A D.C. Programming Approach to the Sparse Generalized Eigenvalue Problem

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

University of California, San Diego

OPT 2009

Generalized Eigenvalue Problem

Given a matrix pair, $(\mathbf{A}, \mathbf{B})$, find a pair $(\lambda, \mathbf{x})$ such that

$$\mathbf{Ax} = \lambda \mathbf{Bx},$$

where $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{n \times n}$, $\mathbb{C}^n \ni \mathbf{x} \neq \mathbf{0}$ and $\lambda \in \mathbb{C}$.

Variational formulation:

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

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

- Popular in multivariate statistics and machine learning.
  - *Classification* : Fisher discriminant analysis
  - *Dimensionality reduction* : Principal component analysis, Canonical correlation analysis
  - *Clustering* : Spectral clustering

Applications

- *Fisher Discriminant Analysis (FDA)*
  - $\mathbf{A} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$ is the *between-cluster variance*.
  - $\mathbf{B} = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$ is the *within-cluster variance*.
- *Principal Component Analysis (PCA)*
  - $\mathbf{A} = \boldsymbol{\Sigma}$ is the *covariance matrix*.
  - $\mathbf{B}$ is the *identity matrix.*
- *Canonical Correlation Analysis (CCA)*
  - $\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{S}_{xy} \\ \mathbf{S}_{yx} & \mathbf{0} \end{pmatrix}$.
  - $\mathbf{B} = \begin{pmatrix} \mathbf{S}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{yy} \end{pmatrix}$, where $S_{..}$ represents the cross-covariance matrix.

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, document translation etc.
- FDA: *feature selection* aids generalization performance by promoting sparse solutions.
- Sparse representation ⇒ *better interpretation, better generalization and reduced computational costs.*

Sparse Generalized Eigenvalue Problem

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

$$
\begin{aligned}
\max_{\mathbf{x}} \quad & \mathbf{x}^T\mathbf{A}\mathbf{x} \\
s.t. \quad & \mathbf{x}^T\mathbf{B}\mathbf{x} = 1 \\
& \|\mathbf{x}\|_0 \leq k,
\end{aligned}
\tag{2}
$$

  where $1 \leq k \leq n$ and $\|\mathbf{x}\|_0 := \sum_{i=1}^{n} \mathbb{1}_{\{|x_i| \neq 0\}}$ is the *cardinality* of $\mathbf{x}$.

- (2) is *non-convex,* NP-hard and therefore intractable.

- Usually, the *ℓ_1-norm approximation* is used for the cardinality constraint, i.e., replace $\|\mathbf{x}\|_0 \leq k$ by $\|\mathbf{x}\|_1 \leq k$.

- The problem is *still computationally hard*.

Sparse Generalized Eigenvalue Problem

- (2) can be written as

$$\begin{aligned} \max_{\mathbf{x}} \quad & \mathbf{x}^T\mathbf{A}\mathbf{x} - \tilde{\rho}\,||\mathbf{x}||_0 \\ \text{s.t.} \quad & \mathbf{x}^T\mathbf{B}\mathbf{x} \leq 1, \end{aligned} \tag{3}$$

  where $\tilde{\rho} \geq 0$.

- Approximate $||\mathbf{x}||_0$ by $\|\mathbf{x}\|_\varepsilon := \sum_{i=1}^n \frac{\log(1+|x_i|\varepsilon^{-1})}{\log(1+\varepsilon^{-1})}$ for *sufficiently small* $\varepsilon > 0$ as

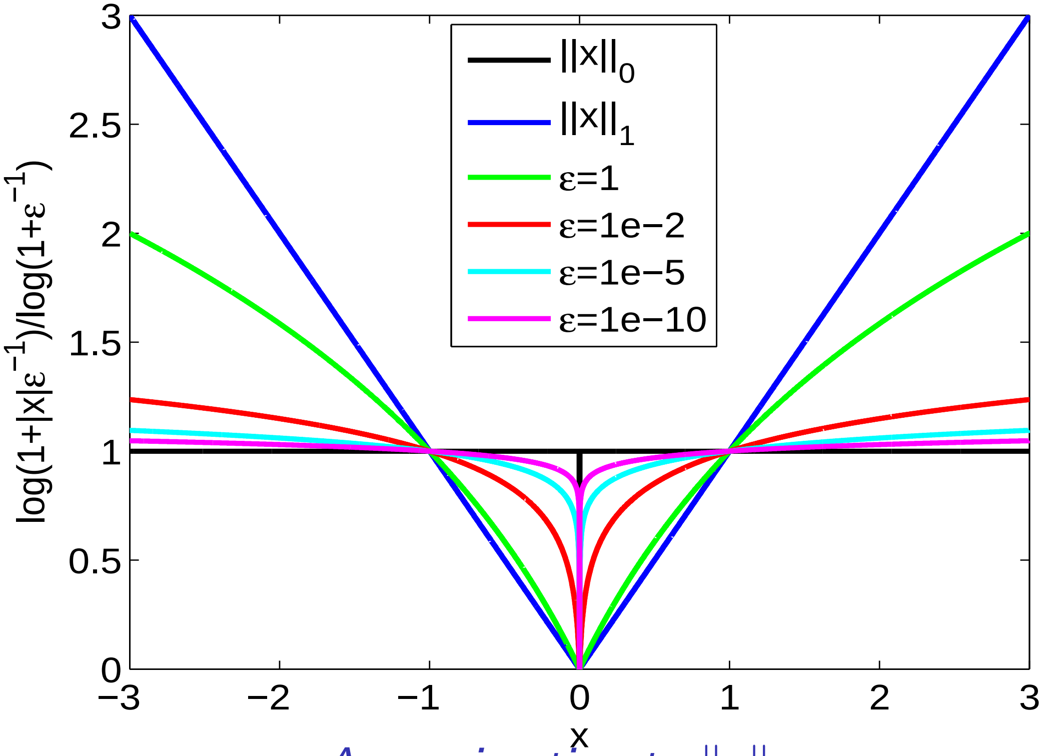
$$\|\mathbf{x}\|_0 = \lim_{\varepsilon \to 0} \sum_{i=1}^n \frac{\log(1+|x_i|\varepsilon^{-1})}{\log(1+\varepsilon^{-1})}. \tag{4}$$

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

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

  and computing its *negative log-likelihood.*

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



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

As $\varepsilon \longrightarrow 0$, $\|\mathbf{x}\|_\varepsilon \longrightarrow \|\mathbf{x}\|_0$ and as $\varepsilon \longrightarrow \infty$, $\|\mathbf{x}\|_\varepsilon \longrightarrow \|\mathbf{x}\|_1$.

Sparse Generalized Eigenvalue Problem

- (3) reduces to the *approximate program,*

$$\begin{aligned} \max_{\mathbf{x}} \quad & \mathbf{x}^T\mathbf{A}\mathbf{x} - \rho_\varepsilon \sum_{i=1}^{n} \log(|x_i| + \varepsilon) \\ \text{s.t.} \quad & \mathbf{x}^T\mathbf{B}\mathbf{x} \leq 1, \end{aligned} \tag{5}$$

  where $\rho_\varepsilon := \frac{\tilde{\rho}}{\log(1+\varepsilon^{-1})}$.

- The task reduces to solving the *approximate program* in (5) with a *small value* of ε.

- (5) can be written as

$$\begin{aligned} \min_{\mathbf{x}} \quad & \tau||\mathbf{x}||^2 - \left(\mathbf{x}^T(\mathbf{A} + \tau\mathbf{I})\mathbf{x} - \rho \sum_{i=1}^{n} \log(|x_i| + \varepsilon)\right) \\ s.t. \quad & \mathbf{x}^T\mathbf{B}\mathbf{x} \leq 1, \end{aligned} \tag{6}$$

  where $\tau \geq \max(0, -\lambda_{min}(\mathbf{A}))$.

- The objective in (6) is a *difference of two convex functions.*

Majorization-Minimization (MM)

- Suppose we want to minimize f over $\Omega \subset \mathbb{R}^n$. Construct a *majorization function*, g over $\Omega \times \Omega$ such that

$$f(x) \leq g(x, y), \forall x, y \in \Omega \qquad \textit{and} \qquad f(x) = g(x, x), \forall x \in \Omega.$$

- The majorization algorithm corresponding to g updates x at iteration l by

$$x^{(l+1)} \in \arg\min_{x \in \Omega} g(x, x^{(l)}), \tag{7}$$

unless we already have

$$x^{(l)} \in \arg\min_{x \in \Omega} g(x, x^{(l)}),$$

in which case the algorithm stops.

- $f(x^{(l+1)}) \leq g(x^{(l+1)}, x^{(l)}) \leq g(x^{(l)}, x^{(l)}) = f(x^{(l)}).$

- MM algorithms can be thought of as a generalization of the *EM algorithm.*

Sparse Generalized Eigenvalue Algorithm

Proposition

The following function

$$
\begin{aligned}
g(\mathbf{x},\mathbf{y}) \quad = \quad & \tau\|\mathbf{x}\|_2^2 - 2\mathbf{x}^T(\mathbf{A}+\tau\mathbf{I}_n)\mathbf{y} + \mathbf{y}^T(\mathbf{A}+\tau\mathbf{I}_n)\mathbf{y} + \rho_\varepsilon \sum_{i=1}^{n} \log(\varepsilon + |y_i|) \\
& + \rho_\varepsilon \sum_{i=1}^{n} \frac{|x_i| - |y_i|}{|y_i| + \varepsilon}, \qquad (8)
\end{aligned}
$$

majorizes the objective function in (6).

By following the minimization step in (7) with g as in (8), the *sparse GEV algorithm* is obtained as

$$
\begin{aligned}
\mathbf{x}^{(l+1)} = \arg\min_{\mathbf{x}} \quad & \tau\|\mathbf{x}\|_2^2 - 2\mathbf{x}^T(\mathbf{A}+\tau\mathbf{I}_n)\mathbf{x}^{(l)} + \rho_\varepsilon \sum_{i=1}^{n} \frac{|x_i|}{|x_i^{(l)}| + \varepsilon} \\
s.t. \quad & \mathbf{x}^T\mathbf{B}\mathbf{x} \le 1, \qquad (9)
\end{aligned}
$$

which is a sequence of quadratically constrained quadratic programs (QCQPs).

Sparse Generalized Eigenvalue Program

- (9) can also be written as

$$\begin{aligned} \mathbf{x}^{(l+1)} = \arg\min_{\mathbf{x}} \quad & \|\mathbf{x} - (\tau^{-1}\mathbf{A} + \mathbf{I}_n)\mathbf{x}^{(l)}\|_2^2 + \frac{\rho}{\tau}\|\mathbf{W}^{(l)}\mathbf{x}\|_1 \\ s.t. \quad & \mathbf{x}^T\mathbf{B}\mathbf{x} \leq 1, \end{aligned} \tag{10}$$

  where $w_i^{(l)} := \frac{1}{|x_i^{(l)}|+\varepsilon}$, $\mathbf{w}^{(l)} := (w_1^{(l)}, \ldots, w_n^{(l)})$ and $\mathbf{W}^{(l)} := \mathrm{diag}(\mathbf{w}^{(l)})$.

- (10) is very similar to *LASSO* [Tibshirani, 1996] except for the *weighted ℓ_1-norm* penalty and the quadratic constraint.

- When $\mathbf{A} \succeq 0$, $\mathbf{B} = \mathbf{I}_n$ and $\tau = 0$, (9) reduces to a very simple iterative rule:

$$x_i^{(l+1)} = \frac{\left[\left|(\mathbf{A}\mathbf{x}^{(l)})_i\right| - \frac{\rho_\varepsilon}{2} w_i^{(l)}\right]_+ \mathrm{sign}((\mathbf{A}\mathbf{x}^{(l)})_i)}{\sqrt{\sum_{i=1}^n \left[\left|(\mathbf{A}\mathbf{x}^{(l)})_i\right| - \frac{\rho_\varepsilon}{2} w_i^{(l)}\right]_+^2}}, \forall\, i, \tag{11}$$

  where $[a]_+ := \max(0, a)$, which we call as DC-PCA.

Convergence Analysis

Theorem

Let $\{\mathbf{x}^{(l)}\}_{l=0}^{\infty}$ be any sequence generated by the sparse GEV algorithm in (9). Then, all the limit points of $\{\mathbf{x}^{(l)}\}_{l=0}^{\infty}$ are stationary points of the program in (5),

$$\rho_\varepsilon \sum_{i=1}^{n} \log(\varepsilon + |x_i^{(l)}|) - [\mathbf{x}^{(l)}]^T \mathbf{A}\mathbf{x}^{(l)} \to \rho_\varepsilon \sum_{i=1}^{n} \log(\varepsilon + |x_i^*|) - [\mathbf{x}^*]^T \mathbf{A}\mathbf{x}^* := L^*,$$

for some stationary point $\mathbf{x}^*$, $\|\mathbf{x}^{(l+1)} - \mathbf{x}^{(l)}\| \to 0$, and either $\{\mathbf{x}^{(l)}\}_{l=0}^{\infty}$ converges or the set of limit points of $\{\mathbf{x}^{(l)}\}_{l=0}^{\infty}$ is a connected and compact subset of $\mathscr{S}(L^*)$, where

$$\mathscr{S}(a) := \{\mathbf{x} \in \mathscr{S} : \mathbf{x}^T \mathbf{A}\mathbf{x} - \rho_\varepsilon \sum_{i=1}^{n} \log(\varepsilon + |x_i|) = -a\}$$

and $\mathscr{S}$ is the set of stationary points of (5). If $\mathscr{S}(L^*)$ is finite, then any sequence $\{\mathbf{x}^{(l)}\}_{l=0}^{\infty}$ generated by (9) converges to some $\mathbf{x}^*$ in $\mathscr{S}(L^*)$.

Convergence Analysis

Corollary

Let $\tilde{\rho} = 0$ and $\lambda_{max}(\mathbf{A}, \mathbf{B}) > 0$. Then, any sequence $\{\mathbf{x}^{(l)}\}_{l=0}^{\infty}$ generated by (9) converges to some $\mathbf{x}^*$ such that $\lambda_{max}(\mathbf{A}, \mathbf{B}) = [\mathbf{x}^*]^T \mathbf{A}\mathbf{x}^*$ and $[\mathbf{x}^*]^T \mathbf{B}\mathbf{x}^* = 1$.

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

Corollary

Let $\mathbf{A} \succeq 0$, $\tau = 0$ and $\tilde{\rho} = 0$. Then, any sequence $\{\mathbf{x}^{(l)}\}_{l=0}^{\infty}$ generated by the following algorithm

$$\mathbf{x}^{(l+1)} = \frac{\mathbf{B}^{-1}\mathbf{A}\mathbf{x}^{(l)}}{\sqrt{[\mathbf{x}^{(l)}]^T \mathbf{A}\mathbf{B}^{-1}\mathbf{A}\mathbf{x}^{(l)}}} \qquad (12)$$

converges to some $\mathbf{x}^*$ such that $\lambda_{max}(\mathbf{A}, \mathbf{B}) = [\mathbf{x}^*]^T \mathbf{A}\mathbf{x}^*$ and $[\mathbf{x}^*]^T \mathbf{B}\mathbf{x}^* = 1$.

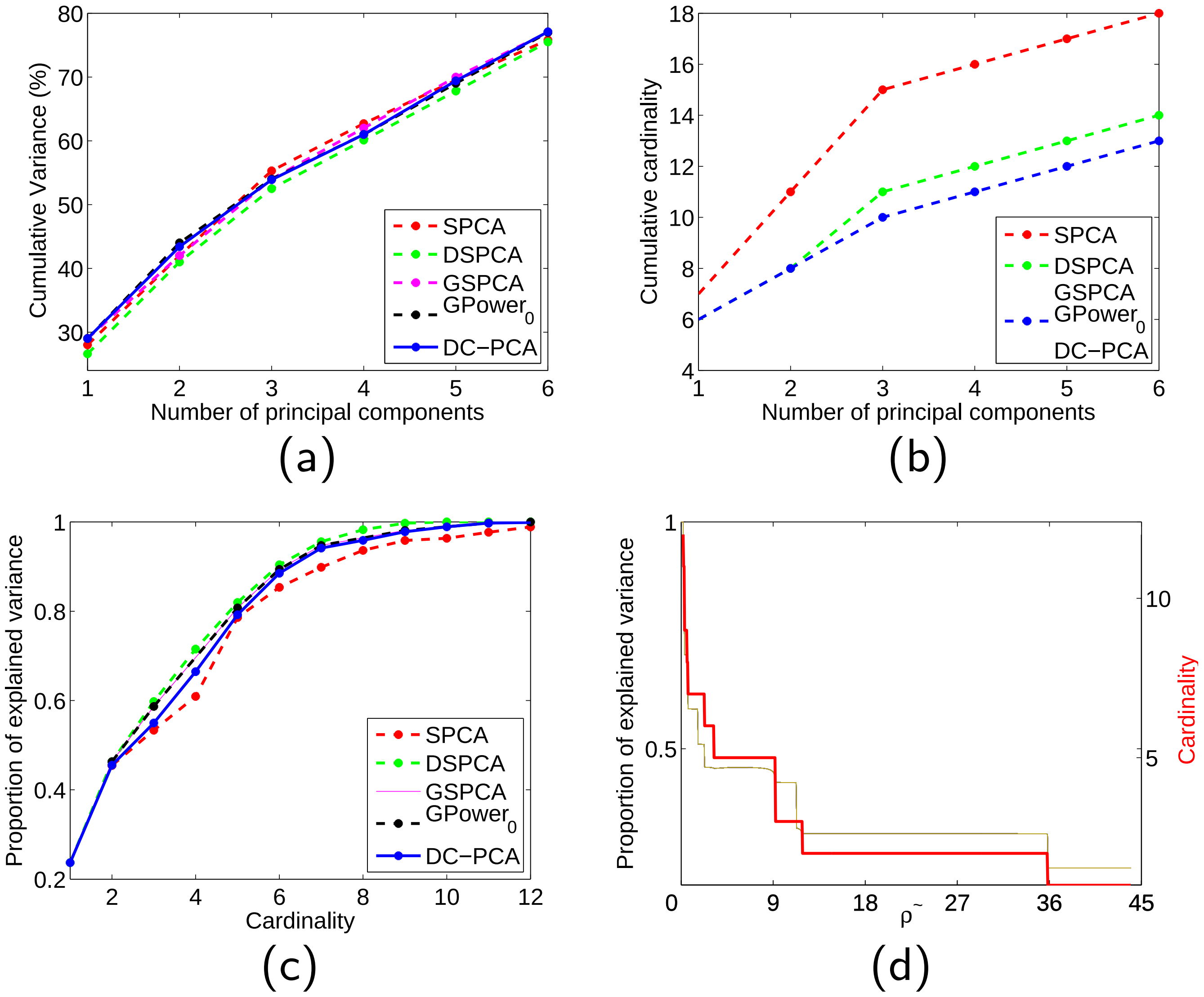
- With $\mathbf{B} = \mathbf{I}_n$, (12) reduces to the *power method* for computing $\lambda_{max}(\mathbf{A})$.

Applications: Sparse PCA

- *Sparse PCA algorithms:* Proposed (DC-PCA), SDP relaxation (DSPCA [d'Aspremont et al., 2005]), greedy approach (GSPCA [Moghaddam et al., 2007]), regression based approach (SPCA [Zou et al., 2006]) and generalized power method (GPower$_{\ell_0}$ [Journée et al., 2008]).
- *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%
GSPCA	(6,2,2,1,1,1)	13	77.1%
GPower$_{\ell_0}$	(6,2,2,1,1,1)	13	77.1%
DC-PCA	*(6,2,2,1,1,1)*	*13*	*77.1%*

Pit Props



(a) (b) (c) (d)

Figure: (a) cumulative variance and (b) cumulative cardinality for the first 6 sparse PCs; (c) proportion of explained variance (PEV) vs. cardinality for the first sparse PC; (d) dependence of sparsity and PEV on $\tilde{\rho}$ for the first sparse PC computed with DC-PCA.

Gene Datasets

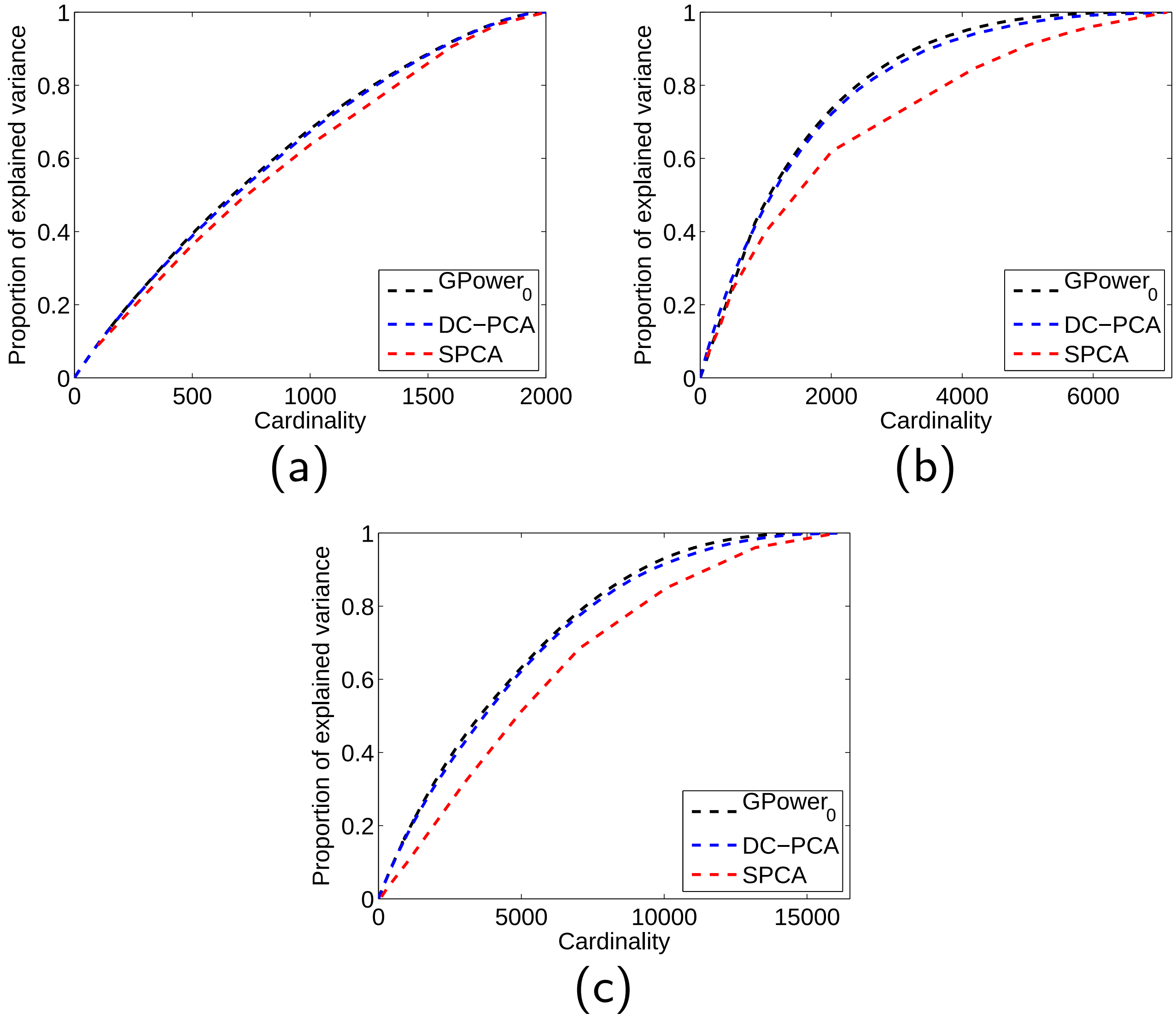
Table: Gene expression datasets

Dataset	Samples (p)	Genes (n)	Reference
Colon cancer	62	2000	[Alon et al., 1999]
Leukemia	38	7129	[Golub et al., 1999]
Ramaswamy	127	16063	[Ramaswamy et al., 2001]

Table: Computation time (in seconds) to obtain the first sparse PC, averaged over cardinalities ranging from 1 to n, for the Colon cancer, Leukemia and Ramaswamy datasets.

	Colon cancer	Leukemia	Ramaswamy
n	2000	7129	16063
SPCA	2.057	3.548	38.731
GPower$_{\ell_0}$	0.182	0.223	2.337
DC-PCA	0.034	0.156	0.547

Gene Datasets



(a)

(b)

(c)

Figure: Trade-off curves between explained variance and cardinality for (a) Colon cancer, (b) Leukemia and (c) Ramaswamy datasets. The proportion of variance explained is computed on the first sparse principal component.

Scalability

- *Complexity*
  - DC-PCA, GPower$_{\ell_0}$: $O(mn^2)$, where m is the number of iterations before convergence.
  - SPCA : $O(mn^3)$
  - GSPCA : $O(n^4)$
  - DSPCA : $O(n^4\sqrt{\log n})$
- Randomly chosen problems of size n ranging from 10 to 10000.
- Linux 3 GHz, 4 GB RAM workstation.

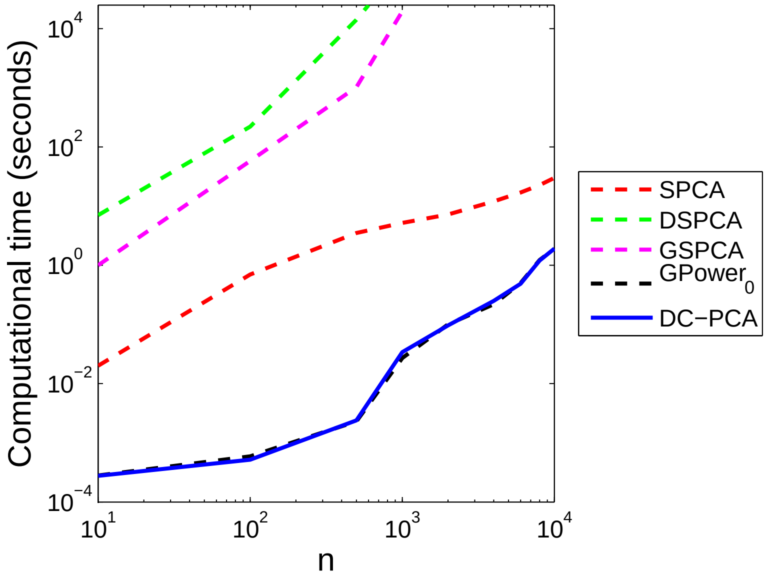


Figure: Average computation time (seconds) for the first sparse PC of **A** vs. problem size, n, over 100 randomly generated matrices **A**.

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.

- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M. K., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. (1999).
  Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring.
  Science, 286:531–537.

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

- Journée, M., Nesterov, Y., Richtárik, P., and Sepulchre, R. (2008).
  Generalized power method for sparse principal component analysis.
  http://arxiv.org/abs/0811.4724v1.

- Moghaddam, B., Weiss, Y., and Avidan, S. (2007).
  Spectral bounds for sparse PCA: Exact and greedy algorithms.
  In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, Cambridge, MA. MIT Press.

- Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J., Poggio, T., Gerald, W., Loda, M., Lander, E., and Golub, T. (2001).
  Multiclass cancer diagnosis using tumor gene expression signature.
  Proceedings of the National Academy of Sciences, 98:15149–15154.

- Tibshirani, R. (1996).
  Regression shrinkage and selection via the LASSO.
  Journal of Royal Statistical Society, Series B, 58(1):267–288.

- Zou, H., Hastie, T., and Tibshirani, R. (2006).
  Sparse principal component analysis.
  Journal of Computational and Graphical Statistics, 15:265–286.