THE HIDDEN CONVEXITY OF SPECTRAL CLUSTERING

Mikhail Belkin, Ohio State University,
Department of Computer Science and Engineering,
Department of Statistics

Joint work with Luis Rademacher and James Voss

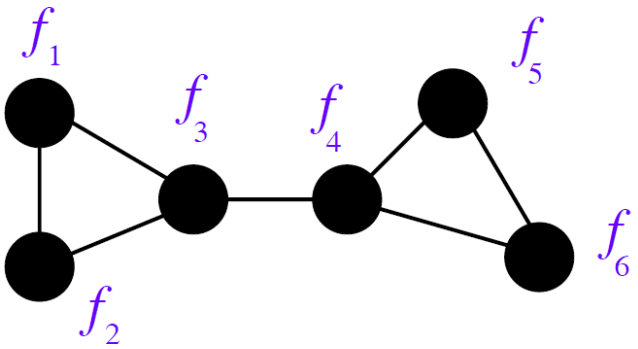
This talk

- A new approach to multiway spectral clustering.
- Based on ICA-like contrast functions. Complete description of admissible contrasts.
- General simplex recovery.
- ICA as simplex recovery.

What is spectral clustering?

1. Take a graph.
2. Construct graph Laplacian matrix.
3. Do something with its bottom eigenvectors to get clusters.

Spectral clustering of a graph



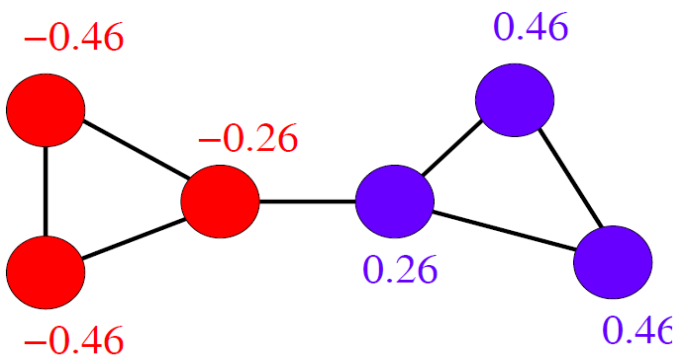
$$\mathbf{L} = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{pmatrix}$$

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

Relaxation gives eigenvectors.

$$\mathbf{L}v = \lambda v$$

Spectral clustering on 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:

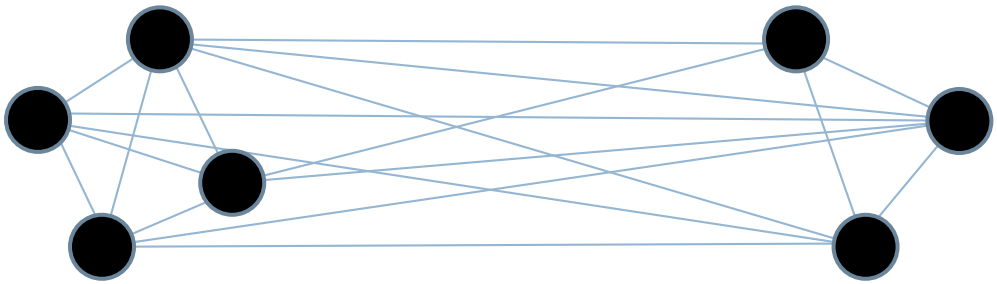
$$L\mathbf{e_1} = \lambda_1 \mathbf{e_1} \qquad \mathbf{e_1} = [-0.46, -0.46, -0.26, 0.26, 0.46, 0.46]$$

Normalized clustering:

$$L\mathbf{e_1} = \lambda_1 D\mathbf{e_1} \qquad \mathbf{e_1} = [-0.31, -0.31, -0.18, 0.18, 0.31, 0.31]$$

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 not that useful.

Multiway 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.

Multiway clustering with k-means

Graph $\rightarrow \mathbb{R}^k$

ϕ: $x_i \rightarrow ((e_1)_i, (e_2)_i, \ldots, (e_k)_i)$

- Apply k-means in the embedding space.

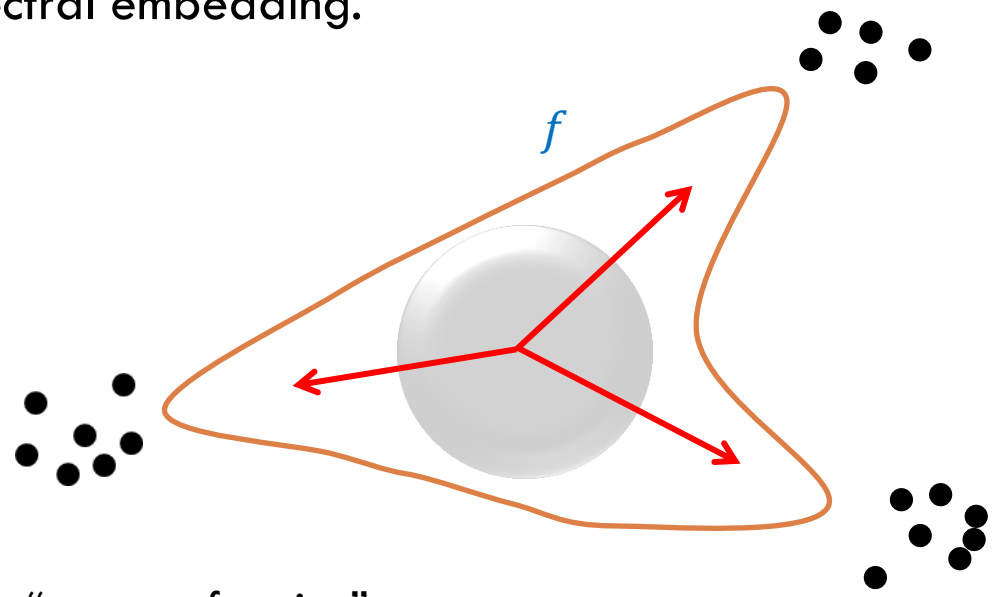
(Shi, Malik 00, Ng, et al, 01, Yu, Shi 03, Bach, Jordan, 06…)

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.

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

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

Allowable contrast functions

Conditions:

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

Equivalently $\frac{g'(x)}{x}$ strictly increasing on $(0, \infty)$.

Some examples:

$-|x|$

$|x^p|, \mathrm{p} > 2$

$\exp(-x^2)$

$\log(\cosh x)$ [ICA]

Algorithms

Input: $x_1, \dots x_n, k$

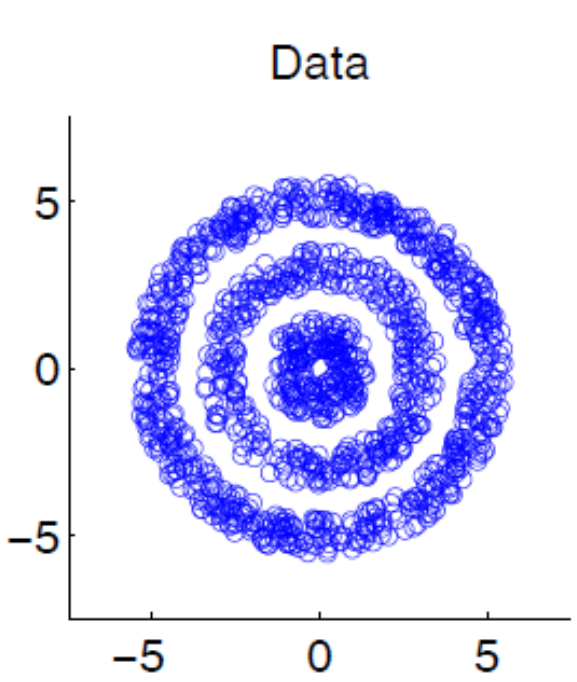
I. Construct graph Laplacian L and spectral embedding $\phi = (e_1, \dots, e_k)$

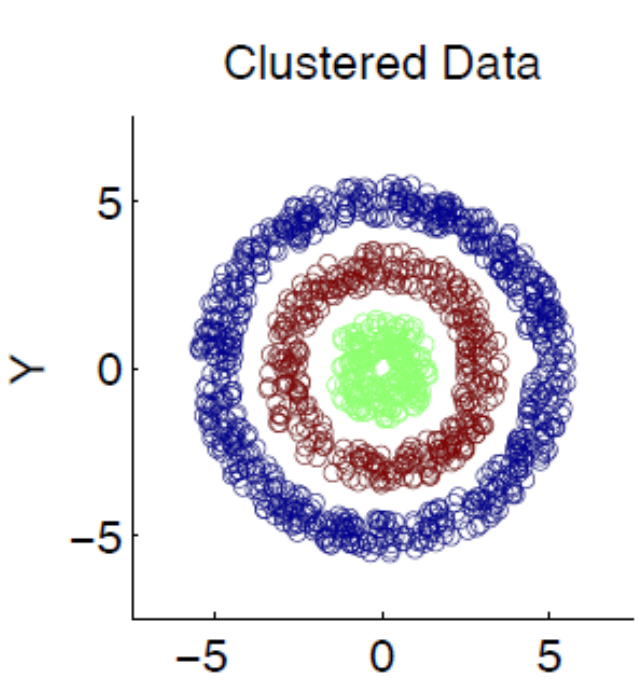
II. Take $f(v) = \frac{1}{n}\sum_{i=1}^{n} g(\langle \mathrm{v}, \phi(x_i) \rangle)$

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
5
0
−5
−5
0
5


Clustered Data
Y
5
0
−5
−5
0
5

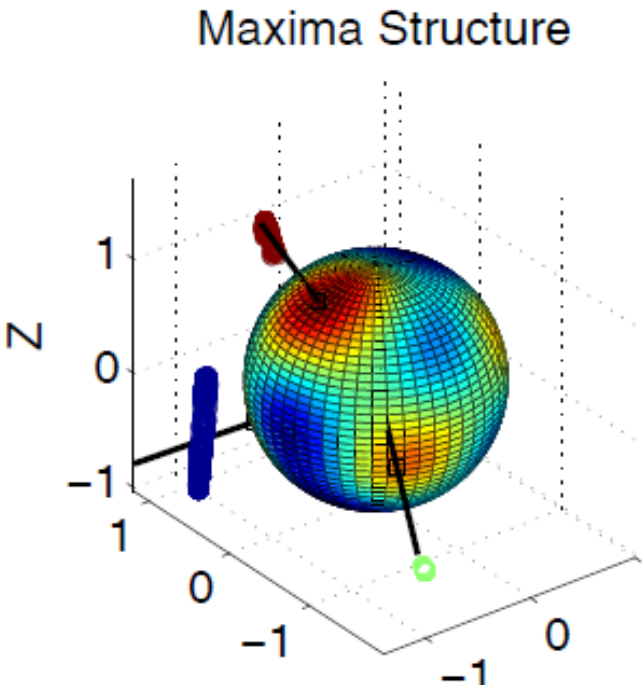

Maxima Structure
Z
1
0
−1
1
0
−1
−1
0

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 a simplex

Eigenvectors corresponding to eigenvalue zero (un-normalized Laplacian) are locally constant functions.

k perfect clusters means that $\phi\colon x_i \to ((e_1)_i, (e_2)_i, \ldots, (e_k)_i)$

maps to a $k-1$ dimensional simplex

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

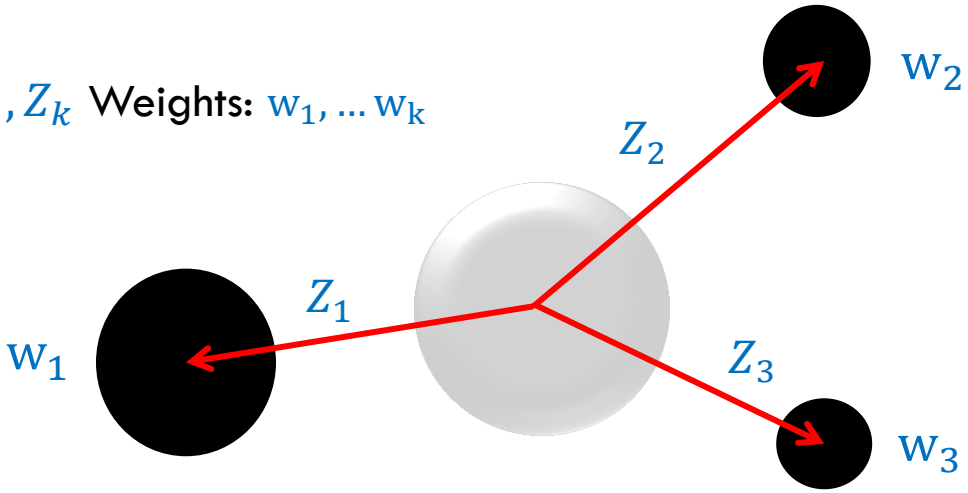
Recovering simplex vertices = recovering clusters.

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

Simplex structure

Weighted discrete simplex.

Vertices: $Z_1, \ldots, Z_k$ Weights: $w_1, \ldots w_k$



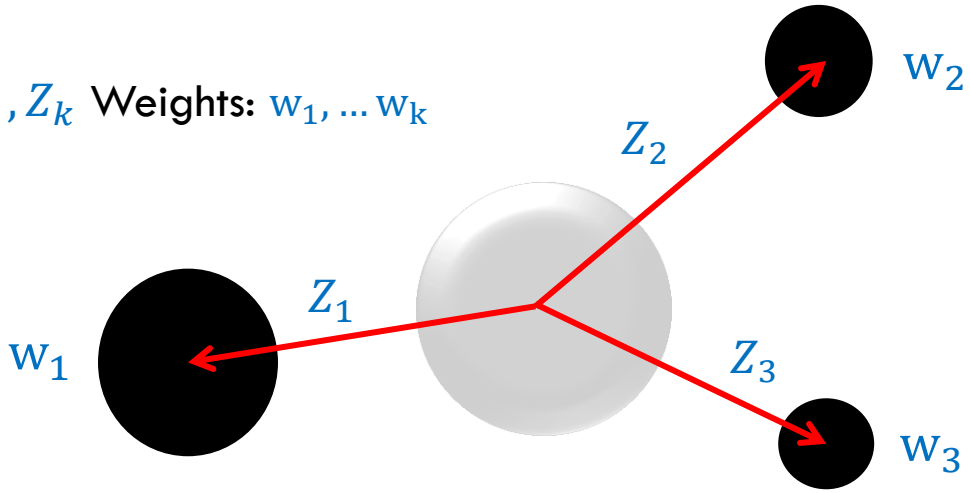
Claim:

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

Simplex structure

Weighted discrete simplex.

Vertices: $Z_1, \ldots, Z_k$ Weights: $w_1, \ldots w_k$



Key identity:

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

Geometric recovery

Theorem 1.

If g is admissible (symmetric and $g(\sqrt{x})$ is convex) and (Z_i, w_i) is a discrete orthogonal simplex. Then the only possible local maxima of f are $\pm \frac{Z_i}{\|Z_i\|}$.

Theorem 2.

The admissibility condition on g is necessary.

Geometric recovery

Recall properties:

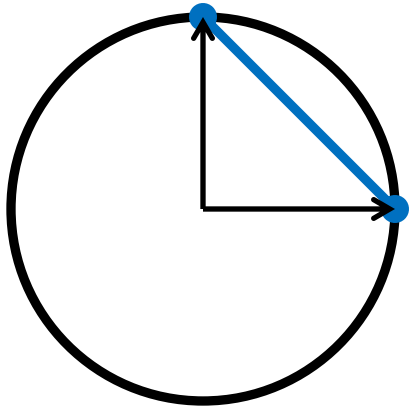
1. $\langle Z_i, Z_j \rangle = 0, \ i \neq j$
2. $\langle Z_i, Z_i \rangle = \frac{n}{n_i}$
3. $w_i = \frac{n_i}{n}$

Theorem 3. If, additionally, properties 1-3 hold, then $\pm \frac{Z_i}{\|Z_i\|}$ is a complete enumeration of the local maxima.

Hidden convexity

τ: $(x_1, \ldots, x_k) \rightarrow \sqrt{x_1}, \ldots, \sqrt{x_k}$

simplex → sphere



Choose the coordinates corresponding to Z_i.

$\mathrm{f}(\tau^{-1}(\mathrm{v})) = \sum_{i=1}^{k} w_i g\left(\sqrt{\langle \mathrm{v}, Z_i \rangle}\right)$ is a sum of convex functions.

Max over sphere → Max over simplex

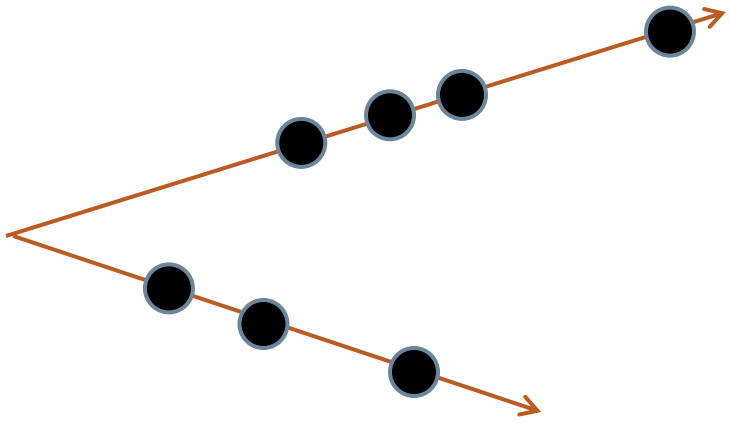
Local maxima of convex functions at extreme points.

Normalized Laplacian

- Normalized Laplacian $D^{-\frac{1}{2}}LD^{-\frac{1}{2}}$

- Eigenvectors with eigenvalue 0 not locally constant.

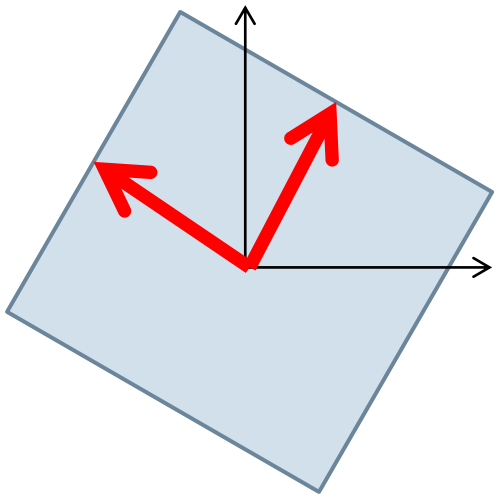
Spectral map to union of 1-D intervals in $\mathbb{R}^k$.



The result still holds!

Independent Component Analysis (Comon, 94)

- Recover independent variables by observing linear combinations. (Cocktail party problem.)
- (whitening) Can assume covariance matrix is I.
- Reduces to recovering a rotation.



Cumulants

- Cumulant generating function $\mathrm{h(t)} = \log(E\exp(tx))$
- $h(t) = \sum \frac{1}{l!} k_l t^l$
- Polynomial in moments:

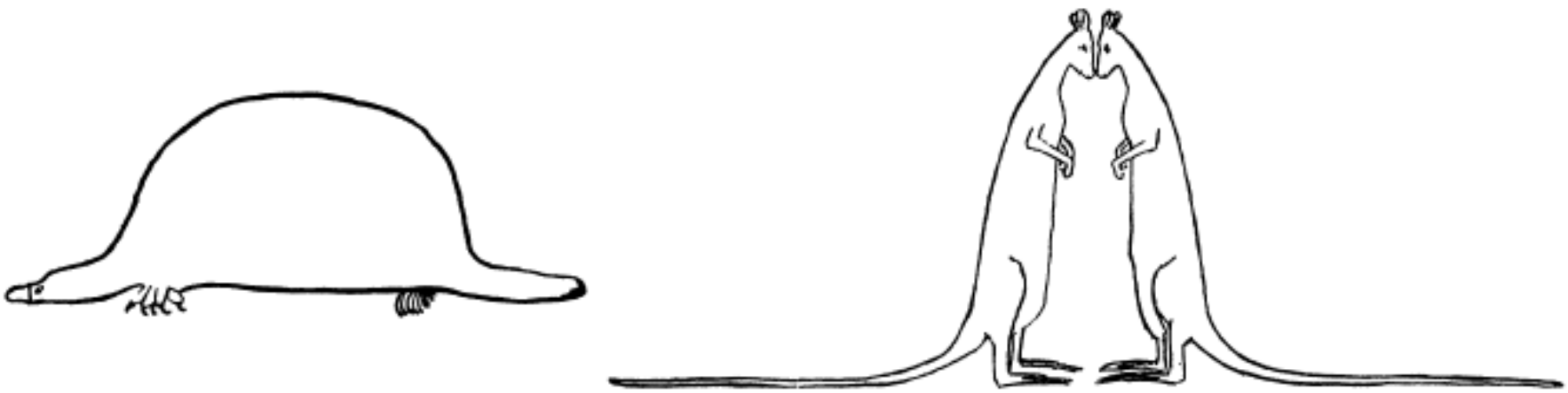
$$k_2 = \mu_2, k_3 = \mu_3, k_4 = \mu_4 - 3\mu_2 \ldots$$

- Key property: $k_l(aX + bY) = a^l k_l(X) + b^l k_l(Y)$ for independent X, Y.

Recent rebirth of moment/cumulant methods in Theoretical CS and Machine Learning. E.g. Hsu, Kakade,12 for learning Gaussian mixtures.

Kurtosis

* In case any of my readers may be unfamiliar with the term "kurtosis" we may define mesokurtic as "having β_2 equal to 3," while platykurtic curves have $\beta_2 < 3$ and leptokurtic > 3. The important property which follows from this is that platykurtic curves have shorter "tails" than the

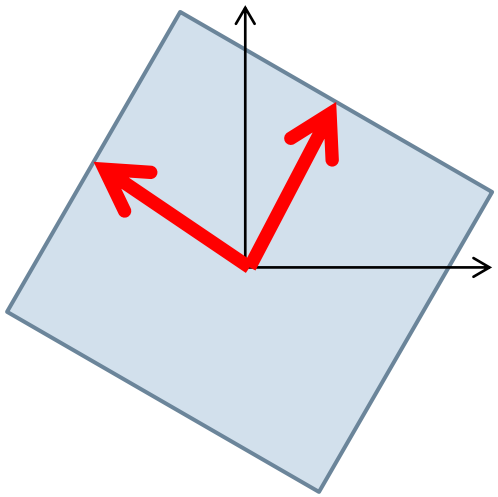


normal curve of error and leptokurtic longer "tails." I myself bear in mind the meaning of the words by the above *memoria technica*, where the first figure represents platypus, and the second kangaroos, noted for "lepping," though, perhaps, with equal reason they should be hares!

Student's drawing , 1927, from the web site of Karl L. Wuensch.

Independent Component Analysis

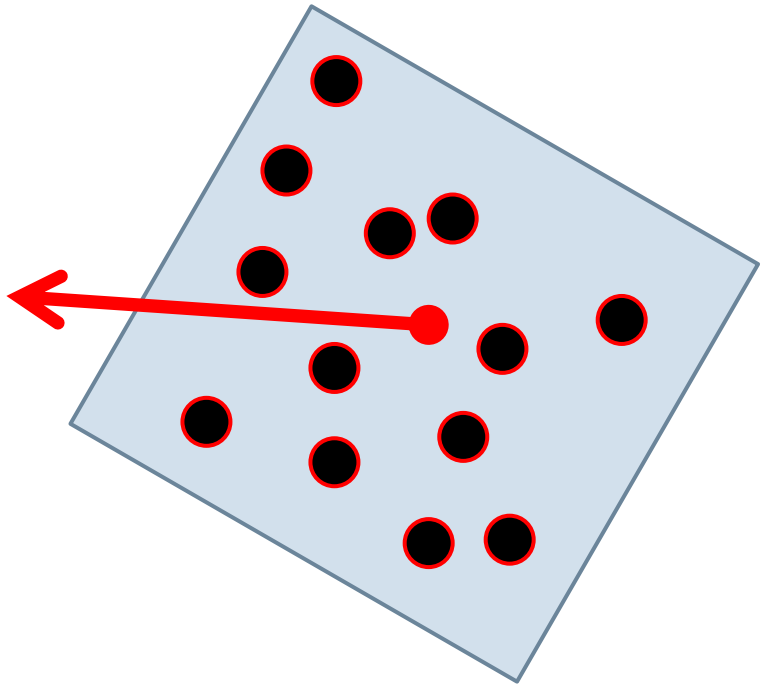
- Cumulant generating function $\mathrm{h(t)} = \log(E\exp(tx))$
- $h(t) = \sum \frac{1}{l!} k_l t^l$
- Define $f(v) = E_x\, k_l(\langle \mathrm{v}, x \rangle)\,, \mathrm{l} > 2$



Theorem: the only maxima of $|f(v)|$ correspond to the original coordinate directions.

Estimating from data

- $f(v) = E_x\, k_l(\langle \mathrm{v}, x \rangle) \approx \frac{1}{n} \sum k_l(\langle \mathrm{v}, x_i \rangle)$



Other contrast functions are also used in practice but only cumulants are guaranteed to work.

ICA as simplex learning

From cumulant properties: $\mathrm{v} = \sum \mathrm{a}_i e_i$

$k_l(v) = k_l(\sum \mathrm{a}_i e_i) = \sum \mathrm{a}_i^l k_l(e_i)$

Put $w_i = k_l(e_i)$, $g(x) = x^l$, $Z_i = e_i$

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

The simplex recovery theorem guarantees ICA recovery as x^l
Is an admissible contrast for $l > 2$.

Cheating slightly – cumulants can be negative. Can be fixed.

Summary

- New algorithms for spectral clustering.
- Simple and (nearly) initialization independent.
- Complete characterization for a large (many more than ICA) class of provable contrast functions. Many are “nice”.
- Simplex learning is of independent interest.