Nonlinear Eigenproblems in Data Analysis and Graph Partitioning

Matthias Hein

Department of Mathematics and Computer Science
Saarland University, Saarbrücken, Germany

Minisymposium: Modern matrix methods for large scale data and networks
SIAM Conference on Applied Linear Algebra, Valencia
19.06.2012

Motivation: Eigenvalue problems are abundant in data analysis

- **Principal Component Analysis:**
  Largest eigenvectors of covariance matrix of the data
  Usage: Denoising by projection onto largest eigenvectors.
- **Spectral Clustering:**
  Second smallest eigenvector of the graph Laplacian
  Usage: Graph partitioning using thresholded eigenvector.
- **Latent Semantic Analysis:**
  Singular value decomposition of term-document matrix
  Usage: Recover underlying latent semantic structure.
- **Many more ... !**

The Symmetric Linear Eigenproblem

Generalized Symmetric Linear Eigenproblem:

Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric and B positive definite. Then

$$Ax = \frac{\langle x, Ax\rangle}{\langle x, Bx\rangle}\, Bx \quad \Longleftrightarrow \quad x \text{ critical point of } \quad \frac{\langle x, Ax\rangle}{\langle x, Bx\rangle}.$$

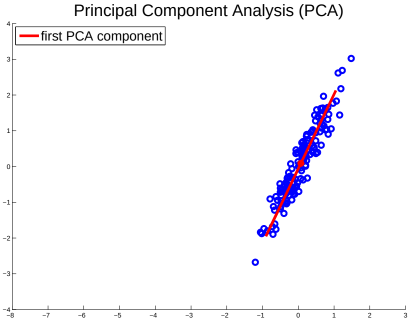
Variational Principle:

Courant-Fischer min-max theorem yields n eigenvalues:

$$\lambda_m = \min_{U_m \in \mathcal{U}_m} \max_{x \in U_m} \frac{\langle x, Ax\rangle}{\langle x, Bx\rangle}, \quad m = 1, \ldots, n,$$

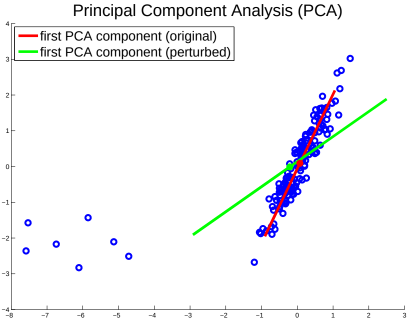
where $\mathcal{U}_m$ is the class of all m-dimensional subspaces of $\mathbb{R}^n$.

Critical point theory for ratios of quadratic functions



	PCA
Type of eigenproblem	**Linear**
Ratio	$\dfrac{\sum_{i=1}^{n} \left\langle w, X_i - \frac{1}{n}\sum_{j=1}^{n} X_j \right\rangle^2}{\|w\|_2^2}$

Robust PCA

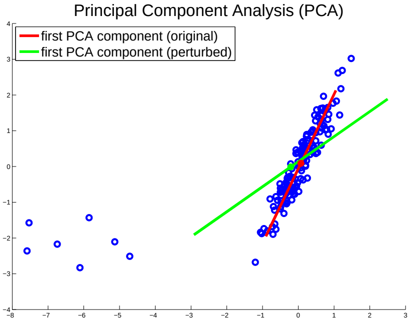


Source of outliers

- noisy data
- adversarial manipulation

	PCA
Type of eigenproblem	**Linear**
Ratio	$\dfrac{\sum_{i=1}^{n} \left\langle w, X_i - \frac{1}{n}\sum_{j=1}^{n} X_j \right\rangle^2}{\|w\|_2^2}$
Robustness	**no**

Robust PCA



Source of outliers

- noisy data
- adversarial manipulation

	PCA		
Type of eigenproblem	**Linear**		**Nonlinear**
Ratio	$\dfrac{\sum_{i=1}^{n} \left\langle w, X_i - \frac{1}{n}\sum_{j=1}^{n} X_j \right\rangle^2}{\lVert w \rVert_2^2}$	$\Rightarrow$	$\dfrac{V\big(\langle w, X_1\rangle, \ldots, \langle w, X_n\rangle\big)}{\lVert w \rVert_2}$
Robustness	**no**		**yes**

The Symmetric Linear Eigenproblem

Pros:

- Fast solvers available

Cons:

- Restriction to ratio of quadratic functionals $\Longrightarrow$ limited modeling abilities
- Quadratic functionals are non-robust against outliers (PCA).
- Quadratic functionals cannot induce eigenvectors which are sparse.

Idea:
Replace quadratic functionals by convex p-homogeneous functions !

The Nonlinear Eigenproblem

(Homogeneous) Nonlinear Eigenproblem:

Let $R, S : \mathbb{R}^n \to \mathbb{R}$ be convex, even and p-homogeneous $(R(\gamma x) = |\gamma|^p R(x))$ and $S(x) = 0 \Leftrightarrow x = 0$. Then

$$0 \in \partial R(x) - \frac{R(x)}{S(x)} \partial S(x) \quad \Longleftarrow \quad x \text{ critical point of } \quad \frac{R(x)}{S(x)}.$$

Variational Principle:

Lusternik-Schnirelmann min-max theorem yields n nonlinear eigenvalues:

$$\lambda_m = \min_{K \in \mathcal{K}_m} \max_{x \in K} \frac{R(x)}{S(x)}, \quad m = 1, \ldots, n,$$

where $\mathcal{K}_m$ is the class of all compact symmetric subsets of $\{x \in \mathbb{R}^n \,|\, S(x) > 0\}$ with Krasnoselskii genus greater or equal to m.

New: general more than n eigenvectors exist.

The Nonlinear Eigenproblem II

Pros:

- Stronger modeling power using non-quadratic functions R and S
- Specific properties of eigenvectors like **robustness** against outliers or **sparsity** can be induced by nonsmooth choices of S respectively R.

Challenges:

- Optimization problems for eigenproblems are typically **nonconvex** and **nonsmooth**.
- Need for new efficient algorithms !

Inverse Power Method for Linear Eigenproblems

$$A f_{k+1} = B\, f_k \quad \Longleftrightarrow \quad f_{k+1} = \underset{u \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2} \langle u, Au \rangle - \left\langle u, Bf^k \right\rangle \right\}$$

Sequence f_k converges to smallest eigenvector of generalized eigenproblem.

Inverse Power Method for Nonlinear Eigenproblems (H.,B.(2010))

$p > 1$	$p = 1$
$g_{k+1} = \underset{u \in \mathbb{R}^n}{\arg\min} \{ R(u) - \langle u, s(f_k) \rangle \}$	$f_{k+1} = \underset{\\|u\\|_2 \leq 1}{\arg\min} \{ R(u) - \lambda_k \langle u, s(f_k) \rangle \}$
$f_{k+1} = g_{k+1} / S(g_{k+1})^{1/p}$	
$s(f_{k+1}) \in \partial S(f_{k+1})$	$s(f_{k+1}) \in \partial S(f_{k+1})$
$\lambda_{k+1} = \frac{R(f_{k+1})}{S(f_{k+1})}$	$\lambda_{k+1} = \frac{R(f_{k+1})}{S(f_{k+1})}$

Theorem (Hein, Bühler (2010)): It holds either

$$\lambda_{k+1} < \lambda_k$$

or $\lambda_{k+1} = \lambda_k$ and the sequence terminates. Moreover, for every cluster point f^* of the sequence f_k one has

$$0 \in \partial R(f^*) - \lambda^* \, \partial S(f^*), \quad \text{where} \quad \lambda^* = \frac{R(f^*)}{S(f^*)}.$$

Guarantees:

- monotonic descent method
- convergence guaranteed to some nonlinear eigenvector **but** not necessarily the one associated with the smallest eigenvalue.

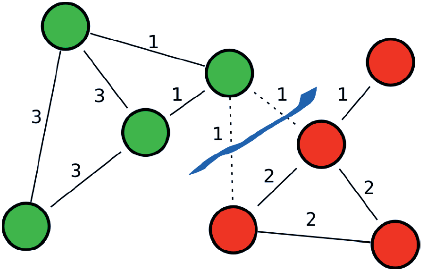
Benefits of Nonlinear Eigenproblems

	Linear EP	Nonlinear EP
Modeling power	low	high
Relaxation of combinatorial problems	loose	tight

Cheeger cut: $(C, \overline{C})$ is a partition of the weighted, undirected graph

$$\phi(C) = \frac{\mathrm{cut}(C, \overline{C})}{\min\{|C|, |\overline{C}|\}}, \quad \text{where} \quad \mathrm{cut}(A, B) = \sum_{i \in A, j \in B} w_{ij}$$

Optimal Cheeger cut, $\phi^* = \min_C \phi(C)$, is NP-hard



Clustering/Community detection

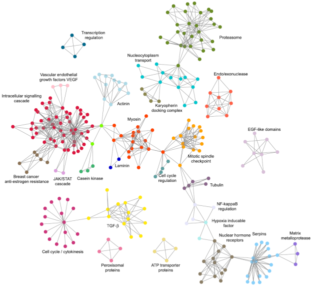
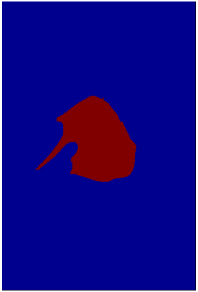
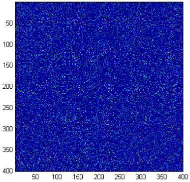
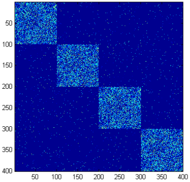


Image Segmentation



Parallel Computing (Matrix Reordering)





Relaxation of Cheeger Cut Problem

Relaxation into semi-definite program with $|V|^3$ constraints:
Best known (worst case) approximation guarantee: $O(\sqrt{\log |V|})$.

Spectral Relaxation based on graph Laplacian L

$$L = D - W,$$

Isoperimetric inequality (Alon, Milman (1984))

$$\frac{(\phi^*)^2}{2 \max_i d_i} \leq \lambda_2(L) \leq 2\phi^*.$$

- there are graphs known which realize lower bound
- bipartitioning obtained by optimal thresholding of second eigenvector

1-graph Laplacian:

The nonlinear graph 1-Laplacian Δ_1 induces the functional $F_1(f)$,

$$F_1(f) := \frac{\langle f, \Delta_1 f\rangle}{\|f\|_1} = \frac{\frac{1}{2}\sum_{i,j=1}^n w_{ij}|f_i - f_j|}{\|f\|_1}.$$

Theorem (Hein,Bühler (2010)): Let G be connected, then

$$\min_C \frac{\mathrm{cut}(C,\overline{C})}{\min\{|C|,|\overline{C}|\}} = \min_{\substack{f \text{ nonconstant} \\ \text{median(f)=0}}} F_1(f) = \lambda_2(\Delta_1),$$

where $\lambda_2(\Delta_1)$ is the second smallest eigenvalue of Δ_1. The second eigenvector of Δ_1 is the indicator vector of the optimal partition.

Tight relaxation of the optimal Cheeger cut !

Quality Guarantee

Tight relaxation of Cheeger cut:

Minimization of continuous relaxation is as hard as original Cheeger cut problem $\Longrightarrow$ **non-convex** and **non-smooth**

No guarantee that one obtains optimal solution by NIPM !

Tight relaxation of Cheeger cut:
Minimization of continuous relaxation is as hard as original Cheeger cut problem $\Longrightarrow$ **non-convex** and **non-smooth**

No guarantee that one obtains optimal solution by NIPM !

but

Quality guarantee:

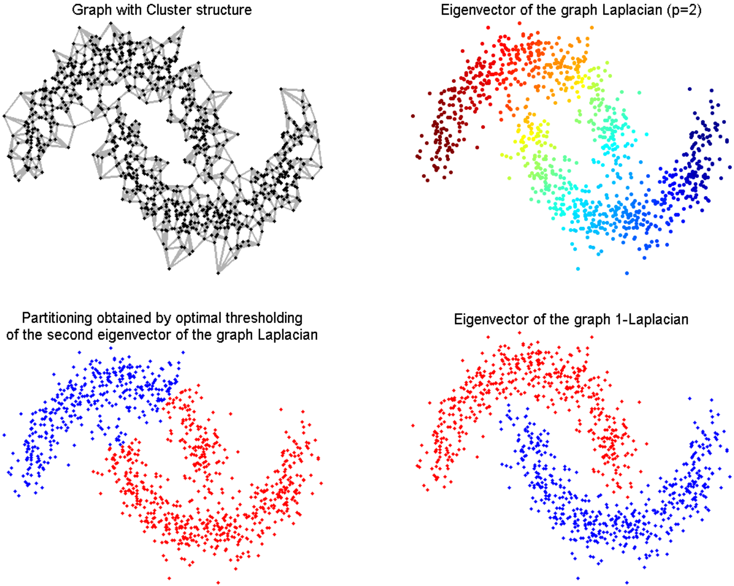
Theorem

Let $(A, \overline{A})$ be a given partition of V. If one uses as initialization of NIPM, $f^0 = \mathbf{1}_A$, then either NIPM terminates after one step or it yields an f^1 which after optimal thresholding gives a partition $(B, \overline{B})$ which satisfies

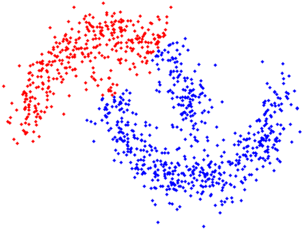
$$\frac{\mathrm{cut}(B, \overline{B})}{\min\{|B|, |\overline{B}|\}} < \frac{\mathrm{cut}(A, \overline{A})}{\min\{|A|, |\overline{A}|\}}.$$

Next Goal: Global approximation guarantees.

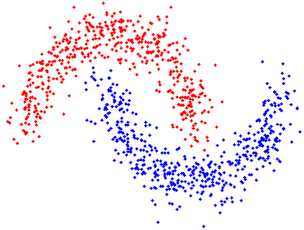
Cheeger Cut: 1-Laplacian (NLEP) vs. 2-Laplacian (LEP)



Result of LEP Relaxation



Result of NLEP Relaxation

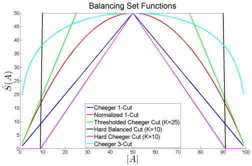


	Linear	Nonlinear
Ratio	$\frac{\sum_{i,j=1}^{n} w_{ij}(x_i - x_j)^2}{\|x\|_2^2}$	$\frac{\sum_{i,j=1}^{n} w_{ij}\lvert x_i - x_j\rvert}{\|x\|_1}$
Approximation Guarantee	**loose**	**tight !** Hein, Bühler(2010)
Convergence	globally optimal	locally optimal
Scalability	✓	✓
Quality	+	+++

1-Spectral Clustering beats state of the art methods on graph partitioning benchmark

Balanced graph cut problem:

$$\min_{A \subset V} \frac{\mathrm{cut}(A, \overline{A})}{\hat{S}(A)}.$$



Balancing set function $\hat{S}$:

Name	$\hat{S}(A)$
Cheeger cut	$\min\{\lvert A\rvert, \lvert\overline{A}\rvert\}$
Ratio cut	$\lvert A\rvert\lvert\overline{A}\rvert$
Hard balanced cut	$\begin{cases} 1, & \text{if } \min\{\lvert A\rvert, \lvert\overline{A}\rvert\} \geq K \\ 0, & \text{else.} \end{cases}$

Modeling of different bias towards balanced partitions via choice of $\hat{S}$.

Do there exist tight relaxations for all balancing set functions ?

Definition

Let $f \in \mathbb{R}^V$ be ordered in increasing order $f_1 \leq f_2 \leq \ldots \leq f_n$ and define $C_i = \{j \in V \,|\, f_j > f_i\}$.Then $S : \mathbb{R}^V \to \mathbb{R}$ given by

$$S(f) \;=\; \sum_{i=1}^{n} f_i \Big(\hat{S}(C_{i-1}) - \hat{S}(C_i) \Big) \;=\; \sum_{i=1}^{n-1} \hat{S}(C_i)(f_{i+1} - f_i) + f_1 \hat{S}(V)$$

is the **Lovasz extension** of $\hat{S} : 2^V \to \mathbb{R}$. One has $S(\mathbf{1}_A) = \hat{S}(A)$, $\forall A \subset V$.

Definition

A set function $\hat{F} : 2^V \to \mathbb{R}$ is **submodular** if for all $A, B \subset V$,

$$\hat{F}(A \cup B) + \hat{F}(A \cap B) \;\leq\; \hat{F}(A) + \hat{F}(B).$$

Proposition

Every set function can be written as difference of two submodular functions.

Theorem (Hein, Setzer (2011))
It holds

$$\min_{f\in\mathbb{R}^V} \frac{\frac{1}{2}\sum_{i,j=1}^n w_{ij}|f_i - f_j|}{S(f)} = \min_{A\subset V} \frac{\mathrm{cut}(A,\overline{A})}{\hat{S}(A)},$$

if either one of the following two conditions holds

1. S is positively one-homogeneous, even, convex and $S(f+\alpha\mathbf{1}) = S(f)$ for all $f\in\mathbb{R}^V, \alpha\in\mathbb{R}$ and $\hat{S}$ is defined as $\hat{S}(A) = S(\mathbf{1}_A)$, $\forall\, A\subset V$.
2. S is the Lovasz extension of the non-negative, symmetric set function $\hat{S}$ with $\hat{S}(\emptyset) = 0$.

Let $f\in\mathbb{R}^V$ and $C_t := \{i\in V \,|\, f_i > t\}$, then it holds in both cases,

$$\min_{t\in\mathbb{R}} \frac{\mathrm{cut}(C_t,\overline{C_t})}{\hat{S}(C_t)} \leq \frac{\frac{1}{2}\sum_{i,j=1}^n w_{ij}|f_i - f_j|}{S(f)}.$$

Minimization of a non-negative ratio of 1-homogeneous d.c. functions

$$\min_{f\in\mathbb{R}^n} \frac{R_1(f) - R_2(f)}{S_1(f) - S_2(f)}.$$

- Note that for a 1-homogeneous convex function

$$S(f) \geq \langle u, f\rangle, \forall f \in \mathbb{R}^n,\ g \in \mathbb{R}^n,\ u \in \partial S(g)$$

- Minimize at each step the convex-concave ratio

$$\frac{R_1(f) - \langle r_2, f\rangle}{\langle f, s_1\rangle - S_2(f)}, \qquad \text{where } r_2 \in \partial R_2(f^k),\ s_1 \in \partial S_1(f^k)$$

  via Dinkelbach's method. This yields the convex opt. problem:

$$\min_{f\in D} R_1(f) - \langle r_2, f\rangle + \lambda^k \left(S_2(f) - \langle s_1, f\rangle\right)$$

- Monotonic descent and convergence to critical point as for NIPM

Latest result:

$$\min_{C \subset V} \quad \frac{\hat{R}(C)}{\hat{S}(C)}$$
$$\text{subj. to}: \quad \widehat{M}_i(C) \geq k_i, \quad i = 1, \ldots, K$$

has a **tight relaxation** into a **nonlinear eigenproblem** if

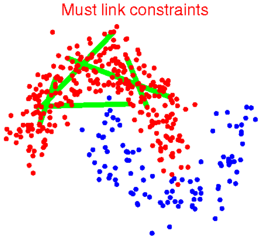
- $\hat{R}, \hat{S}$ are non-negative set functions
- $\hat{R}(\emptyset) = \hat{S}(\emptyset) = 0$
- The constraint functions $\hat{M}_i$ underlie no restrictions

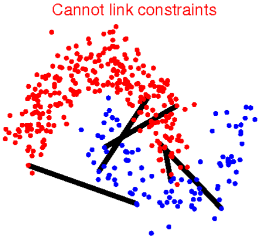
Integration of prior knowledge in clustering/community detection problems via constraints !

Constrained Normalized Cut

Clustering with prior knowledge (Rangapuram, Hein (2012))

- must-link and cannot-link constraints





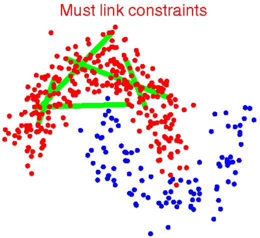
- a partition is called consistent if all constraints are satisfied

Constrained ratio cut problem:

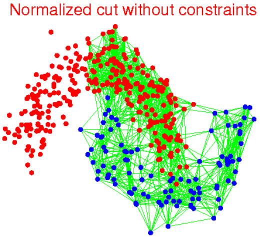
$$\min_{(C,\overline{C})\text{ consistent}} \frac{\mathrm{cut}(C,\overline{C})}{\mathrm{vol}(C)\,\mathrm{vol}(\overline{C})}$$

- previous methods can not guarantee that constraints are satisfied

Must-link and cannot-link constraints

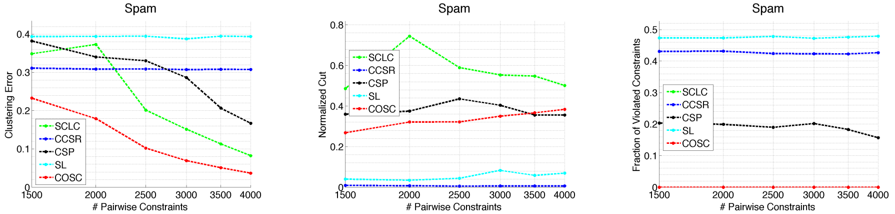


Result of unconstrained 1-Spectral Clustering (left) and constrained normalized cut (right)

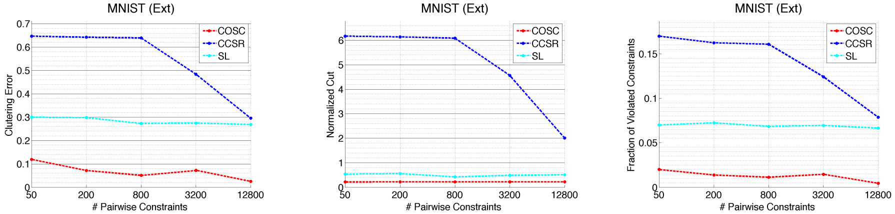


Our NLEP formulation: COSC

Binary-partitioning problem (Spam dataset $|V| = 4207$):



Multi-partitioning problem (extended MNIST dataset, $|V| = 630000$):



Conclusion and Outlook

Benefits of Nonlinear Eigenproblems

- better integration of modeling goals using additional degrees of freedom
- generalized inverse power method makes computation of nonlinear eigenvectors feasible
- Tight relaxation of combinatorial problems as nonlinear eigenproblems

Open Problems in Nonlinear Eigenproblems:

- What is a suitable min-max principle for nonlinear eigenvectors ?
- Computation of higher-order eigenvectors
- Theory of modeling properties of eigenvectors via choice of R and S
- Approximation guarantees for tight relaxations of combinatorial problems
- ...

European Research Council

Established by the European Commission

Supporting top researchers from anywhere in the world

Ph.D. and Postdoc positions

- ERC Starting Grant starting in Autumn

  Nonlinear Eigenproblems for Data Analysis

- Desired background in one or more of the following areas
  1. convex (and non-convex) optimization
  2. machine learning/statistics
  3. functional analysis, variational problems