Sparse Principal Component Analysis

Formulations And Algorithms

Thomas Rusch Norbert Walchhofer, Department of Finance, Accounting and Statistics
WU Vienna
June 20, 2011

Outline

- Background
  - Reviw of Principal Component Analysis (PCA)?
- Generalized Power Method for Sparse PCA
- Problem Formulations and Reformulations
  - Single-unit sparsePCA
    - Single-unit sparsePCA via ℓ_1-penalty
    - Single-unit sparsePCA via ℓ_0-penalty
  - Block sparsePCA
  - Power Method
- Proposed Algorithms and their Evaluation
  - Examplary Algorithm

Principle Component Analysis - Motivation

- Method for dimension reduction
- Orthogonal transformation of possibly correlated variables into uncorrelated principal components
- Project a centered data matrix A or a (sample) covariance matrix thereof $\Sigma = A^T A$ from $\boldsymbol{R}^p$ into $\boldsymbol{R}^m$ where $q \leq p$
- Aims at finding a few linar combinations the p variables, pointing in orthogonal directions explaining as much variance as possible.

Principal Component Analysis - Problem

PCA - Formulation

$$z^\star = \max_{z^T z \leq 1} z^T \Sigma z$$

Extracting the first principal component can be done in two ways:

- computing the first eigenvector of Σ
- or the first right singular value of A.

Usually principal components are linear combinations of all input variables with loading vector $z^\star$ (score).

Principal Component Analysis - Disadvantage

PCA aimes to reduce complexity, however there are some drawbacks:

- principal components depend on many variables
- interpretation of components can be agonizing
- individual loadings can be negligible

Sparse PCA - Motivation

Sparse PCA simplifies mass of loadings and therefore

- highlights the most essential structures,
- is easier to interpret,
- amount of input variables can be controled for
- and it provides a reasonable *trade-off* between **explained variance** and **usability**.

Overview

Journée et al. (2010) provide following contributions:

- Formulations of for single-unit sparse PCA via ℓ_1 & cardinality (ℓ_0)-penalty
- Formulations of for block sparse PCA via ℓ_1 & cardinality-penalty
- Reformulations to **convex** optimization problems
- Application of the Power Method for sparse PCA
- Development of algorithms to solve the reformulated optimization problems

Single-unit optimization tries to find sparse loadings for one principal component, before calculating the next one.

Consider following optimization problem

$$\Phi_{\ell_1}(\gamma) \stackrel{\text{def}}{=} \max_{z \in \mathcal{B}^n} \sqrt{z^T \Sigma z} - \gamma \|z\|_1 \tag{1}$$

with sparsity-controlling parameter $\gamma \geq 0$ and sample covariance matrix $\Sigma = A^T A$.
By setting $\gamma = 0$ there can be shown that $\Phi_{\ell_1}(0)$ leads to

$$\gamma < \|a_{i^\star}\|_2, \tag{2}$$

defining the upper bound for γ where $i^\star$ is obtained by $\max_i \|a_i\|_2$

Reformulation I

Reformulating the problem

$$
\begin{aligned}
\Phi_{\ell_1}(\gamma) &= \max_{z \in B^n} \|Az\|_2 - \gamma \|z\|_1 \\
&= \max_{z \in B^n} \max_{x \in B^n} x^T A z - \gamma \|z\|_1 && (3) \\
&= \max_{x \in B^n} \max_{z \in B^n} \sum_{i=1}^{n} z_i (a_i^T x) - \gamma \|z\|_1 \\
&= \max_{x \in B^n} \max_{z' \in B^n} \sum_{i=1}^{n} |z_i'| (|a_i^T x| - \gamma) && (4)
\end{aligned}
$$

where $z_i = \mathrm{sign}(a_i^T x) z_i'$.
Equation 2 proofs that there is a $x \in B^n$ for which $a_i^T x > \gamma$.

Reformulation II

Further reformulating the problem

In view of 2, there is some $x \in \mathrm{B}^{\mathrm{n}}$ for which $a_i^T x > \gamma$. By fixing x, solving the inner maximization problem for z' we obtain a closed solution for $z^\star$:

$$z_i^\star = z_i^\star(\gamma) = \frac{\mathrm{sign}(a_i^T x)\left[|a_i^T x| - \gamma\right]_+}{\sqrt{\sum_{k=1}^{n}\left[|a_k^T x| - \gamma\right]_+^2}}, \qquad i = 1, \ldots, n. \tag{5}$$

Reformulation III

Adjusting the objective function

Therefore Eq. 4 can be written as

$$\Phi_{\ell_1}^2(\gamma) = \max_{x \in S^p} \sum_{i=1}^{n} \left[|a_i^T x| - \gamma \right]_+^2 . \tag{6}$$

This results in a differentiable and **convex** objective function, where all local and global maximal must lie in den Eucledian sphere S^p , **reducing the search space** of our inital problem formulation (see Eq. 8) to dimension p with $p \ll n$!

Reformulation Summary

What really happend...

- By introducing a vector x the optimization problem is split in two, solving x and z, respecitvely.
- x is solved in Eq. 6 providing a sparsity pattern for $z^\star$.
- This sparsity pattern indicates which z_i are active, i.e. are not 0.
- Therefore loadings only have to be calculated for p of the n variables of A (for one component).

Optimization Problem

In contrast to the ℓ_1-penalty (soft constraint) the ℓ_0 or cardinality-penalty directly penalizes the number of non-zero components of vector z (hard constraint).

Optimization problem formulated in d'Aspremont et al. (2008)

$$\Phi_{\ell_0}(\gamma) \overset{\text{def}}{=} \max_{z \in B^n} \sqrt{z^T \Sigma z} - \gamma \|z\|_0 \tag{7}$$

Analogue to the ℓ_1 case with derive the boundary for γ, optimization for $z_i^\star$ and x:

$$\gamma < \|a_{i^\star}\|_2^2,$$

$$z_i^\star = z_i^\star(\gamma) = \frac{\left[\operatorname{sign}(a_i^T x)^2 - \gamma\right]_+ a_i^T x}{\sqrt{\left[\operatorname{sign}(a_i^T x)^2 - \gamma\right]_+ (a_i^T x)^2}}, \qquad i = 1, \ldots, n,$$

$$\Phi_{\ell_1}^2(\gamma) = \max_{x \in S^p} \sum_{i=1}^{n} \left[(a_i^T x)^2 - \gamma\right]_+ .$$

Optimization Problem

Block optimization tries to find sparse loadings for m principal components.

Consider following generalization of Eq. 3

$$\Phi_{\ell_1,m}(\gamma) \stackrel{\text{def}}{=} \max_{X \in \mathrm{S}^{\mathrm{p}}_{\mathrm{m}}} \max_{Z \in [S^n]^m} \mathrm{Tr}(X^T AZN) - \sum_{j=1}^{m} \gamma_j \sum_{i=1}^{n} |z_i j| \tag{8}$$

where $\gamma = [\gamma_1, \dots, \gamma_m]^T \quad \forall \gamma_j \geq 0$ and $N = \mathrm{Diag}(\mu_1, \dots, \mu_m) \quad \forall \mu_j > 0$.

Each γ_j controls the sparsity for the corresponding component. For positve γ_j columns of Z are not expected to be orthogonal anymore! Note that distinct values of μ_j ensure the columns of $X^\star$ being the dominant m components, while also pursing more sparse and orthogonal vectors.

Reformulation

Since the columns of Z are decoupled the reformulation can be done analogue to the single-unit case. Hence, for every column of X every row element is optimzed, indicating the 'active-status' for each component of Z (i.e. variable of A) of each row Z.
If $\mu_j|a_i^T x_j^\star| > \gamma_j$ is fullfilled $z_ij^\star$ is active.

Power Method

The power method is a eigenvalue algorithm, given a matrix A trying to find the dominant eigenvalue λ and it corresponding eigenvector v such that $Av = \lambda v$. By avoiding a matrix decomposition it is very favorable for large sparse matricen since the computation afford is very low. The scalar $q = x^T x$ converges linearily against the dominant eigenvalue.

$$x_{k+1} = \frac{Ax_k}{\|Ax_k\|} \tag{9}$$

x_0 can be an approxmiation or a random vector. The method works under following assumptions:

- A has an eigenvalue strictly greater than others
- Starting vector x_0 has a non-zero component in the direction of the eigenvector of the dominant eigenvalue.

Generalization of the Power Method

Based on gradient method for maximizing convex functions the authors show that a convex function

$$f^{\star} = \max_{x \in \mathrm{Q}} f(x) \tag{10}$$

can iteratively maximized by a subgradient, even if that $f(x)$ is not assumed to be differentiable.
In our case we have to solve a quadratic objective function $f(x) = \frac{1}{2}x^T Cx$ for $C \in \mathrm{S}^{\mathrm{p}}_{++}$, which can be solved by

$$x_{k+1} = \frac{C_{x_k}}{\| C_{x_k} \|}, \quad k \geq 0. \tag{11}$$

Algorithm for Block sparse PCA

Algorithm 4: Block sparse PCA algorithm based on the ℓ_1-penalty (16)

input : Data matrix $A \in \mathbf{R}^{p \times n}$
Sparsity-controlling vector $[\gamma_1, \ldots \gamma_m]^T \geq 0$
Parameters $\mu_1, \ldots, \mu_m > 0$
Initial iterate $X \in S_m^p$

output: A locally optimal sparsity pattern P

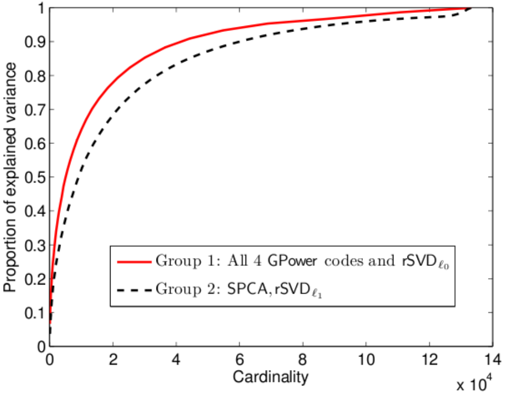
begin
- **repeat**
  - **for** $j = 1, \ldots, m$ **do**
    - $x_j \longleftarrow \sum_{i=1}^{n} \mu_j [\mu_j |a_i^T x_j| - \gamma_j]_+ \operatorname{sign}(a_i^T x) a_i$
  - $X \longleftarrow \operatorname{Polar}(X)$
- **until** *a stopping criterion is satisfied*
- Construct matrix $P \in \{0,1\}^{n \times m}$ such that $\begin{cases} p_{ij} = 1 & \text{if } \mu_j |a_i^T x_j| > \gamma_j \\ p_{ij} = 0 & \text{otherwise.} \end{cases}$

end

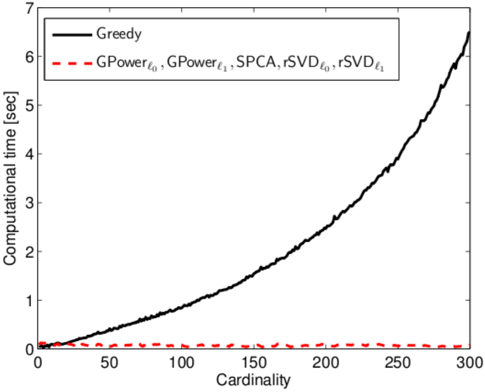
Algorithms Evaluated

- All four GPower algorithms, two single-unit and two block sparse PCA each with ℓ_0 and ℓ_1 penalty
- Greedy search algorithm of d'Aspremont et al. (2008) (non-convex)
- *SPCA* from Zhou et al. (2006) (lasso penalty)
- $rSVD_{\ell_0}$ and $rSVD_{\ell_1}$ by Shen and Huang (2008)

Evaluation on Real Data - Explained Variance



Evaluation on Real Data - Computation Time I



$p \times n$	50×500	100×1000	250×2500	500×5000	750×7500
GPower_{ℓ_1}	0.22	0.56	4.62	12.6	20.4
GPower_{ℓ_0}	0.06	0.17	2.15	6.16	10.3
$\text{GPower}_{\ell_1,m}$	0.09	0.28	3.50	12.4	23.0
$\text{GPower}_{\ell_0,m}$	0.05	0.14	2.39	7.7	12.4
SPCA	0.61	1.47	13.4	48.3	113.3
rSVD_{ℓ_1}	0.29	1.12	7.72	22.6	46.1
rSVD_{ℓ_0}	0.28	1.03	7.21	20.7	41.2

Table 8: Average computational time for the extraction of $m = 5$ components (in seconds).

Evaluation on Real Data - Conclusion

GPower algorithms show competitve behavior in terms of

- explained variance
- computation time
- control for sparsity pattern
- usability (data matrix and sample covariance matrix)

References

A. d'Aspremont, F. Bach, and L. El Ghaoui. Optimal solutions for sparse principal component analysis. *Journal of Machine Learning*, 9:1269–1294, 2008.

Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11: 517–553, 2010.

Haipeng Shen and Jianhua Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivar. Anal.*, 99:1015–1034, July 2008. ISSN 0047-259X. doi: 10.1016/j.jmva.2007.06.007.

H. Zhou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15:265–286, 2006.

Thank you for your attention

Thomas Rusch
Norbert Walchhofer

WU Wirtschaftsuniversität Wien
Augasse 2–6, A-1090 Wien