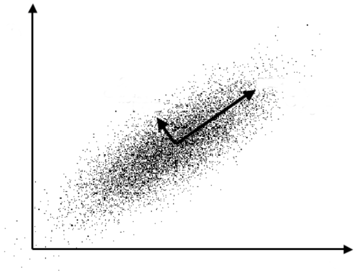
# Weakly Correlated Sparse Components with Nearly Orthonormal Loadings

GSI 2015

Matthieu Genicot — *Université Catholique de Louvain*
Wen Huang — *Université Catholique de Louvain*
Nickolay T. Trendafilov — *Open University*

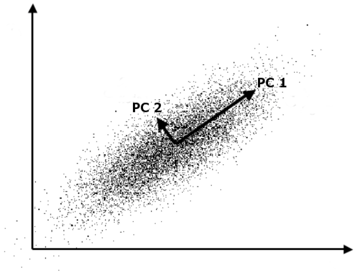
# Principal Components Analysis (PCA)

Goal of PCA: Reducing high dimensional data to a lower dimension for visualization purpose or to reveal hidden patterns



# Principal Components Analysis (PCA)

Goal of PCA: Reducing high dimensional data to a lower dimension for visualization purpose or to reveal hidden patterns

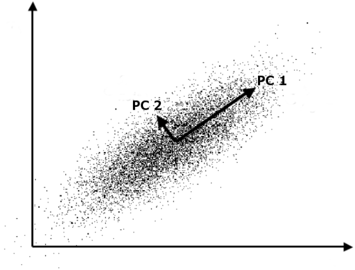



Express the data $\mathrm{X} \in \mathbb{R}^{n \times p}$ in a new space: $\mathrm{Y} = \mathrm{XA}$

- linear combinations of all the $p$ variables of $\mathrm{X}$
- orthogonal loadings $(\mathrm{A})$
- uncorrelated components $(\mathrm{Y})$

# Principal Components Analysis (PCA)

Goal of PCA: Reducing high dimensional data to a lower dimension for visualization purpose or to reveal hidden patterns




Express the data $\mathrm{X} \in \mathbb{R}^{n \times p}$ in a new space: $\mathrm{Y} = \mathrm{XA}$

- linear combinations of **all** the $p$ variables of $\mathrm{X}$

$\Rightarrow$ Difficulty to interpret the results

# Motivations for sparse PCA

**Gene expression analysis**
20000 genes, $\sim$ 200 samples
The components can have a biological interpretation

**Financial applications**
To manage the stocks efficiently
Every non-zero loading has a cost (e.g., a transaction cost)

# Motivations for sparse PCA

**Gene expression analysis**
20000 genes, $\sim$ 200 samples
The components can have a biological interpretation

**Financial applications**
To manage the stocks efficiently
Every non-zero loading has a cost (e.g., a transaction cost)

$\Rightarrow$ trade-off between statistical fidelity (i.e., variance explained) and interpretability/utility (i.e., number of variables used)

How to reduce the number of variables used for each component?

$\Rightarrow$ How to achieve sparseness?

# Motivations for sparse PCA

**Gene expression analysis**
20000 genes, $\sim$ 200 samples
The components can have a biological interpretation

**Financial applications**
To manage the stocks efficiently
Every non-zero loading has a cost (e.g., a transaction cost)

$\Rightarrow$ trade-off between statistical fidelity (i.e., variance explained) and interpretability/utility (i.e., number of variables used)

How to reduce the number of variables used for each component?

$\Rightarrow$ How to achieve sparseness?

**Other applications**
Image processing, multiscale data processing etc.

# Problem formulation

Sparse optimizers $a$ of $f(a)$ with $\ell_1$ norm:

- Weighted form: $\min f(a) + \tau \|a\|_1$, for some $\tau > 0$
- $\ell_1$-constrained form: $\min f(a)$ subject to $\|a\|_1 \leq \tau$
- Function-constrained form: $\min \|a\|_1$ subject to $f(a) \leq \bar{f}$

# Problem formulation

Classic PCA problem:

$$\begin{aligned} \underset{a_i}{\text{maximize}} \quad & f(a_i) = a_i^\top \mathrm{R} a_i \\ \text{subject to} \quad & a_i^\top a_i = 1 \\ & a_i^\top a_j = 0, \ i \neq j \end{aligned}$$

Sparse optimizers $a$ of $f(a)$ with $\ell_1$ norm:

- Weighted form: $\min f(a) + \tau \|a\|_1$, for some $\tau > 0$
- $\ell_1$-constrained form: $\min f(a)$ subject to $\|a\|_1 \leq \tau$
- Function-constrained form: $\min \|a\|_1$ subject to $f(a) \leq \bar{f}$

# Problem formulation

Classic PCA problem:

$$\begin{aligned} \underset{a_i}{\text{maximize}} \quad & f(a_i) = a_i^\top \mathrm{R} a_i \\ \text{subject to} \quad & a_i^\top a_i = 1 \\ & a_i^\top a_j = 0, \ i \neq j \end{aligned}$$

Sparse optimizers $a$ of $f(a)$ with $\ell_1$ norm for sparse PCA:

- **Weighted form**: $\max a^\top Ra + \tau \|a\|_1$, for some $\tau > 0$
- $\ell_1$-constrained form: $\min f(a)$ subject to $\|a\|_1 \leq \tau$
- Function-constrained form: $\min \|a\|_1$ subject to $f(a) \leq \bar{f}$

## Problem formulation

Classic PCA problem:

$$\begin{aligned} \underset{a_i}{\text{maximize}} \quad & f(a_i) = a_i^\top \mathrm{R} a_i \\ \text{subject to} \quad & a_i^\top a_i = 1 \\ & a_i^\top a_j = 0, \; i \neq j \end{aligned}$$

Sparse optimizers $a$ of $f(a)$ with $\ell_1$ norm for sparse PCA:

- Weighted form: $\max a^\top Ra + \tau \|a\|_1$, for some $\tau > 0$
- $\ell_1$-**constrained form**: $\max a^\top Ra$ subject to $\|a\|_1 \leq \tau$
- Function-constrained form: $\min \|a\|_1$ subject to $f(a) \leq \bar{f}$

# Problem formulation

Classic PCA problem:

$$\begin{aligned} \underset{a_i}{\text{maximize}} \quad & f(a_i) = a_i^\top \mathrm{R} a_i \\ \text{subject to} \quad & a_i^\top a_i = 1 \\ & a_i^\top a_j = 0,\ i \neq j \end{aligned}$$

Sparse optimizers $a$ of $f(a)$ with $\ell_1$ norm for sparse PCA:

- Weighted form: $\max a^\top Ra + \tau \|a\|_1$, for some $\tau > 0$
- $\ell_1$-constrained form: $\max a^\top Ra$ subject to $\|a\|_1 \leq \tau$
- **Function-constrained form**: $\min \|a\|_1$ subject to $a^\top Ra \geq \lambda_{max} - \epsilon$

Trendafilov (2014)

# Problem formulation

Optimization on the oblique manifold $\mathcal{OB}(p, r)$:

$$\min_{\mathcal{OB}(p,r)} \|A\|_1 + \mu\|A^\top RA - D^2\|_F^2$$

Make the problem smooth:

$$\|A\|_1 \approx \sum_{ij} \left(\sqrt{A_{ij}^2 + \epsilon^2} - \epsilon\right)$$

At the minimum:

$$A^\top RA \approx D, \text{ then } A^\top A \approx I$$

Final cost function:

$$\min_{\mathcal{OB}(p,r)} \sum_{ij} \left(\sqrt{A_{ij}^2 + \epsilon^2} - \epsilon\right) + \mu\|A^\top RA - D^2\|_F \ ,$$

# Tests

**Dataset**

Real DNA methylation dataset available online on the NCBI website
2000 genes randomly selected and $\sim 150$ samples

Tests with 10 components

**Measures of interest**

- Variance explained
- Correlation of the Components
- Orthogonality of the loadings

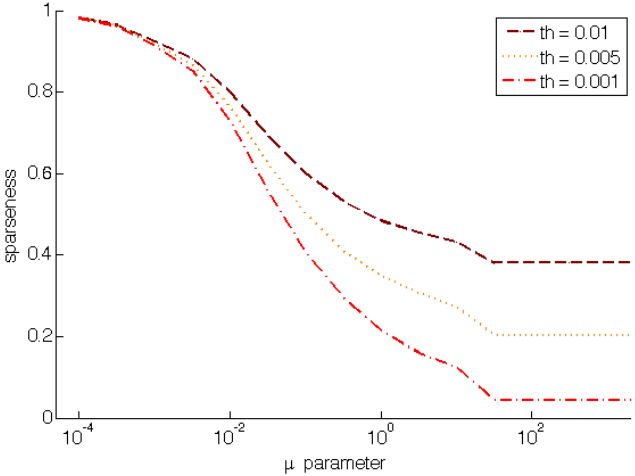
Comparison to the method of Journée et al. (2010) with both $\ell_0$ and $\ell_1$ norms

# Sparseness

Drawback: Not exactly zero values

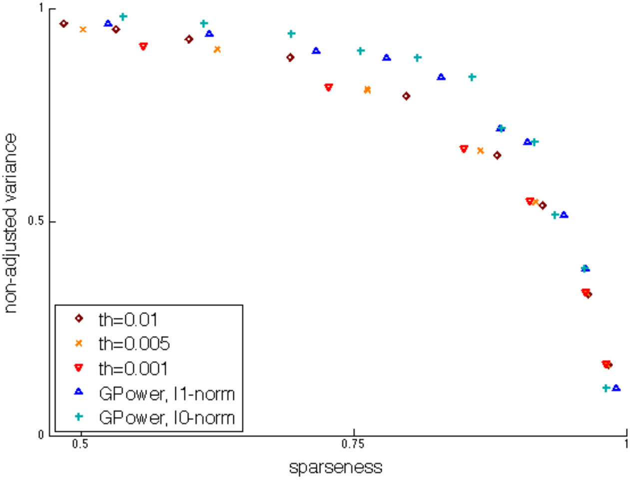
# Sparseness

Drawback: Not exactly zero values



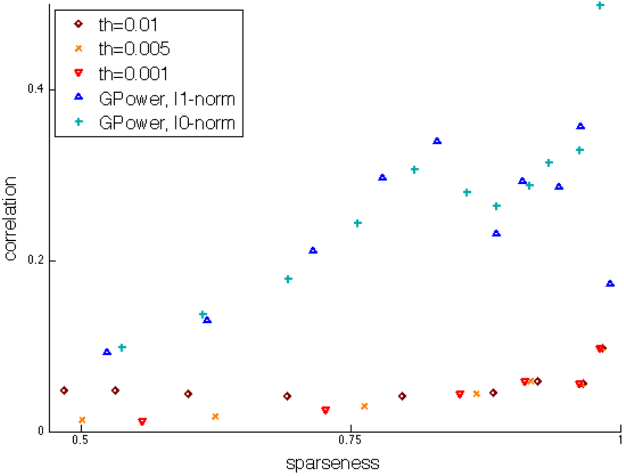
# Naive variance explained

$\mathrm{tr}(A^\top X^\top XA)$



# Correlation

$$\| A^\top XX^\top A - \mathrm{diag}(A^\top XX^\top A) \|_F$$



# Adjusted variance explained

Zou et al. 2006

**Orthogonal loadings**: $\text{tr}(A^\top X^\top XA)$

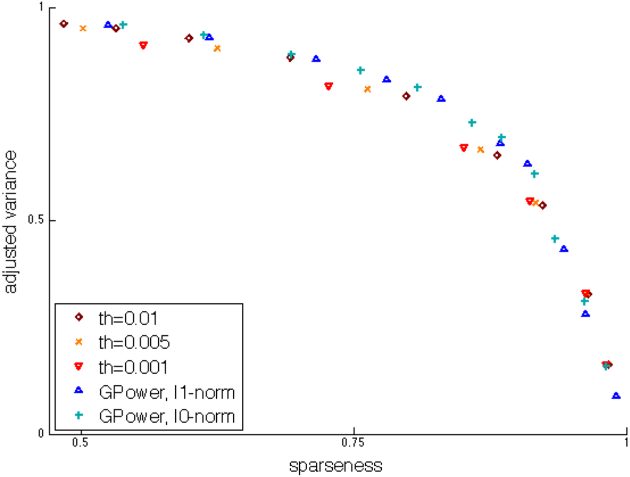
**Non-orthogonal loadings**
For component $i$: Remove variance already explained by components $1, \ldots (i-1)$

Project component $i$ on the space spanning by components $1, \ldots (i-1)$
$\rightarrow$ QR decomposition: $\mathrm{Y} = QR$, with $\mathrm{Y} = \mathrm{XA}$

Residual variance after adjustment: $\text{tr}(R^2)$

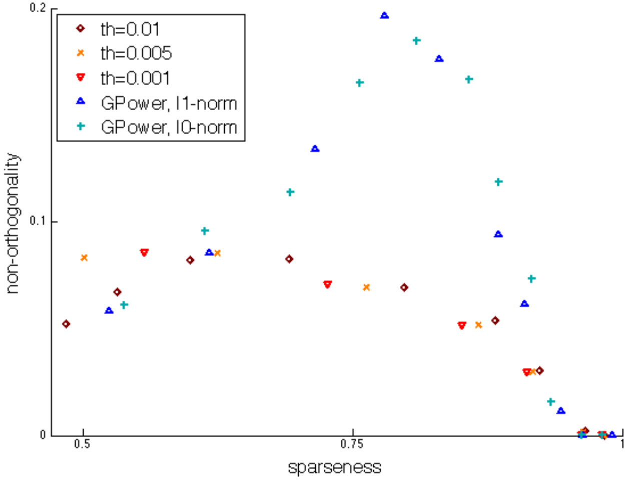
# Adjusted variance explained

Zou et al. 2006



# Orthogonality

$$\| A^\top A - \text{diag}(A^\top A) \|_F$$



# Take-Home message & further work

**Motivation**

With larger and larger datasets collected, sparseness in PCA is more and more needed

**Results**

Our method

- can explain a large part of the variance in the data
- outperforms Journée's method for the uncorrelation between the components
- outperforms Journée's method for the orthogonality of the loadings

**Further work**

Tests at a larger-scale are needed and comparison with more methods are needed