sum_{i=1}^{k} w_i g\Big(\sqrt{\langle \mathrm{v}, z_i \rangle}\Big)$

is a ***strictly convex function.***

Max over sphere → Max over simplex

Maximum principle: Local maxima of strictly convex function only at extreme points of simplex.

Hidden basis recovery as an algorithmic primitive

More generally: Hidden Basis Recovery

- Hidden orthonormal basis: $z_1, \dots, z_k$.
- "Basis Encoding Function" (BEF):

$$F(u) = \sum_{i=1}^{k} g_i(u \cdot z_i)$$

- Problem: given evaluation access to F and derivatives, find z_is.
- Example: spectral clustering, with spectral embedding (x_j):

$$F(u) = \sum_{j=1}^{n} g(u \cdot x_j) = \sum_{i=1}^{k} g_i(u \cdot z_i)$$

with $g_i(t) = n_i g(b_i t)$,
b_i =length of embedded vectors and
n_i =size of ith cluster.

More examples

Orthogonal tensor decomposition:

Given $T = T_{jlmt} = \sum_i w_i \, z_i \otimes \; z_i \otimes z_i \otimes z_i$, Basis

Encoding Function is

$$F(u) = T(u, u, u, u) = \sum_i w_i \, (u \cdot z_i)^4$$

$$= \sum_i g_i(u \cdot z_i)$$

with $g_i(t) = w_i t^4$.

What makes tensor power iteration work?

Multi-linear algebra can be replaced by another explanation:

- □ “Hidden convexity”: $g_i(\sqrt{t}) = w_i t^2$ is strictly convex.
- □ As in “hidden convexity” for spectral clustering, $\{\pm$hidden basis$\}$ is complete enumeration of local maxima of $T(u,u,u,u)$ over sphere.
- □ Power iteration can be interpreted as projected gradient ascent with an (automatic) adaptive step size.

More examples: ICA

- Independent Component Analysis [Comon]: Given samples from x given by $x = As$, with
  - x, s d-dim. random vectors,
  - s with independent coordinates,
  - A square invertible matrix.

    Recover A.
- After whitening/isotropy, can assume A is unitary (i.e. columns are orthonormal basis).
- BEF: $F(u) = \kappa_4(u \cdot x) = \sum_i \kappa_4(s_i)(u \cdot A_i)^4$ with $g_i(t) = \kappa_4(s_i)t^4$.
  (where κ_4 is the fourth cumulant, here $\kappa_4(T) = E(T^4) - 3$)

Our results: Practical algorithm to find hidden basis

- “Gradient Iteration”: a fixed point iteration of the gradient:

$$u_{new} = \frac{\nabla F(u_{old})}{\|\nabla F(u_{old})\|}.$$

Gradient iteration

- “Gradient Iteration” is an extension of tensor power iteration to a functional setting without multi-linear algebra:
For example: $F(u) = T(u, u, u, u)$, then tensor power iteration is $u_{new} = \frac{T(u,u,u,\cdot)}{\|T(u,u,u,\cdot)\|}$
Gradient iteration is $u_{new} = \frac{\nabla F(u)}{\|\nabla F(u)\|}$
with $\nabla F(u) = c\, T(u, u, u,\cdot)$

Algorithm to find hidden basis

Under hidden convexity assumptions on contrasts g_i:

- **Thm:** The set of stable fixed points of gradient iteration is exactly $\{\pm z_i\}$, the hidden basis vectors.
- **Thm:** A provably correct refinement of gradient iteration can enumerate hidden basis vectors efficiently, **even under additive perturbation of basis encoding function.**
- We generalize conditions where power iteration has superlinear convergence:
  Thm: If $t \mapsto g_i\left(\sqrt[r]{t}\right)$ is convex, then convergence of gradient iteration is of order $r - 1$.

More examples

- Parameter estimation for Spherical Gaussian mixture model (inspired by [Hsu Kakade]).

An interesting phenomenon

- "Blessing of dimensionality" for Gaussian Mixture Model with identical components:
  - Estimation generically polynomial time in the smoothed analysis sense for mixtures in R^d with d^m components (for any fixed m).
  - Estimation generically hard in low dimension: generic pairs of sets of k means in R^d support a pair mixtures that are within total variation distance $e^{-k^{1/d}}$.
    Implies sample complexity is at least $e^{k^{1/d}}$.
- In other words, for GMM, "*dimensionality reduction considered harmful*".

Summary

- A new algorithm for multi-way spectral clustering.
- An algorithmic primitive, “hidden basis recovery”, where data is encoded by functions, generalizing tensors.
- An efficient algorithm, Gradient Iteration. Complete characterization of admissible contrasts for spectral clustering.

More details:

- “The hidden convexity of spectral clustering”, arxiv, with M. Belkin and J. Voss.
- “The more, the merrier: the blessing of dimensionality for learning large Gaussian mixtures” COLT 2014, with J. Anderson, M. Belkin, N. Goyal, J. Voss.
- “Learning a hidden basis through imperfect measurements: An algorithmic primitive”, in preparation, with M. Belkin and J. Voss.