Generalized Power Method for Sparse Principal Component Analysis

Peter Richtárik

CORE/INMA – Catholic University of Louvain – Belgium

VOCAL 2008, Veszprém, Hungary

CORE Discussion Paper #2008/70

joint work with M. Journée, Yu. Nesterov and R. Sepulchre

1. Outline

- Sparse PCA
- Optimization reformulations
- Algorithm and complexity analysis
- Numerical experiments

2. Sparse PCA (sPCA)

- **Input:** Matrix $A = [a_1, \ldots, a_n] \in \mathbf{R}^{p \times n}, \quad p \leq n$
- **Goal:** Find unit-norm vector $z^* \in \mathbf{R}^n$ which simultaneously
  1. **maximizes variance** $z^T A^T A z$
  2. is **sparse**

If sparsity is **not** required, z^* is the **dominant right singular vector** of A:

$$\max_{z^T z \leq 1} z^T A^T A z = \lambda_{\mathsf{max}}(A^T A) = (\sigma_{\mathsf{max}}(A))^2.$$

Extracting more components: Discussion above is about the **single-unit case ($m = 1$)**. Often more components (sparse dominant singular directions) are needed: **block case ($m > 1$)**.

Applications: gene expression, finance, data visualization, signal processing, vision, ...

3. Our approach to sPCA

1. Formulate sPCA as an optimization problem with **sparsity-inducing penalty** (ℓ_1 or ℓ_0) controlled by a single parameter
2. **Reformulate** to get problem of a **suitable form**:
   - suitable for analysis
   - suitable for computation
3. "Solve" reformulation using a simple **gradient scheme**
4. Recover solution of the original problem

Will illustrate steps 1) and 2) and 4) on the single-unit ℓ_1 penalized case and then jump to general analysis of step 3).

4. Three observations about the ℓ_1 penalty

Notation: $\|z\|_1 = \sum_i |z_i|$.

Penalty formulation of single-unit sPCA:

$$\phi_{\ell_1}(\gamma) \stackrel{\text{def}}{=} \max_{z^T z \le 1} \sqrt{z^T A^T A z} - \gamma \|z\|_1. \tag{1}$$

Observations:

1. $\gamma = 0 \Rightarrow$ **no reason to expect zero coordinates in** z^*
2. $\gamma \ge \|a_{i^*}\|_2 \stackrel{\text{def}}{=} \max_i \|a_i\|$**, then** $z^* = 0$**.** Indeed, since

$$\begin{aligned} \max_{z \ne 0} \frac{\|Az\|_2}{\|z\|_1} &= \max_{z \ne 0} \frac{\|\sum_i z_i a_i\|_2}{\|z\|_1} \\ &\le \max_{z \ne 0} \frac{\sum_i |z_i| \|a_i\|_2}{\sum_i |z_i|} = \max_i \|a_i\|_2. \end{aligned}$$

3. In fact, $\gamma \ge \|a_i\|_2 \;\Rightarrow\; z_i^*(\gamma) = 0$ **for all** i

5. Reformulation

Note that:

$$\phi_{\ell_1}(\gamma) = \max_{z\in\mathcal{B}^n} \|Az\|_2 - \gamma\|z\|_1 = \max_{z\in\mathcal{B}^n} \max_{x\in\mathcal{B}^p} x^T Az - \gamma\|z\|_1$$
$$= \max_{x\in\mathcal{B}^p} \max_{z\in\mathcal{B}^n} \sum_{i=1}^{n} z_i(a_i^T x) - \gamma|z_i|.$$

For fixed x, the inner max-problem has the closed-form solution

$$z_i = \text{sign}(a_i^T x)[|a_i^T x| - \gamma]_+, \qquad z^* = z/\|z\|_2.$$

Hence to solve (1), we only need to **solve this reformulation:**

$$\boxed{\phi_{\ell_1}^2(\gamma) = \max_{\substack{x\in\mathbf{R}^p \\ x^T x=1}} \sum_{i=1}^{n} [|a_i^T x| - \gamma]_+^2,} \tag{2}$$

Note: The objective function of (2) is **convex** and **smooth** and the **feasible region is in $\mathbf{R}^p$ instead of $\mathbf{R}^n$ ($p \ll n$)**.

6. Single-unit sPCA via ℓ_0 penalty

Similar story as in the ℓ_1 case, so only briefly:

Notation: $\|z\|_0 = \mathrm{Card}\{i \ : \ z_i \neq 0\}$.

Penalty formulation:

$$\phi_{\ell_0}(\gamma) \stackrel{\text{def}}{=} \max_{z^T z \leq 1} z^T A^T A z - \gamma \ \|z\|_0, \tag{3}$$

To solve (3), first **solve this reformulation:**

$$\boxed{\phi_{\ell_1}(\gamma) = \max_{\substack{x \in \mathbf{R}^p \\ x^T x = 1}} \sum_{i=1}^{n} [(a_i^T x)^2 - \gamma]_+,} \tag{4}$$

and then set

$$z_i = [\mathrm{sign}((a_i^T x)^2 - \gamma)]_+ a_i^T x, \qquad z^* = z/\|z\|_2.$$

7. Maximizing convex functions

Problems (2) and (4) (and their block generalizations) are of the form

$$\boxed{f^* = \max_{x \in \mathcal{Q}} f(x),} \qquad \text{(P)}$$

where

- $\mathbf{E}$ is a finite-dimensional vector space,
- $f : \mathbf{E} \to \mathbf{R}$ is a **convex function,**
- $\mathcal{Q} \subset \mathbf{E}$ is **compact.**

In particular,

- $\mathcal{Q} =$ unit **Euclidean sphere** in $\mathbf{R}^p$ / Single-unit case ($m = 1$)
- $\mathcal{Q} =$ **Stiefel manifold** in $\mathbf{R}^{p \times m}$, i.e. the set of $p \times m$ matrices with orthonormal columns / Block case ($m > 1$)

How to solve (P)?

8. Gradient algorithm

We solve (P) using this simple **gradient method:**

1. **Input:** Initial iterate $x_0 \in \mathcal{Q}$
2. **For** $k \geq 0$ **repeat**
   - $x_{k+1} \in \operatorname{Arg\,max}\{f(x_k) + \langle f'(x_k), y - x_k \rangle \mid y \in \mathcal{Q}\}$
   - $k \leftarrow k + 1$

This algorithm generalizes the **power method** for computing the largest eigenvalue of a symmetric positive definite matrix C:

$$f(x) = \tfrac{1}{2} x^T C x \quad \rightarrow \quad x_{k+1} = \frac{C x_k}{\|C x_k\|_2}.$$

Hence **"Generalized Power Method" (**GPower**).**

9. Iteration complexity: basic result

At any point $x \in \mathcal{Q}$ we introduce a **measure for the first-order optimality conditions:**

$$\Delta(x) \stackrel{\text{def}}{=} \max_{y \in \mathcal{Q}} \langle f'(x), y - x \rangle.$$

Clearly, $\Delta(x) \geq 0$ and it vanishes only at the points where the gradient $f'(x)$ belongs to the normal cone to $\mathrm{Conv}(\mathcal{Q})$ at x.

Denote $\Delta_k \stackrel{\text{def}}{=} \min_{0 \leq i \leq k} \Delta(x_i)$.

Theorem Let sequence $\{x_k\}_{k=0}^{\infty}$ be generated by GPower as applied to a convex function f. Then the sequence $\{f(x_k)\}_{k=0}^{\infty}$ is **monotonically increasing** and $\lim_{k \to \infty} \Delta(x_k) = 0$. Moreover,

$$\Delta_k \leq \frac{f^* - f(x_0)}{k+1}. \tag{5}$$

10. Strong convexity of functions and sets

Function f is strongly convex if there exists a constant $\sigma_f > 0$ such that for any $x, y \in \mathbf{E}$

$$f(y) \geq f(x) + \langle f'(x), y - x \rangle + \frac{\sigma_f}{2}\|y - x\|^2.$$

The set $\mathrm{Conv}(\mathcal{Q})$ is strongly convex if there exists a constant $\sigma_{\mathcal{Q}} > 0$ such that for any $x, y \in \mathrm{Conv}(\mathcal{Q})$ and $\alpha \in [0, 1]$ the following inclusion holds:

$$\alpha x + (1 - \alpha)y + \frac{\sigma_{\mathcal{Q}}}{2}\alpha(1 - \alpha)\|x - y\|^2 \cdot \mathcal{S} \subset \mathrm{Conv}(\mathcal{Q}).$$

Theorem If $f : \mathbf{E} \to \mathbf{R}$ is nonnegative, has $\sigma_f > 0$ and f' is L_f-Lipschitz, then for any $\omega > 0$, the level set

$$\mathcal{Q}_\omega \stackrel{\text{def}}{=} \{x \mid f(x) \leq \omega\}$$

is strongly convex with parameter $\sigma_{\mathcal{Q}_\omega} = \sigma_f / \sqrt{2\omega L_f}$.

11. Refined analysis under strong convexity

Theorem

Let

- f be convex with strong convexity parameter $\sigma_f \geq 0$, and
- $\mathrm{Conv}(\mathcal{Q})$ be convex with strong convexity parameter $\sigma_{\mathcal{Q}} \geq 0$.

If $0 < \delta_f = \inf_{x \in \mathcal{Q}} \|f'(x)\|_*$ and either $\sigma_f > 0$ or $\sigma_{\mathcal{Q}} > 0$, then

$$\sum_{k=0}^{N} \|x_{k+1} - x_k\|^2 \leq \frac{2(f^* - f(x_0))}{\sigma_{\mathcal{Q}} \delta_f + \sigma_f}.$$

Note: If f is *not* minimized on $\mathcal{Q}$, then $\delta_f > 0$.

12. Computational experiments

We compare the following **Sparse PCA algorithms:**

GPower$_{\ell_1}$	Single-unit sparse PCA via ℓ_1-penalty [1]
GPower$_{\ell_0}$	Single-unit sparse PCA via ℓ_0-penalty [1]
GPower$_{\ell_1,m}$	Block sparse PCA via ℓ_1-penalty [1]
GPower$_{\ell_0,m}$	Block sparse PCA via ℓ_0-penalty [1]
SPCA	SPCA algorithm [2]
Greedy*	Greedy method [3]
rSVD$_{\ell_1}$	Method [4] with ℓ_1-penalty ("soft thresholding")
rSVD$_{\ell_0}$	Method [4] with ℓ_0-penalty ("hard thresholding")

*Greedy slows down dramatically, compared to the other methods, if aimed at obtaining a component of higher cardinality.

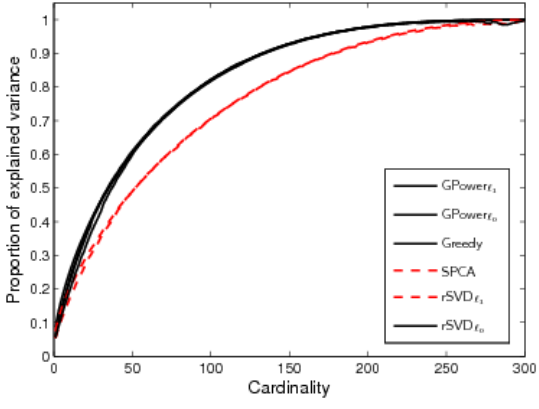
Test Problems:

- Randomly generated

  $A =$ Gaussian with zero mean and unit variance

- Gene-expression data

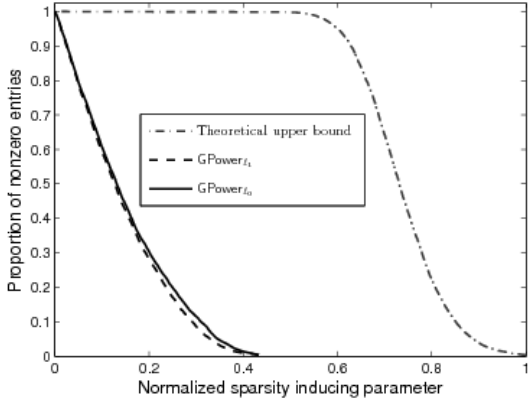
13. Trade-off curves



Trade-off between **explained variance** and **cardinality**. The algorithms aggregate in two groups. The methods GPower$_{\ell_1}$, GPower$_{\ell_0}$, Greedy and rSVD$_{\ell_0}$ do better **(black solid lines)**, and SPCA and rSVD$_{\ell_1}$ do worse **(red dashed lines)**.

Based on 100 random test problems of size $p = 100, n = 300$.

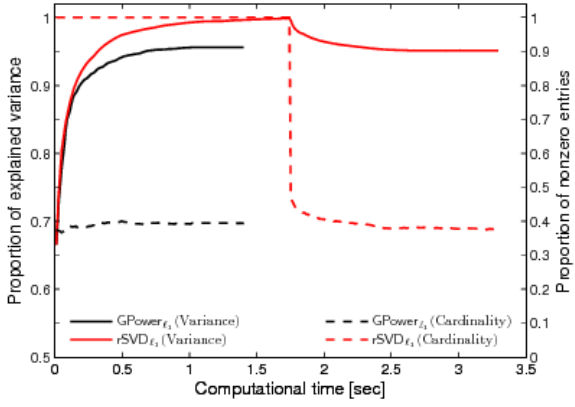
14. Controlling sparsity with γ



Dependence of **cardinality** on the value of the **sparsity-inducing parameter** γ. The horizontal axis shows a normalized interval of reasonable values of γ. The vertical axis shows percentage of nonzero coefficients of the resulting sparse loading vector z^*.

Based on 100 random test problems of size $p = 100, n = 300$.

15. How does the trade-off evolve in time?



Evolution of the **explained variance** (solid lines and left axis) and **cardinality** (dashed lines and right axis) in time for the methods GPower$_{\ell_1}$ and rSVD$_{\ell_1}$.

Based on random test problem of size $p = 250$ and $n = 2500$.

16. Random data: speed

Fixed n/p ratio:

$p \times n$	250×2500	500×5000	750×7500	1000×10000
GPower$_{\ell_1}$	0.85	2.61	3.89	5.32
GPower$_{\ell_0}$	0.46	1.21	2.41	2.93
SPCA	2.77	14.0	41.0	81.6
rSVD$_{\ell_1}$	1.40	6.80	17.8	41.2
rSVD$_{\ell_0}$	1.33	6.20	15.4	36.3

Fixed p, growing n:

$p \times n$	500×2000	500×4000	500×8000	500×16000
GPower$_{\ell_1}$	0.97	1.96	4.30	8.43
GPower$_{\ell_0}$	0.39	0.97	2.01	4.63
SPCA	7.37	11.4	22.4	44.6
rSVD$_{\ell_1}$	2.56	5.27	11.3	26.8
rSVD$_{\ell_0}$	2.30	4.70	10.3	23.8

17. Gene expression data: speed

Data sets (breast cancer cohorts):

Study	Samples (p)	Genes (n)	Reference
Vijver	295	13319	van de Vijver et al. [2002]
Wang	285	14913	Wang et al. [2005]
Naderi	135	8278	Naderi et al. [2006]
JRH-2	101	14223	Sotiriou et al. [2006]

Speed (in seconds):

	Vijver	Wang	Naderi	JRH-2
GPower$_{\ell_1}$	7.72	6.96	2.15	2.69
GPower$_{\ell_0}$	3.80	4.07	1.33	1.73
GPower$_{\ell_1,m}$	5.40	4.37	1.77	1.14
GPower$_{\ell_0,m}$	5.61	7.21	2.25	1.47
SPCA	77.7	82.1	26.7	11.2
rSVD$_{\ell_1}$	46.4	49.3	13.8	15.7
rSVD$_{\ell_0}$	46.8	48.4	13.7	16.5

18. Gene expression data: content

PEI-values based on 536 cancer-related pathways:

	Vijver	Wang	Naderi	JRH-2
PCA	0.0728	0.0466	0.0149	0.0690
GPower$_{\ell_1}$	**0.1493**	0.1026	0.0728	0.1250
GPower$_{\ell_1}$	0.1250	0.1250	0.0672	0.1026
GPower$_{\ell_1,m}$	0.1418	0.1250	**0.1026**	**0.1381**
GPower$_{\ell_0,m}$	0.1362	**0.1287**	0.1007	0.1250
SPCA	0.1362	0.1007	0.0840	0.1007
rSVD$_{\ell_1}$	0.1213	0.1175	0.0914	0.0914
rSVD$_{\ell_0}$	0.1175	0.0970	0.0634	0.1063

Pathway Enrichment Index (PEI) measures the statistical significance of the overlap between two kinds of gene sets.

19. Summary

We have

- developed **4 reformulations** (single unit/block $\times$ ℓ_1/ℓ_0) of the sPCA problem which enabled us to
  - devise a very **fast method** (we work in dimension $p \ll n$ and use only gradients), and
  - analyze the **iteration complexity** of the method;
- analyzed a simple gradient method **(Generalized Power Method)** for maximizing convex functions on compact sets;
- applied GPower to 4 reformulations and ended-up with 4 algorithms for sPCA;
- tested our algorithms on random and gene expression data:
  - they **outperform other methods significantly in speed** (finish before some other algorithms initialize),
  - for the biological data, they produce slightly higher quality of solution in terms of PEI.

20. References

[1] M. Journée, Yu. Nesterov, P. Richtárik, R. Sepulchre. **Generalized Power Method for Sparse Principal Component Analysis (this talk).** *submitted to Journal of Machine Learning Research*, November 2008.

[2] H. Zou, T. Hastie, R. Tibshirani. **Sparse Principal Component Analysis.** *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.

[3] A. d'Aspremont, F. R. Bach, L. El Ghaoui. **Optimal Solutions for Sparse Principal Component Analysis.** *Journal of Machine Learning Research*, 9:1269–1294, 2008.

[4] H. Shen, J. Z. Huang. **Sparse Principal Component Analysis via Regularized Low Rank Matrix Approximation.** *Journal of Multivariate Analysis*, 99(6):1015–1034, 2008.