Sparse Eigen Methods by D.C. Programming

Bharath K. Sriperumbudur David A. Torres
Gert R. G. Lanckriet

University of California, San Diego

International Conference on Machine Learning, 2007.

Generalized Eigenvalue Problem

- Eigenvalue problems are popular in machine learning and statistics.
  - Classification
    - Fisher discriminant analysis (FDA)
  - Dimensionality reduction
    - Principal component analysis (PCA)
    - Canonical correlation analysis (CCA)
- The variational formulation for the generalized eigenvalue problem is given by

$$\begin{aligned} \lambda_{max}(\mathbf{A}, \mathbf{B}) = \max_{\mathbf{x}} \quad & \mathbf{x}^T \mathbf{A} \mathbf{x} \\ \text{s.t.} \quad & \mathbf{x}^T \mathbf{B} \mathbf{x} = 1, \end{aligned} \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{S}^n$ and $\mathbf{B} \succ 0$.

Classification

- Fisher Discriminant Analysis
  - finds a 1-D subspace onto which the projections of data lead to a maximal separation between classes (binary classification setting).
  - The variational formulation for FDA is given by

$$\begin{aligned} \max_{\mathbf{x}} \quad & \mathbf{x}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{x} \\ \text{s.t.} \quad & \mathbf{x}^T(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\mathbf{x} = 1. \end{aligned} \tag{2}$$

  - $\mathbf{A} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$ is the between-cluster variance and $\mathbf{B} = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$ is the within-cluster variance.

Dimensionality Reduction

- Principal Component Analysis
  - identifies the direction of maximal variance.
  - The variational formulation for PCA is given by

$$\begin{aligned} \max_{\mathbf{x}} \quad & \mathbf{x}^T \boldsymbol{\Sigma} \mathbf{x} \\ \text{s.t.} \quad & \mathbf{x}^T \mathbf{x} = 1. \end{aligned} \tag{3}$$

  - $\mathbf{A} = \boldsymbol{\Sigma}$ is the covariance matrix and $\mathbf{B} = \mathbf{I}$ is the identity matrix.

Dimensionality Reduction

- Canonical Correlation Analysis
  - two-view PCA between spaces $\mathscr{X}$ and $\mathscr{Y}$.
  - The variational formulation for CCA is given by

$$\begin{aligned} \max_{\mathbf{w}_x, \mathbf{w}_y} \quad & \mathbf{w}_x^T \mathbf{S}_{xy} \mathbf{w}_y \\ \text{s.t.} \quad & \mathbf{w}_x^T \mathbf{S}_{xx} \mathbf{w}_x = 1 \,, \mathbf{w}_y^T \mathbf{S}_{yy} \mathbf{w}_y = 1. \end{aligned} \tag{4}$$

  - $\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{S}_{xy} \\ \mathbf{S}_{yx} & \mathbf{0} \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} \mathbf{S}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{yy} \end{pmatrix}$ and $\mathbf{x} = \begin{pmatrix} \mathbf{w}_x \\ \mathbf{w}_y \end{pmatrix}$ where $\mathbf{S} = \begin{pmatrix} \mathbf{S}_{xx} & \mathbf{S}_{xy} \\ \mathbf{S}_{yx} & \mathbf{S}_{yy} \end{pmatrix}$ is the covariance matrix between samples from $\mathscr{X}$ and $\mathscr{Y}$.

Why Sparsity?

- Usually, the solutions of FDA, PCA and CCA are not sparse.
- This often makes it difficult to interpret the results.
- PCA/CCA: For better interpretability, few relevant features are required that explain as much variance as possible.
  - Applications: bio-informatics, finance etc.
- FDA: feature selection aids generalization performance by promoting sparse solutions.
- Sparse representation $\Rightarrow$ better interpretation, better generalization and reduced computational costs.

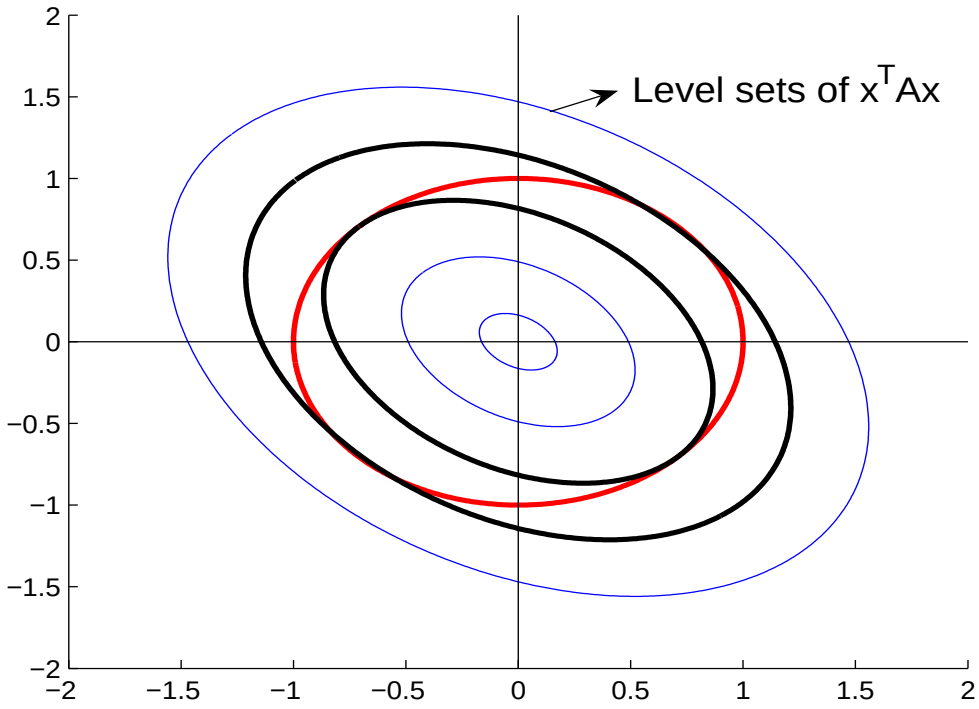
Sparse Generalized Eigenvalue Problem

- The variational formulation for the sparse generalized eigenvalue problem is given by

$$\begin{aligned} \max_{\mathbf{x}} \quad & \mathbf{x}^T \mathbf{A} \mathbf{x} \\ \text{s.t.} \quad & \mathbf{x}^T \mathbf{B} \mathbf{x} = 1 \\ & ||\mathbf{x}||_0 \leq k, \end{aligned} \tag{5}$$

  where $1 \leq k \leq n$ and $||\mathbf{x}||_0$ is the cardinality of $\mathbf{x}$.
- Eq. (5) is non-convex, NP-hard and therefore intractable.
- Usually, the ℓ_1 approximation is used for the cardinality constraint.
- The problem is still computationally hard.

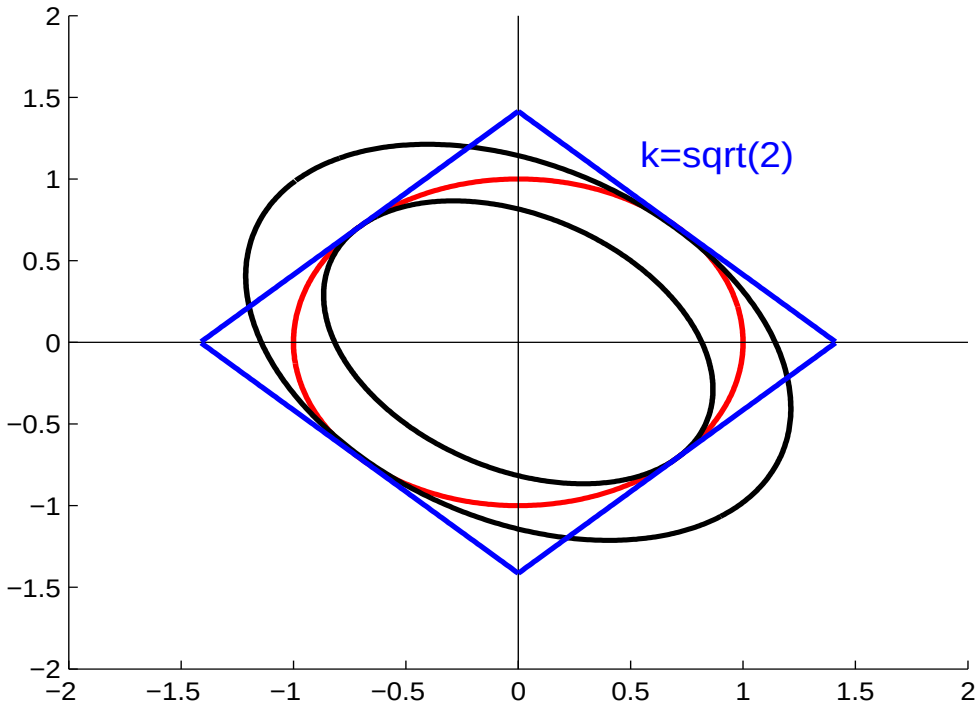
Sparse Generalized Eigenvalue Problem



Geometry of the generalized eigenvalue problem

KKT condition: $\mathbf{Ax} = \lambda \mathbf{Bx}$. The generalized eigenvalue problem is very special.

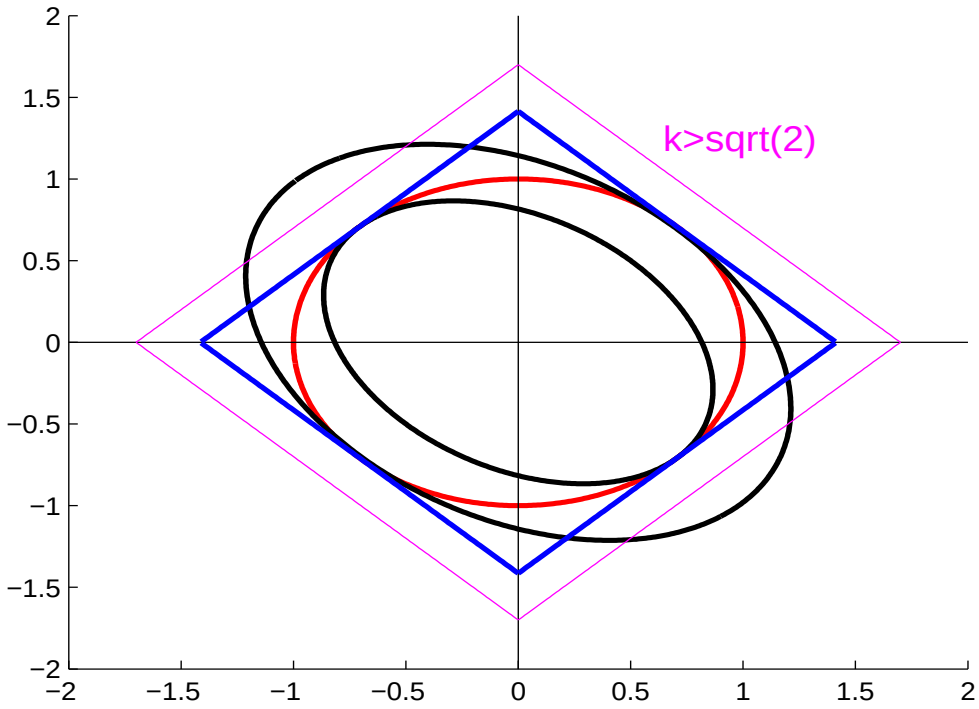
Sparse Generalized Eigenvalue Problem



Geometry of the sparse generalized eigenvalue problem

$$||\mathbf{x}||_1 \leq k$$

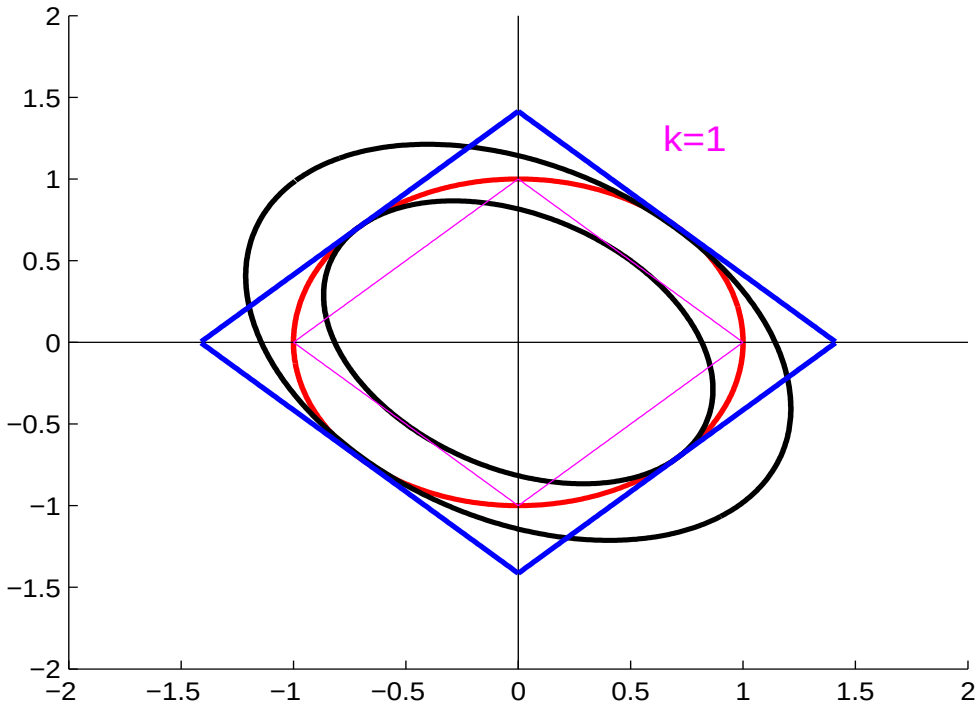
Sparse Generalized Eigenvalue Problem



Geometry of the sparse generalized eigenvalue problem

$||\mathbf{x}||_1 \leq k$; uninteresting

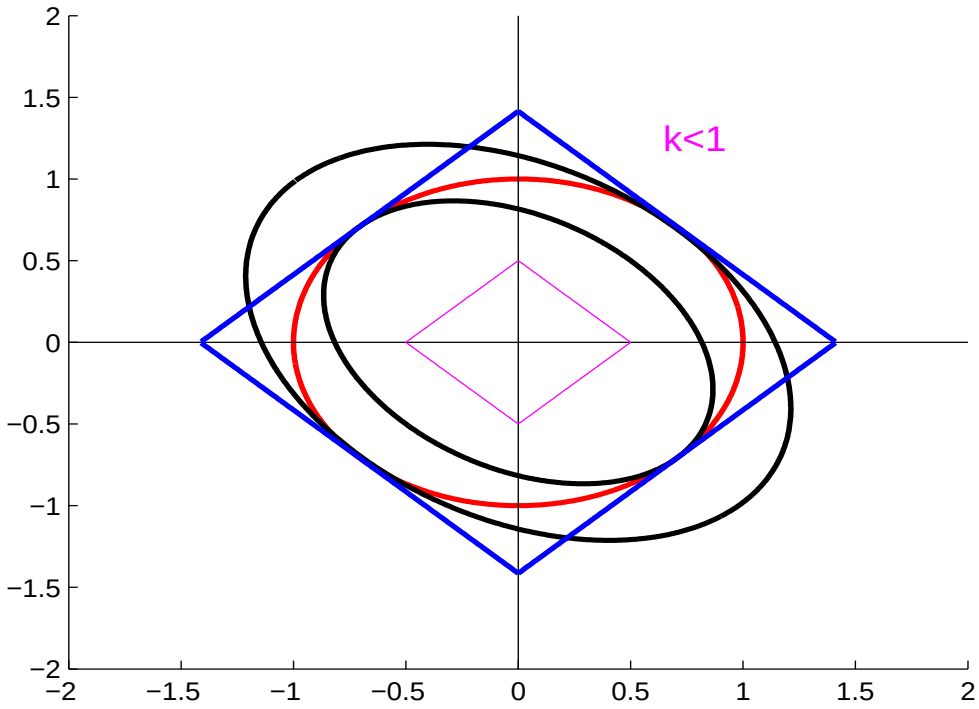
Sparse Generalized Eigenvalue Problem



Geometry of the sparse generalized eigenvalue problem

$||\mathbf{x}||_1 \leq k$; uninteresting

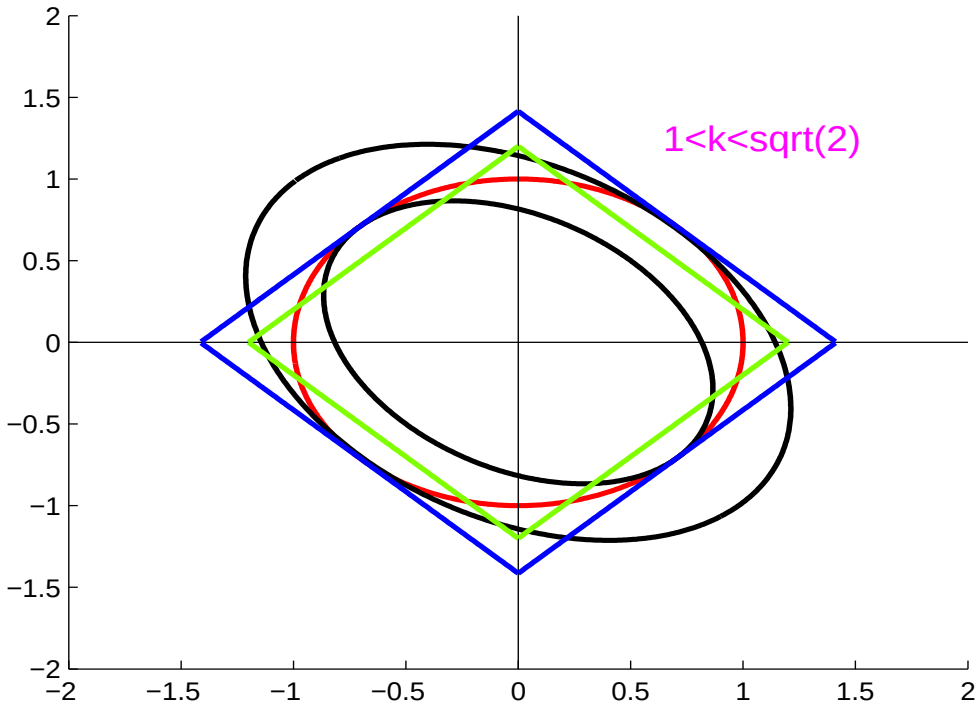
Sparse Generalized Eigenvalue Problem



Geometry of the sparse generalized eigenvalue problem

$||\mathbf{x}||_1 \leq k$; infeasible

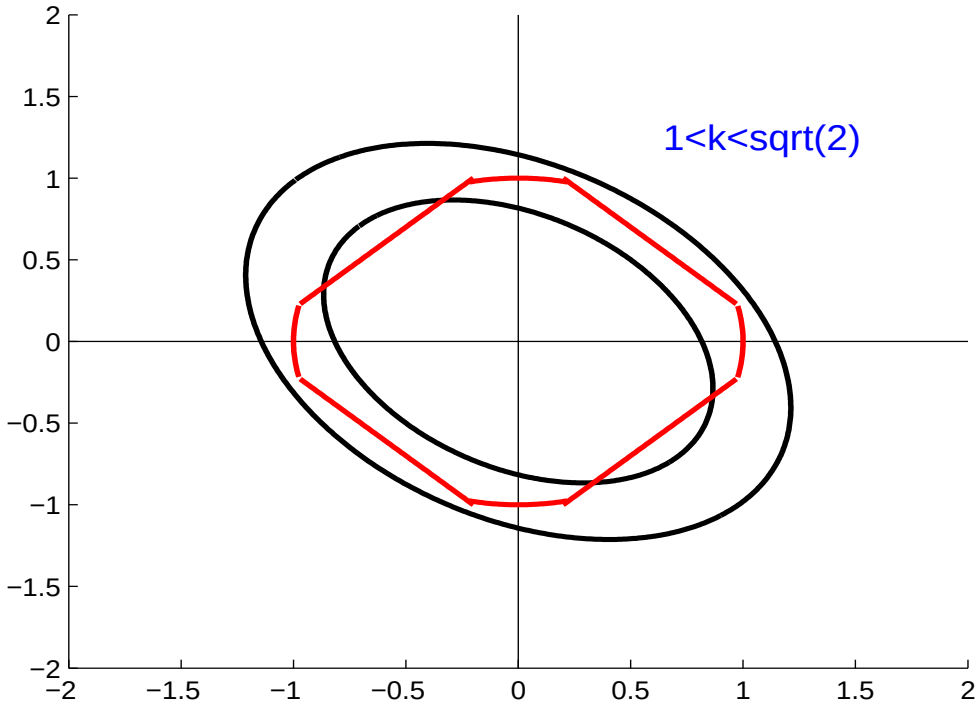
Sparse Generalized Eigenvalue Problem



Geometry of the sparse generalized eigenvalue problem

$||\mathbf{x}||_1 \leq k$; interesting

Sparse Generalized Eigenvalue Problem



Geometry of the sparse generalized eigenvalue problem

- For $\mathbf{x} \in \mathbb{R}^n$
  - $k < 1$, infeasible
  - $k = 1$, uninteresting
  - $1 < k < \sqrt{n}$, interesting but hard
  - $k \geq \sqrt{n}$, uninteresting (generalized eigenvalue problem).

Prior Work

- SCoTLASS [Jolliffe et al., 2003]: ℓ_1 approximation; locally convergent algorithm for $\mathbf{B} = \mathbf{I}$.
- DSPCA [d'Aspremont et al., 2005]: ℓ_1 approximation, followed by lifting and rank constraint relaxation; a convex semidefinite program (SDP) for $\mathbf{B} = \mathbf{I}$.
- The variational formulation for DSPCA is given by

$$\begin{aligned} \max_{\mathbf{X} \succeq 0} \quad & \mathrm{tr}(\mathbf{A}\mathbf{X}) \\ \text{s.t.} \quad & \mathrm{tr}(\mathbf{X}) = 1, \, \mathbf{1}^T |\mathbf{X}| \mathbf{1} \leq k, \end{aligned} \tag{6}$$

- SPCA [Zou et al., 2004]: ℓ_1-penalized regression algorithm for PCA using elastic-net; solved very efficiently using least angle regression.

Prior Work

- DSPCA is computationally very expensive.
  - Interior point methods: $O(n^6 \log(1/\epsilon))$.
  - First-order methods: $O(n^4 \sqrt{\log(n)}/\epsilon)$, where ϵ is the required accuracy on the optimal value.
- Convexity vs. Scalability
  - SDP relaxation is the only possible convex approach.
  - prohibitively expensive for large n.
  - though convex, it is only an approximation to the true solution.
- Options: locally convergent algorithms / expensive mixed-integer programs.
- Currently, SPCA is the only viable option for handling very high-dimensional datasets (on the order of 10,000).

Approximation to $||\mathbf{x}||_0$

- Two observations
  - The ℓ_1-norm relaxation does not simplify Eq. (5) $\Rightarrow$ a better approximation to cardinality would improve sparsity.
  - The convex SDP approximation to Eq. (5) scales terribly in size $\Rightarrow$ use a locally convergent algorithm with better scalability.
- Eq. (5) can be written as

$$\begin{aligned} \max_{\mathbf{x}} \quad & \mathbf{x}^T\mathbf{A}\mathbf{x} - \rho\,||\mathbf{x}||_0 \\ \text{s.t.} \quad & \mathbf{x}^T\mathbf{B}\mathbf{x} \leq 1, \end{aligned} \tag{7}$$

  where $\rho \geq 0$.
- Approximate $||\mathbf{x}||_0$ by $\sum_{i=1}^{n} \log(\varepsilon + |x_i|)$, where $0 \leq \varepsilon \ll 1$ avoids problems when one of the x_i is zero.

Approximation to $||\mathbf{x}||_0$

- The approximation can be interpreted as defining a limiting Student's t-distribution prior over $\mathbf{x}$ (leading to an improper prior) given by

$$p(\mathbf{x}) \propto \prod_{i=1}^{n} \frac{1}{|x_i|}$$

  and computing its negative log-likelihood.

- Eq. (7) reduces to

$$\begin{aligned} \max_{\mathbf{x}} \quad & \mathbf{x}^T \mathbf{A}\mathbf{x} - \rho \sum_{i=1}^{n} \log(\varepsilon + |x_i|) \\ \text{s.t.} \quad & \mathbf{x}^T \mathbf{B}\mathbf{x} \leq 1. \end{aligned} \tag{8}$$

Approximation to $||\mathbf{x}||_0$

Proposition

Let $\hat{\mathbf{x}}$ and $\check{\mathbf{x}}$ be the maximizers of Eq. (7) and Eq. (8) respectively. If $\check{\mathbf{x}}$ is independent of the choice of ε and $\delta \leq |\check{x}_i| < \infty$ for some fixed $\delta > 0$, where $i = \{j : |\check{x}_j| \neq 0, 1 \leq j \leq n\}$, then

$$||\check{\mathbf{x}}||_0 \leq ||\hat{\mathbf{x}}||_0 + \frac{c_\delta}{\log \varepsilon},$$

where c_δ is a constant dependent on δ.

- The global maximizers of Eq. (7) and Eq. (8) have almost the same cardinality.

Approximation to $||\mathbf{x}||_0$

- With $\varepsilon = 0$, Eq. (8) can be written as

$$\begin{aligned} \max_{\mathbf{x},\mathbf{y}} \quad & \mathbf{x}^T\mathbf{A}\mathbf{x} - \rho \sum_{i=1}^{n} \log y_i \\ \text{s.t.} \quad & (\mathbf{x},\mathbf{y}) \in \mathscr{F} = \{(\mathbf{x},\mathbf{y}) : \mathbf{x}^T\mathbf{B}\mathbf{x} \leq 1\,, -\mathbf{y} \preceq \mathbf{x} \preceq \mathbf{y}\}. \end{aligned} \tag{9}$$

- Rewriting Eq. (9), we have

$$\min_{\mathbf{x},\mathbf{y}} \mathcal{I}_{\mathscr{F}}(\mathbf{x},\mathbf{y}) - \left(\mathbf{x}^T\mathbf{A}\mathbf{x} - \rho \sum_{i=1}^{n} \log y_i \right), \tag{10}$$

where the convex function $\mathcal{I}_{\mathscr{F}} = \begin{cases} 0 & (\mathbf{x},\mathbf{y}) \in \mathscr{F} \\ \infty & (\mathbf{x},\mathbf{y}) \notin \mathscr{F} \end{cases}$ is the *indicator function* of the convex set, $\mathscr{F}$.

D.C. Programming

Definition

Let $\mathscr{C}$ be a convex subset of $\mathbb{R}^n$. A real-valued function $f : \mathscr{C} \to \mathbb{R}$ is called difference of convex functions (d.c.) on $\mathscr{C}$, if there exist two convex functions $g, h : \mathscr{C} \to \mathbb{R}$ such that f can be expressed in the form

$$f(x) = g(x) - h(x). \tag{11}$$

- The variational form of d.c. programming problems is given by

$$\begin{aligned} \min_{x \in \mathscr{X}} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0,\ i = 1, \ldots, m, \end{aligned} \tag{12}$$

where $f_i = g_i - h_i$, $i = 0, \ldots, m$, are d.c. functions and $\mathscr{X}$ is a closed convex subset of $\mathbb{R}^n$.

D.C. Programming

- Global optimization approaches such as branch and bound, cutting planes, etc. are not scalable to large-scale d.c. problems.
- A robust and efficient d.c. minimization algorithm (DCA) was proposed by [Tao and An, 1998].
- DCA is a primal-dual subdifferential method for solving large-scale d.c. programs.
- Based on the d.c. duality and the local optimality, DCA solves a sequence of convex programs.
- DCA can be understood as the convex-concave procedure (CCCP) [Yuille and Rangarajan, 2003].

Sparse Generalized Eigenvalue Algorithm

Require: $\mathbf{A} \succeq 0$, $\mathbf{B} \succ 0$ and $\rho \geq 0$

1: Choose $\mathbf{x}_0 \in \{\mathbf{x} : \mathbf{x}^T\mathbf{B}\mathbf{x} \leq 1\}$ arbitrarily
2: **repeat**
3:

$$\begin{aligned} \bar{\mathbf{x}}^* = \arg\max_{\bar{\mathbf{x}}} \quad & \mathbf{x}_l^T\mathbf{A}\mathbf{D}(\mathbf{x}_l)\bar{\mathbf{x}} - \tfrac{\rho}{2}||\bar{\mathbf{x}}||_1 \\ \text{s.t.} \quad & \bar{\mathbf{x}}^T\mathbf{D}(\mathbf{x}_l)\mathbf{B}\mathbf{D}(\mathbf{x}_l)\bar{\mathbf{x}} \leq 1 \end{aligned} \tag{13}$$

4: $\mathbf{x}_{l+1} = \mathbf{D}(\mathbf{x}_l)\bar{\mathbf{x}}^*$
5: **until** $\mathbf{x}_{l+1} = \mathbf{x}_l$
6: **return** $\mathbf{x}_l$, $\bar{\mathbf{x}}^*$

where $\mathbf{D}(\mathbf{x}) = \text{diag}(\mathbf{x})$.

- solves a sequence of convex quadratically constrained quadratic programs (QCQPs).

Sparse Generalized Eigenvalue Algorithm

Proposition

Let $\rho = 0$, $\mathbf{x}_l$ *be the output of sparse generalized eigenvalue algorithm and*

$$\begin{aligned} \breve{\mathbf{x}} = \arg\max_{\mathbf{x}} \quad & \mathbf{x}^T \mathbf{A}\mathbf{x} \\ s.t. \quad & \mathbf{x}^T \mathbf{B}\mathbf{x} = 1. \end{aligned}$$

Then

$$\mathbf{x}_l^T \mathbf{A}\mathbf{x}_l = \lambda_{max}(\mathbf{B}^{-1}\mathbf{A}) \text{ and } \mathbf{x}_l = \breve{\mathbf{x}},$$

where $\lambda_{max}(\mathbf{B}^{-1}\mathbf{A})$ *is the maximum eigenvalue of* $\mathbf{B}^{-1}\mathbf{A}$.

- Local and global solutions are the same for $\rho = 0$.

Sparse Generalized Eigenvalue Algorithm

Proposition

Let $\mathbf{B} = \mathbf{I}$, $\mathbf{A} \succeq 0$ *and* $\rho = 0$. *Then sparse generalized eigenvalue algorithm is the power method for eigenvalue computation.*

- We call the sparse generalized eigenvalue program for $\mathbf{B} = \mathbf{I}$ as DC-PCA.
- Comparison to SCoTLASS
  - Using DCA, the variational formulation for SCoTLASS is given by

$$\begin{aligned} \bar{\mathbf{x}}^* = \arg\max_{\bar{\mathbf{x}}} \quad & \mathbf{x}_l^T \mathbf{A} \bar{\mathbf{x}} - \frac{\rho}{2} ||\bar{\mathbf{x}}||_1 \\ \text{s.t.} \quad & \bar{\mathbf{x}}^T \bar{\mathbf{x}} \leq 1, \end{aligned} \tag{14}$$

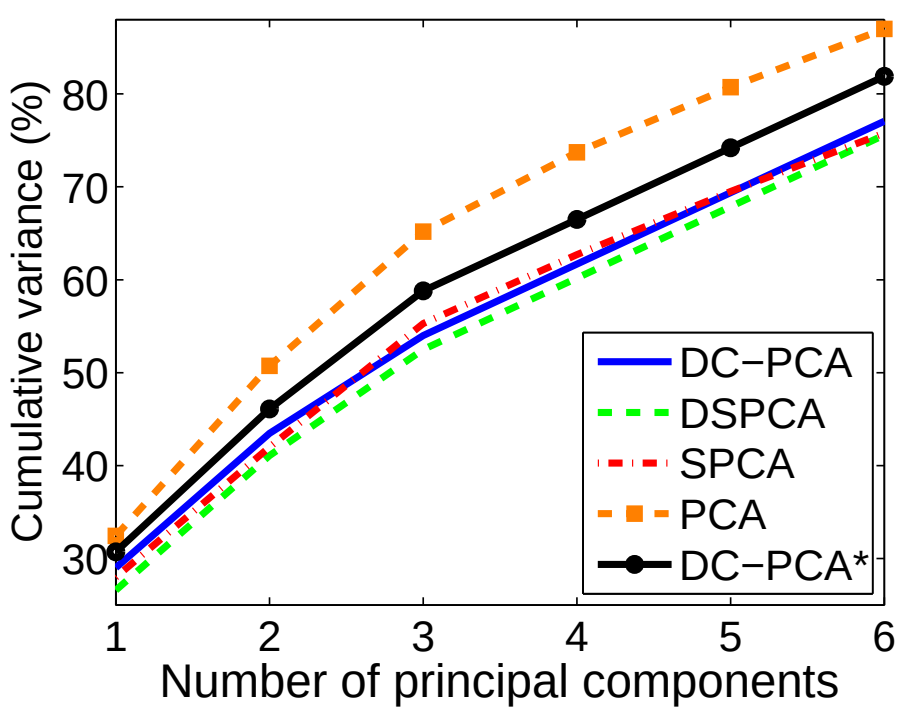
    with $\mathbf{x}_{l+1} = \bar{\mathbf{x}}^*$.
  - Eq. (14) differs from DC-PCA in the multiplicative update. Therefore, DC-PCA yields atleast as much sparsity as SCoTLASS.

Experiments & Results

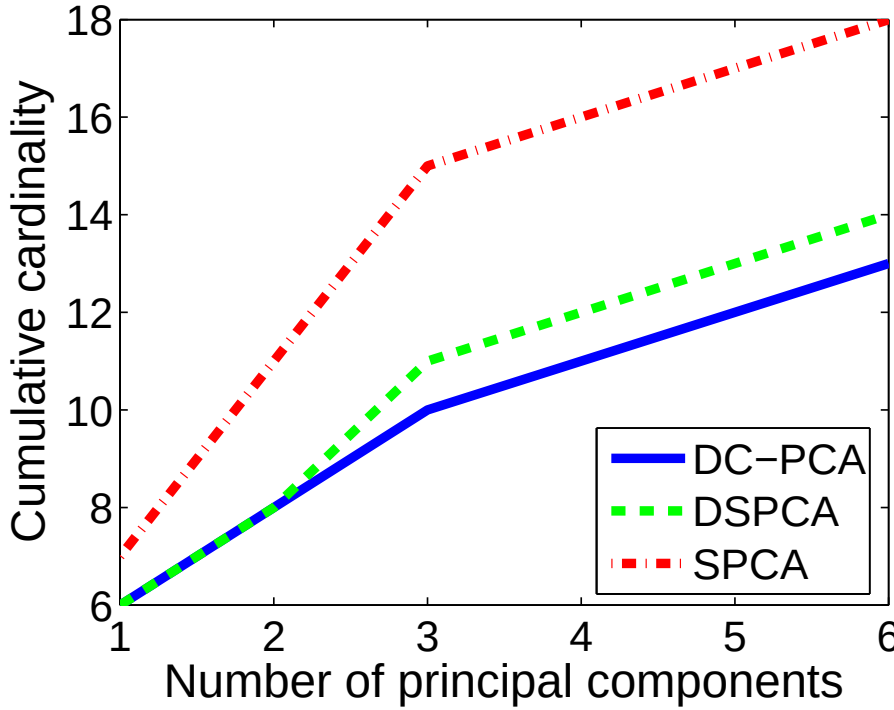
- Pit props data [Jeffers, 1967]
  - A benchmark data to test sparse PCA algorithms.
  - 180 observations and 13 measured variables.
  - 6 principal directions are considered as they capture 87% of the total variance.

Algorithm	Sparsity pattern	Cumulative cardinality	Cumulative variance
SPCA	(7,4,4,1,1,1)	18	75.8%
DSPCA	(6,2,3,1,1,1)	14	75.5%
DC-PCA	(6,2,2,1,1,1)	13	77.1%
DC-PCA	(7,4,4,1,1,1)	18	81.8%

Pit Props



(a)

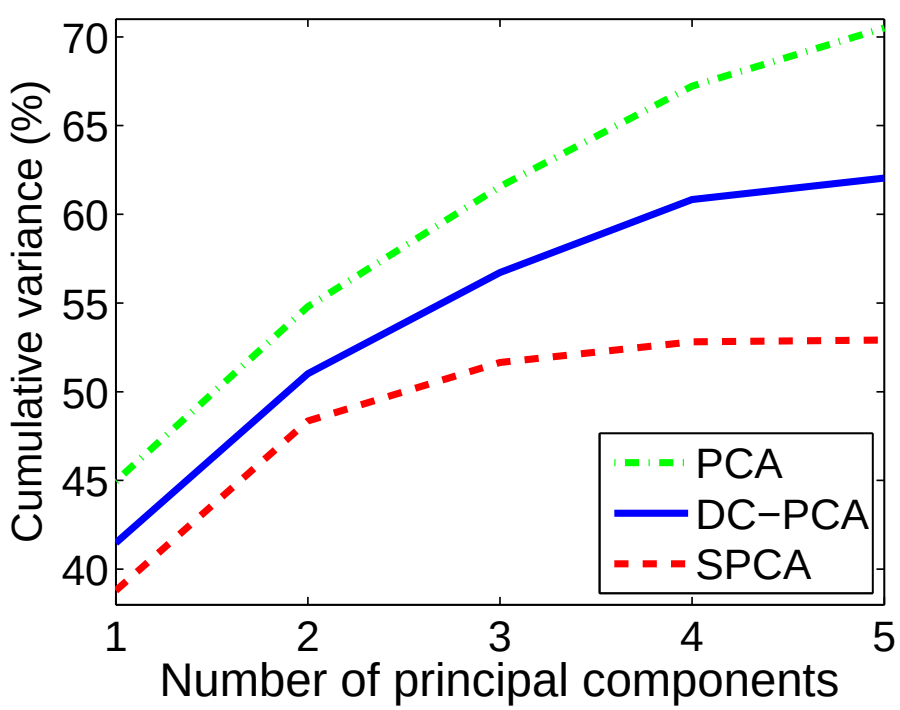


(b)

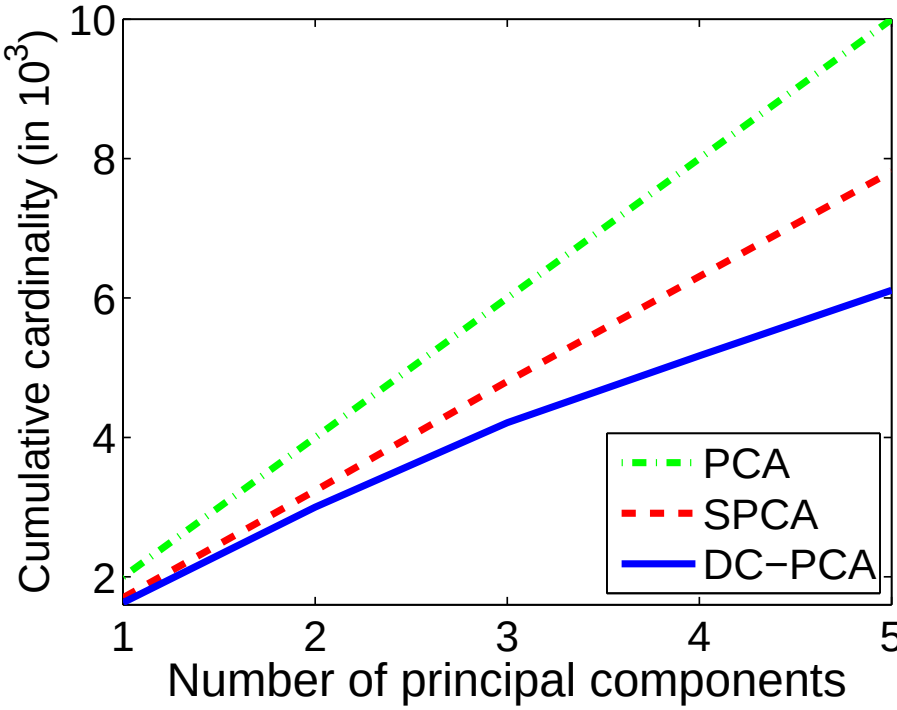
Figure: (a) cumulative variance (b) cumulative cardinality for first 6 sparse principal components (PCs). DC-PCA* in (a) represents DC-PCA evaluated at SPCA's sparsity pattern of (7,4,4,1,1,1).

Colon Cancer [Alon et al., 1999]

- 62 tissue samples (22 normal and 40 cancerous) with gene expression profiles of $n = 2000$ genes extracted from DNA micro-array data.



(a)

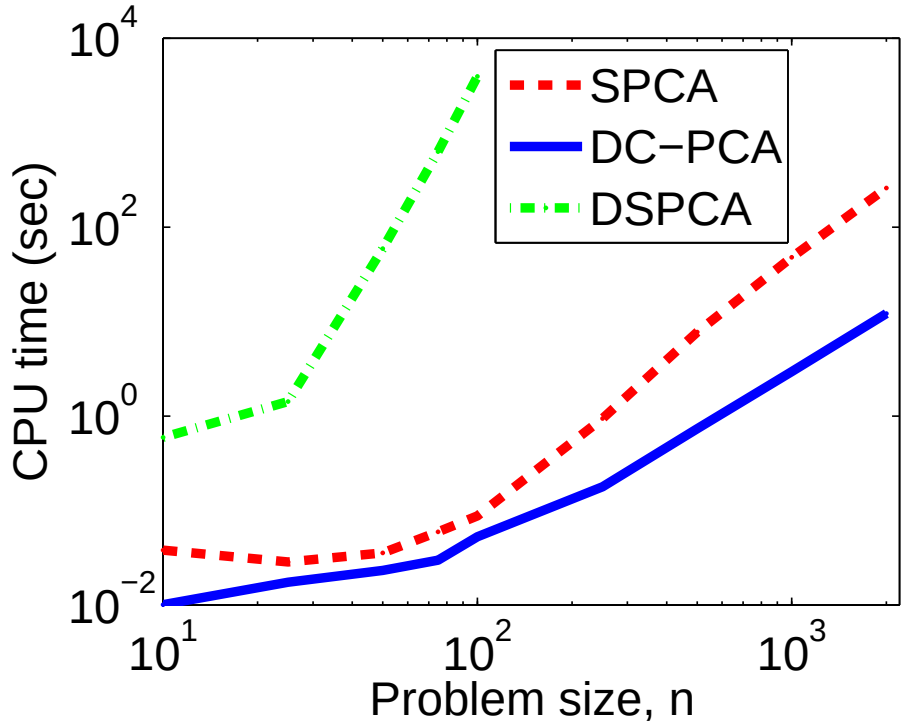


(b)

Figure: (a) cumulative variance (b) cumulative cardinality for first 5 sparse principal components (PC).

Scalability

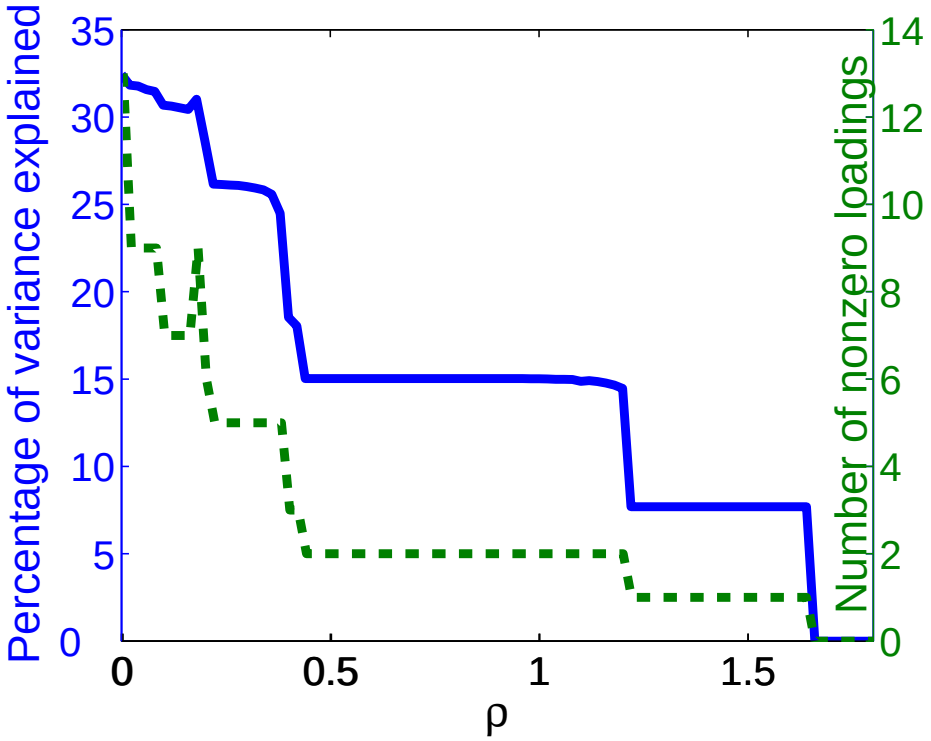
- Randomly chosen problems of size n ranging from 10 to 2000.
- Linux 3 GHz, 4 GB RAM workstation.



The empirical complexity is $O(n^p)$ where $p = 1.46$ for DC-PCA, $p = 1.91$ for SPCA and $p = 3.92$ for DSPCA.

Controlling sparsity with ρ

- DC-PCA & SPCA
  - setting ρ is not straight forward to achieve the required sparsity, k.
- DSPCA
  - k can be directly used in the semidefinite program.



Conclusion

- A sparse generalized eigenvalue algorithm using d.c. programming is proposed.
- The sparse PCA algorithm (DC-PCA) is derived as a special case.
- DC-PCA has better sparsity and scalability than SPCA and DSPCA.
- Sparsity parameter
  - DC-PCA & SPCA: difficult to set ρ.
  - DSPCA: k is explicitly mentioned.
- Quality of approximation
  - Difficult to characterize (theoretically) the quality of DC-PCA solution.
  - The quality of convex relaxation can be studied for DSPCA [El Ghaoui, 2006].

Future Work

- Extend the sparse generalized eigenvalue algorithm to any $\mathbf{A} \in \mathbb{S}^n$.
- Explore path following techniques to efficiently set the sparsity parameter, ρ.
- Investigate the sparsity paradigm for CCA which has interesting applications in dictionary translation, music annotation, etc.

References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999).
Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon cancer tissues.
Cell Biology, 96:6745–6750.

d'Aspremont, A., El Ghaoui, L., Jordan, M. I., and Lanckriet, G. R. G. (2005).
A direct formulation for sparse PCA using semidefinite programming.
In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 41–48, Cambridge, MA. MIT Press.

El Ghaoui, L. (2006).
On the quality of a semidefinite programming bound for sparse principal component analysis.
arXive.org.

Jeffers, J. (1967).
Two case studies in the application of principal components.
Applied Statistics, 16:225–236.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003).
A modified principal component technique based on the LASSO.
Journal of Computational and Graphical Statistics, 12:531–547.

Tao, P. D. and An, L. T. H. (1998).
D.c. optimization algorithms for solving the trust region subproblem.
SIAM J. Optim., pages 476–505.

Yuille, A. L. and Rangarajan, A. (2003).
The concave-convex procedure.
Neural Computation, pages 915–936.

Zou, H., Hastie, T., and Tibshirani, R. (2004).
Sparse principal component analysis.
Technical report, Statistics Department, Stanford University.

Questions

Thank You