On Interpretability in Data Analytics

Emilio Carrizosa, IMUS - Instituto de Matemáticas de la Universidad de Sevilla

Nizhny Novgorod, 21st November 2020

Contents

The team

- Sandra Benítez-Peña, US
- Rafael Blanquero, US
- Emilio Carrizosa, US
- Marcela Galvis, CBS
- Vanesa Guerrero-Lozano, UC3M
- M. Asunción Jiménez-Cordero, UMa
- Kseniia Kurishchenko, CBS
- Belén Martín-Barragán, UEd
- Cristina Molero-Río, US
- Alba V. Olivares-Nadal, Chicago Booth
- Pepa Ramírez-Cobo, UCa
- Dolores Romero Morales, CBS
- Remedios Sillero-Denamiel, US

The tool

$$\min_{x \in \mathcal{X}} f(x)$$

The team

- Sandra Benítez-Peña, US
- Rafael Blanquero, US
- Emilio Carrizosa, US
- Marcela Galvis, CBS
- Vanesa Guerrero-Lozano, UC3M
- M. Asunción Jiménez-Cordero, UMa
- Kseniia Kurishchenko, CBS
- Belén Martín-Barragán, UEd
- Cristina Molero-Río, US
- Alba V. Olivares-Nadal, Chicago Booth
- Pepa Ramírez-Cobo, UCa
- Dolores Romero Morales, CBS
- Remedios Sillero-Denamiel, US

The tool

$$\min_{x \in \mathcal{X}} f(x)$$

Introduction

Mathematical Optimization

$$\min_{x \in \mathcal{X}} f(x)$$

- x^* sought with $f(x^*) \leq f(x) \quad \forall x \in \mathcal{X}$
- $\mathcal{X} \subset \mathbb{R}^n$

Taxonomy (neos Guide)

$$\min_{x \in \mathcal{X}} f(x)$$

- x^* sought with $f(x^*) \leq f(x) \quad \forall x \in \mathcal{X}$
- $\mathcal{X} \subset \mathbb{R}^n$

Taxonomy (neos Guide)

Mathematical Optimization

$$\min_{x \in \mathcal{X}} f(x)$$

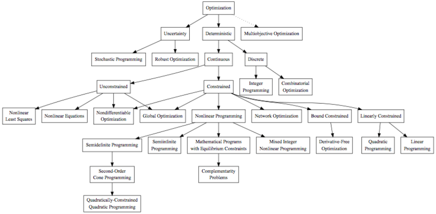
- x^* sought with $f(x^*) \leq f(x) \quad \forall x \in \mathcal{X}$
- $\mathcal{X} \subset \mathbb{R}^n$

Taxonomy (neos Guide)

Optimization Taxonomy

Back to Types of Optimization Problems

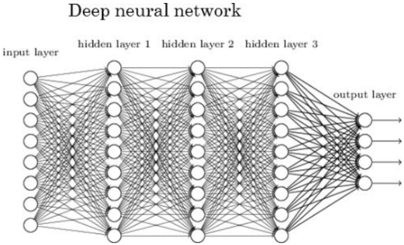
It is difficult to provide a taxonomy of optimization because many of the subfields have multiple links. Shown here is one perspective, focused mainly on the subfields of deterministic optimization with a single objective function.



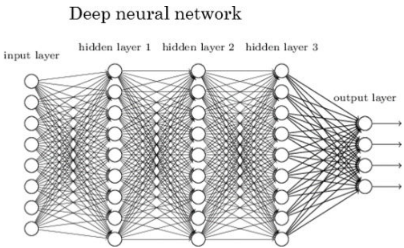
E. Carrizosa and M.D. Romero Morales
Supervised Classification and Mathematical Optimization
Computers & OR, 2013.

A Pedro Duarte Silva
Optimization Approaches to Supervised Classification
EJOR, 2017.

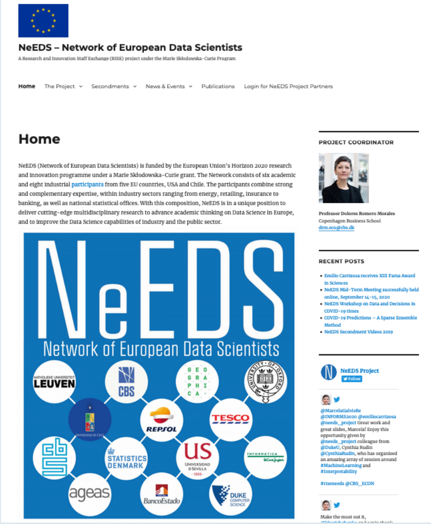
Claudio Gambella, Bissan Ghaddar Joe Naoum-Sawaya
Optimization problems for machine learning: A survey
EJOR, 2020.



https://riseneeds.eu



https://riseneeds.eu



Sparse
Principal Component Analysis

PCA. A very quick introduction

PCA

- Principal Component Analysis (PCA): way of projecting properly a data set $\subset \mathbb{R}^n$ into a vector space V of smaller dimension

K. Pearson
On Lines and Planes of Closest Fit to Systems of Points in Space
Philosophical Magazine, 1901.

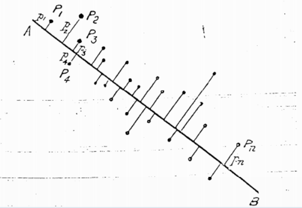
PCA. A very quick introduction

PCA

- Principal Component Analysis (PCA): way of projecting properly a data set $\subset \mathbb{R}^n$ into a vector space V of smaller dimension

K. Pearson
On Lines and Planes of Closest Fit to Systems of Points in Space
Philosophical Magazine, 1901.

(y' being the ordinate of the theoretical line at the point x which corresponds to y), had we wanted to determine the best-fitting line in the usual manner.



PCA

- Principal Component Analysis (PCA): way of projecting properly a data set $\subset \mathbb{R}^n$ into a vector space V of smaller dimension

Projections

Given $c_1, \ldots, c_k \in \mathbb{R}^n$, denote

- $span(\{c_1, \ldots, c_k\})$: vector space spanned by $c_1, \ldots, c_k$
- $\pi_{\{c_1,\ldots,c_k\}}$: projection onto $span(\{c_1, \ldots, c_k\})$:

$$\pi_{\{c_1,\ldots,c_k\}}(x) = \arg \min_{y \in span(\{c_1,\ldots,c_k\})} \|y - x\|$$

PCA. A very quick introduction

PCA

- Principal Component Analysis (PCA): way of projecting properly a data set $\subset \mathbb{R}^n$ into a vector space V of smaller dimension

Projections

Given $c_1, \ldots, c_k \in \mathbb{R}^n$, denote

- $span(\{c_1, \ldots, c_k\})$: vector space spanned by $c_1, \ldots, c_k$
- $\pi_{\{c_1,\ldots,c_k\}}$: projection onto $span(\{c_1, \ldots, c_k\})$:

$$\pi_{\{c_1,\ldots,c_k\}}(x) = \arg \min_{y \in span(\{c_1,\ldots,c_k\})} \|y - x\|$$

PCA. A very quick introduction

PCA

- Principal Component Analysis (PCA): way of projecting properly a data set $\subset \mathbb{R}^n$ into a vector space V of smaller dimension

Projections

Given $c_1, \ldots, c_k \in \mathbb{R}^n$, denote

- $span(\{c_1, \ldots, c_k\})$: vector space spanned by $c_1, \ldots, c_k$
- $\pi_{\{c_1,\ldots,c_k\}}$: projection onto $span(\{c_1, \ldots, c_k\})$:

$$\pi_{\{c_1,\ldots,c_k\}}(x) \quad = \quad \arg \min_{y \in span(\{c_1, \ldots, c_k\})} \|y - x\|$$

PCA. A very quick introduction

- We're given $\{u_1, \ldots, u_p\} \subset \mathbb{R}^n$, wlog, $\frac{1}{p}\sum_{i=1}^{p} u_i = 0_n$
- We're seeking orthonormal vectors $c_1, \ldots, c_k$ s.t.

PCA. A very quick introduction

- We're given $\{u_1, \ldots, u_p\} \subset \mathbb{R}^n$, wlog, $\frac{1}{p}\sum_{i=1}^{p} u_i = 0_n$
- We're seeking orthonormal vectors $c_1, \ldots, c_k$ s.t.
  - $u_i \approx \pi_{\{c_1,\ldots,c_k\}}(u_i) \qquad \forall i = 1, 2, \ldots, p$:

$$\min_{c_1,\ldots,c_k:\ \text{orthonormal}} \frac{1}{p}\sum_{i=1}^{p} \left\| u_i - \pi_{\{c_1,\ldots,c_k\}}(u_i) \right\|^2$$

PCA. A very quick introduction

- We're given $\{u_1, \ldots, u_p\} \subset \mathbb{R}^n$, wlog, $\frac{1}{p}\sum_{i=1}^{p} u_i = 0_n$
- We're seeking orthonormal vectors $c_1, \ldots, c_k$ s.t.
  - $u_i \approx \pi_{\{c_1,\ldots,c_k\}}(u_i) \qquad \forall i = 1, 2, \ldots, p:$

$$\min_{c_1,\ldots,c_k:\ \text{orthonormal}} \frac{1}{p}\sum_{i=1}^{p} \left\| u_i - \pi_{\{c_1,\ldots,c_k\}}(u_i) \right\|^2$$

PCA. A very quick introduction

$$\min_{c_1,\ldots,c_k:\ \text{orthonormal}} \frac{1}{p} \sum_{i=1}^{p} \left\| u_i - \pi_{\{c_1,\ldots,c_k\}}(u_i) \right\|^2$$

- $V := \frac{1}{p}(u_1|u_2|\ldots|u_p) \cdot (u_1|u_2|\ldots|u_p)^\top$ (covariance matrix), an sdp matrix
- Problem equivalent to

PCA. A very quick introduction

$$\min_{c_1,\dots,c_k:\ \text{orthonormal}} \frac{1}{p}\sum_{i=1}^{p}\left\|u_i - \pi_{\{c_1,\dots,c_k\}}(u_i)\right\|^2$$

- $V := \frac{1}{p}\,(u_1|u_2|\dots|u_p)\cdot(u_1|u_2|\dots|u_p)^\top$ (covariance matrix), an sdp matrix
- Problem equivalent to

PCA. A very quick introduction

$$\min_{c_1,\ldots,c_k:\ \text{orthonormal}} \frac{1}{p}\sum_{i=1}^{p}\left\|u_i - \pi_{\{c_1,\ldots,c_k\}}(u_i)\right\|^2$$

- $V := \frac{1}{p}\left(u_1|u_2|\ldots|u_p\right)\cdot\left(u_1|u_2|\ldots|u_p\right)^\top$ (covariance matrix), an sdp matrix
- Problem equivalent to

$$\begin{array}{ll} \min & \frac{1}{p}\sum_{i=1}^{p}\|u_i\|^2 - \frac{1}{p}\sum_{j=1}^{k} c_j^\top \cdot V \cdot c_j \\ & c_i^\top c_j = \delta_{ij} \qquad \forall i,j = 1\ldots k \end{array}$$

PCA. A very quick introduction

$$\min_{c_1,\ldots,c_k:\ \text{orthonormal}} \frac{1}{p}\sum_{i=1}^{p}\left\| u_i - \pi_{\{c_1,\ldots,c_k\}}(u_i)\right\|^2$$

- $V := \frac{1}{p}(u_1|u_2|\ldots|u_p)\cdot(u_1|u_2|\ldots|u_p)^\top$ (covariance matrix), an sdp matrix
- Problem equivalent to

$$\frac{1}{p}\sum_{i=1}^{p}\|u_i\|^2 \quad - \quad \begin{array}{l} \max \frac{1}{p}\sum_{j=1}^{k} c_j^\top \cdot V \cdot c_j \\ c_i^\top c_j = \delta_{ij} \qquad \forall i,j = 1\ldots k \end{array}$$

PCA. A very quick introduction

$$\min_{c_1,\dots,c_k:\ \text{orthonormal}} \frac{1}{p}\sum_{i=1}^{p} \left\| u_i - \pi_{\{c_1,\dots,c_k\}}(u_i) \right\|^2$$

- $V := \frac{1}{p}(u_1|u_2|\dots|u_p)\cdot(u_1|u_2|\dots|u_p)^\top$ (covariance matrix), an sdp matrix
- Problem equivalent to

$$\frac{1}{p}\sum_{i=1}^{p}\|u_i\|^2 \quad - \quad \overbrace{\max \frac{1}{p}\sum_{j=1}^{k} c_j^\top \cdot V \cdot c_j}^{\text{var explained } v_k}$$

$$c_i^\top c_j = \delta_{ij} \qquad \forall i,j = 1\dots k$$

PCA. A very quick introduction

$$\begin{array}{ll} \min & \frac{1}{p}\sum_{i=1}^{p}\|u_i\|^2 - \frac{1}{p}\sum_{j=1}^{k} c_j^\top \cdot V \cdot c_j \\ & c_i^\top c_j = \delta_{ij} \qquad \forall i,j = 1\dots k \end{array}$$

Calculating principal components

- Optimal $c_1, c_2, \dots, c_k$: unit eigenvectors associated with the k largest eigenvalues of the sdp matrix V

Hearing Loss

PCA			
.40	-.32	-.16	-.33
.42	-.23	.05	-.48
.37	.24	.47	-.28
.28	.47	-.43	-.16
.34	-.39	-.26	.49
.41	-.23	.03	.37
.31	.32	.56	.39
.25	.51	-.43	.16
Var 87.4 %			

Interpreting principal components

Hastie et al, 2009

We often interpret principal components by examining the direction vectors c_j, also known as loadings, to see which variables play a role. Often this interpretation is made easier if the loadings are sparse.

Jolliffe et al, 2003

A common approach is to effectively ignore (treat as zero) any coefficients less than some threshold value, so that the function becomes simple and the interpretation becomes easier for the users. Such a procedure can be misleading.

Interpreting principal components

Hastie et al, 2009

We often interpret principal components by examining the direction vectors c_j, also known as loadings, to see which variables play a role. Often this interpretation is made easier if the loadings are sparse.

Jolliffe et al, 2003

A common approach is to effectively ignore (treat as zero) any coefficients less than some threshold value, so that the function becomes simple and the interpretation becomes easier for the users. Such a procedure can be misleading.

Interpreting principal components

- Several attempts
  - Simple Component Analysis
  - Rotation procedures (varimax, . . .)
  - Lasso-based procedures (SCoTLASS, . . .)
  - SDP-based (DSPCA)
  - . . .

Sparse PCA. A few references

d'Aspremont, A., El Ghaoui, L., Jordan, M., and Lanckriet, G.
A Direct Formulation for Sparse PCA Using Semidefinite Programming, *SIAM Review*, 2007.

Carrizosa, E., and Guerrero, V.
rs-Sparse principal component analysis: A mixed integer nonlinear programming approach with VNS *Computers & Operations Research* 2014.

Carrizosa, E., and Guerrero, V.
Biobjective sparse principal component analysis. Journal of Multivariate Analysis 132, 2014.

Jolliffe, I.T., N. T. Trendafilov and M. Uddin
"A Modified Principal Component Technique Based on the LASSO", *J. of Computational and Graphical Statistics*, 2003.

McCabe, G. P.
"Principal Variables", *Technometrics*, 26, 1984.

Vines, S. K.
"Simple Principal Components", *Applied Statistics*, 2000.

Zou, H., T. Hastie and R. Tibshirani
"Sparse Principal Component Analysis", *J. of Computational and Graphical Statistics*, 2006.

Sparse PCA. Our proposal

PCA

$$\min \quad \frac{1}{p}\sum_{i=1}^{p} \left\| u_i - \pi_{\{c_1,\ldots,c_k\}}(u_i) \right\|^2$$
$$c_1, \ldots, c_k : \text{ orthonormal}$$

Global sparsity constraints

1. Each variable is nonzero in at most r components c_j
2. Each c_j has at most s nonzero elements

Hard constraints

Sparse PCA. Our proposal

Sparse PCA

$$\min \quad \frac{1}{p}\sum_{i=1}^{p}\left\|u_i - \pi_{\{c_1,\ldots,c_k\}}(u_i)\right\|^2$$

$c_1, \ldots, c_k$: orthonormal

$+$ global sparsity constraints:

Global sparsity constraints

1. Each variable is nonzero in at most r components c_j
2. Each c_j has at most s nonzero elements

Hard constraints

Sparse PCA. Our proposal

Sparse PCA

$$\min \quad \frac{1}{p}\sum_{i=1}^{p} \left\| u_i - \pi_{\{c_1,\ldots,c_k\}}(u_i) \right\|^2$$

$c_1, \ldots, c_k$: orthonormal

$+$ global sparsity constraints:

Global sparsity constraints

1. Each variable is nonzero in at most r components c_j
2. Each c_j has at most s nonzero elements

Hard constraints

Sparse PCA. Our proposal

Sparse PCA

$$\min \quad \frac{1}{p}\sum_{i=1}^{p}\left\|u_i - \pi_{\{c_1,\ldots,c_k\}}(u_i)\right\|^2$$

$c_1, \ldots, c_k$: orthonormal

\+ global sparsity constraints:

Global sparsity constraints

1. Each variable is nonzero in at most r components c_j
2. Each c_j has at most s nonzero elements

Hard constraints

Define: $z_{il} = \begin{cases} 1 & \text{if} \quad c_{il} \neq 0 \\ 0 & \text{else} \end{cases} \quad i = 1 \ldots k, l = 1 \ldots n$

$$\sum_{i=1}^{k} z_{il} \leq r \quad \forall l = 1 \ldots n$$

$$\sum_{l=1}^{n} z_{il} \leq s \quad \forall i = 1 \ldots k$$

Sparse PCA. MINLP formulation

Define: $z_{il} = \begin{cases} 1 & \text{if} \quad c_{il} \neq 0 \\ 0 & \text{else} \end{cases} \quad i = 1 \dots k, l = 1 \dots n$

$$|c_{il}| \leq M z_{il} \qquad \forall i, l$$

$$\sum_{i=1}^{k} z_{il} \leq r \qquad \forall l = 1 \dots n$$

$$\sum_{l=1}^{n} z_{il} \leq s \qquad \forall i = 1 \dots k$$

Sparse PCA. MINLP formulation

Define: $z_{il} = \begin{cases} 1 & \text{if} \quad c_{il} \neq 0 \\ 0 & \text{else} \end{cases} \quad i = 1 \dots k, l = 1 \dots n$

$$|c_{il}| \leq 1 z_{il} \qquad \forall i, l$$

$$\sum_{i=1}^{k} z_{il} \leq r \quad \forall l = 1 \dots n$$

$$\sum_{l=1}^{n} z_{il} \leq s \quad \forall i = 1 \dots k$$

Sparse PCA. MINLP formulation

Define: $z_{il} = \begin{cases} 1 & \text{if} \quad c_{il} \neq 0 \\ 0 & \text{else} \end{cases} \quad i = 1 \ldots k, l = 1 \ldots n$

$$|c_{il}| \leq 1 z_{il} \qquad \forall i, l$$

$$\sum_{i=1}^{k} z_{il} \leq r \quad \forall l = 1 \ldots n$$

$$\sum_{l=1}^{n} z_{il} \leq s \quad \forall i = 1 \ldots k$$

Sparse PCA. MINLP formulation

Define: $z_{il} = \begin{cases} 1 & \text{if} \quad c_{il} \neq 0 \\ 0 & \text{else} \end{cases} \quad i = 1 \dots k, l = 1 \dots n$

$$|c_{il}| \leq 1 z_{il} \qquad \forall i, l$$

$$\sum_{i=1}^{k} z_{il} \leq r \quad \forall l = 1 \dots n$$

$$\sum_{l=1}^{n} z_{il} \leq s \quad \forall i = 1 \dots k$$

MINLP formulation

$$
\begin{array}{lll}
\min & \frac{1}{p}\sum_{i=1}^{p} \left\| u_i - \pi_{\{c_1,\ldots,c_k\}}(u_i) \right\|^2 & \\
 & c_i^\top c_j = \delta_{ij} & \forall i,j \\
 & |c_{il}| \leq z_{il} & \forall i,l \\
 & \sum_{i=1}^{k} z_{il} \leq r & \forall l = 1\ldots n \\
 & \sum_{l=1}^{n} z_{il} \leq s & \forall i = 1\ldots k \\
 & z_{il} \in \{0,1\} & \forall i,l
\end{array}
$$

MINLP formulation

$$
\begin{array}{lll}
\max & \sum_{j=1}^{k} c_j^\top \cdot V \cdot c_j & \\
& c_i^\top c_j = \delta_{ij} & \forall i, j \\
& |c_{il}| \leq z_{il} & \forall i, l \\
& \sum_{i=1}^{k} z_{il} \leq r & \forall l = 1 \ldots n \\
& \sum_{l=1}^{n} z_{il} \leq s & \forall i = 1 \ldots k \\
& z_{il} \in \{0, 1\} & \forall i, l
\end{array}
$$

MINLP formulation

$$
\begin{array}{lll}
\max & \sum_{j=1}^{k} c_j^\top \cdot V \cdot c_j & \\
 & c_i^\top c_j = \delta_{ij} & \forall i, j \\
 & |c_{il}| \leq z_{il} & \forall i, l \\
 & \sum_{i=1}^{k} z_{il} \leq r & \forall l = 1 \ldots n \\
 & \sum_{l=1}^{n} z_{il} \leq s & \forall i = 1 \ldots k \\
 & z_{il} \in \{0, 1\} & \forall i, l
\end{array}
$$

Simmetry issues?

$$
\begin{array}{lll}
\max & \sum_{j=1}^{k} c_j^\top \cdot V \cdot c_j & \\
 & c_i^\top c_j = \delta_{ij} & \forall i, j \\
 & |c_{il}| \leq z_{il} & \forall i, l \\
 & \sum_{i=1}^{k} z_{il} \leq r & \forall l = 1 \ldots n \\
 & \sum_{l=1}^{n} z_{il} \leq s & \forall i = 1 \ldots k \\
 & z_{il} \in \{0, 1\} & s\forall i, l
\end{array}
$$

Simmetry issues?

$$
\begin{array}{lll}
\max & \sum_{j=1}^{k} c_j^\top \cdot V \cdot c_j & \\
& c_i^\top c_j = \delta_{ij} & \forall i, j \\
& |c_{il}| \leq z_{il} & \forall i, l \\
& \sum_{i=1}^{k} z_{il} \leq r & \forall l = 1 \dots n \\
& \sum_{l=1}^{n} z_{il} \leq s & \forall i = 1 \dots k \\
& z_{il} \in \{0, 1\} & \forall i, l \\
& u^\top c_i \leq u^\top c_{i+1} & \forall i = 1 \dots k-1
\end{array}
$$

$r = 1$

$$
\begin{array}{lll}
\max & \sum_{j=1}^{k} c_j^\top \cdot V \cdot c_j & \\
 & c_i^\top c_j = \delta_{ij} & \forall i, j \\
 & |c_{il}| \leq z_{il} & \forall i, l \\
 & \sum_{i=1}^{k} z_{il} = 1 & \forall l = 1 \dots n \\
 & \sum_{l=1}^{n} z_{il} = s & \forall i = 1 \dots k \\
 & z_{il} \in \{0, 1\} & s \forall i, l
\end{array}
$$

$r = 1$

$$
\begin{array}{lll}
\max & \sum_{j=1}^{k} c_j^\top \cdot V \cdot c_j & \\
& c_i^\top c_i = 1 & \forall i \\
& |c_{il}| \leq z_{il} & \forall i, l \\
& \sum_{i=1}^{k} z_{il} = 1 & \forall l = 1 \dots n \\
& \sum_{l=1}^{n} z_{il} = s & \forall i = 1 \dots k \\
& z_{il} \in \{0, 1\} & s \forall i, l
\end{array}
$$

Fixing z ...

$$
\begin{array}{lll}
\max & \sum_{j=1}^{k} c_j^\top \cdot V \cdot c_j & \\
& c_i^\top c_i = 1 & \forall i \\
& |c_{il}| \leq z_{il} & \forall i, l \\
& \sum_{i=1}^{k} z_{il} = 1 & \forall l = 1 \dots n \\
& \sum_{l=1}^{n} z_{il} = s & \forall i = 1 \dots k \\
& z_{il} \in \{0, 1\} & \forall i, l
\end{array}
$$

Resulting problem ...

- Separable in k problems (of classical PCA-type)
- Amounts to solving largest eigenvalue and associated eigenvector of k submatrices of V

Fixing z ...

$$
\begin{array}{lll}
\max & \sum_{j=1}^{k} c_j^\top \cdot V \cdot c_j & \\
 & c_i^\top c_i = 1 & \forall i \\
 & c_{il} = 0 & \forall i, l : z_{il} = 0
\end{array}
$$

Resulting problem ...

- Separable in k problems (of classical PCA-type)
- Amounts to solving largest eigenvalue and associated eigenvector of k submatrices of V

Fixing z ...

$$\begin{array}{lll} \max & \sum_{j=1}^{k} c_j^\top \cdot V \cdot c_j & \\ & c_i^\top c_i = 1 & \forall i \\ & c_{il} = 0 & \forall i, l : z_{il} = 0 \end{array}$$

Resulting problem ...

- Separable in k problems (of classical PCA-type)
- Amounts to solving largest eigenvalue and associated eigenvector of k submatrices of V

Fixing z ...

$$\begin{array}{lll} \max & \sum_{j=1}^{k} c_j^\top \cdot V \cdot c_j & \\ & c_i^\top c_i = 1 & \forall i \\ & c_{il} = 0 & \forall i, l : z_{il} = 0 \end{array}$$

Resulting problem ...

- Separable in k problems (of classical PCA-type)
- Amounts to solving largest eigenvalue and associated eigenvector of k submatrices of V

A first heuristic

1. "Judiciously" choose z
2. Find the optimal c of z fixed (by calculating k eigenvalues and eigenvectors)

Choosing z

- Easily available: $c_1^*, \ldots, c_k^*$, principal components
- Controlled rounding of $c_1^*, \ldots, c_k^*$:

$$\begin{array}{lll} \max & \sum_{i=1}^{n} \sum_{l=1}^{k} |c_{il}^*| z_{il} & \\ & \sum_{l=1}^{k} z_{il} = 1 & \forall i = 1, \ldots, n \\ & \sum_{i=1}^{n} z_{il} \leq s & \forall l = 1, \ldots, k \\ & \sum_{i=1}^{n} z_{il} \geq 1 & \forall l = 1, \ldots, k \\ & z_{il} \geq 0 & \forall i, l \end{array}$$

A first heuristic

1. "Judiciously" choose z
2. Find the optimal c of z fixed (by calculating k eigenvalues and eigenvectors)

Choosing z

- Easily available: $c_1^*, \ldots, c_k^*$, principal components
- Controlled rounding of $c_1^*, \ldots, c_k^*$:

$$
\begin{array}{lll}
\max & \sum_{i=1}^{n} \sum_{l=1}^{k} |c_{il}^*| z_{il} & \\
 & \sum_{l=1}^{k} z_{il} = 1 & \forall i = 1, \ldots, n \\
 & \sum_{i=1}^{n} z_{il} \leq s & \forall l = 1, \ldots, k \\
 & \sum_{i=1}^{n} z_{il} \geq 1 & \forall l = 1, \ldots, k \\
 & z_{il} \geq 0 & \forall i, l
\end{array}
$$

A first heuristic

1. "Judiciously" choose z
2. Find the optimal c of z fixed (by calculating k eigenvalues and eigenvectors)

Choosing z

- Easily available: $c_1^*, \ldots, c_k^*$, principal components
- Controlled rounding of $c_1^*, \ldots, c_k^*$:

$$
\begin{array}{lll}
\max & \sum_{i=1}^{n} \sum_{l=1}^{k} |c_{il}^*| z_{il} & \\
 & \sum_{l=1}^{k} z_{il} = 1 & \forall i = 1, \ldots, n \\
 & \sum_{i=1}^{n} z_{il} \leq s & \forall l = 1, \ldots, k \\
 & \sum_{i=1}^{n} z_{il} \geq 1 & \forall l = 1, \ldots, k \\
 & z_{il} \geq 0 & \forall i, l
\end{array}
$$

A metaheuristic

- Solution (c, z) so obtained:
  - feasible (sparse + orthonormal)
  - may not be optimal to the MINLP
  - starting point of a search procedure (exchange algorithm, VNS algorithm, with natural definition of neighborhoods)

For arbitrary r

For arbitrary r

$$
\begin{array}{lll}
\max & \sum_{j=1}^{k} c_j^\top \cdot V \cdot c_j & \\
 & c_i^\top c_j = \delta_{ij} & \forall i, j \\
 & |c_{il}| \leq z_{il} & \forall i, l \\
 & \sum_{i=1}^{k} z_{il} \leq r & \forall l = 1 \ldots n \\
 & \sum_{l=1}^{n} z_{il} \leq s & \forall i = 1 \ldots k \\
 & z_{il} \in \{0, 1\} & s\forall i, l
\end{array}
$$

1. "Judiciously" choose z (controlled rounding: flow problem)
2. For z fixed, solve the NLP problem
3. Solution so obtained
    - hopefully feasible
    - starting point of a search procedure (exchange algorithm, VNS algorithm, with natural definition of neighborhoods)

Data sets

Data from Rousson-Gasser, 2003

Name	n	k
Hearing Loss	8	4
Reflexes	10	5
Pitprop	13	6
Movements	22	4
Musclestrength	51	6

Benchmarks

- PCA
- Varimax. (Kaiser, 1958)
- SCA (Rousson-Gasser, 2003)
- SPCA (Zou-Hastie-Tibshirani, 2006)

Data sets

Data from Rousson-Gasser, 2003

Name	n	k
`Hearing Loss`	8	4
`Reflexes`	10	5
`Pitprop`	13	6
`Movements`	22	4
`Musclestrength`	51	6

Benchmarks

- PCA
- Varimax. (Kaiser, 1958)
- SCA (Rousson-Gasser, 2003)
- SPCA (Zou-Hastie-Tibshirani, 2006)

Hearing Loss

PCA			
.40	-.32	-.16	-.33
.42	-.23	.05	-.48
.37	.24	.47	-.28
.28	.47	-.43	-.16
.34	-.39	-.26	.49
.41	-.23	.03	.37
.31	.32	.56	.39
.25	.51	-.43	.16
Var 87.4 %			

Varimax			
.60	-.09	.03	.15
.67	.11	-.03	-.03
.29	.61	.02	-.19
.13	-.01	.70	-.10
.03	-.16	.02	.74
.07	.15	-.02	.58
-.26	.75	-.01	.21
-.13	.02	.71	.09
Var 68.4 %			

SCA			
.35	-.35	.00	-.35
.35	-.35	.00	-.35
.35	.35	-.50	-.35
.35	.35	.50	-.35
.35	-.35	.00	.35
.35	-.35	.00	.35
.35	.35	-.50	.35
.35	.35	.50	.35
Var 85.4 %			

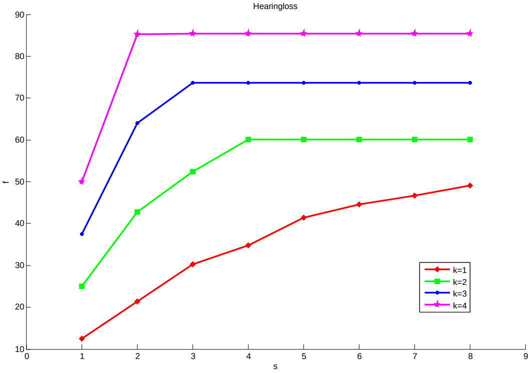
SPCA			
-.58	.00	.00	.00
-.71	.00	.00	.00
.00	.00	.61	.00
.00	-.71	.00	.00
.00	.00	.00	-1.00
-.40	.00	.00	.00
.00	.00	.79	.00
.00	-.71	.00	.00
Var 84.08 %			

Algorithms tested

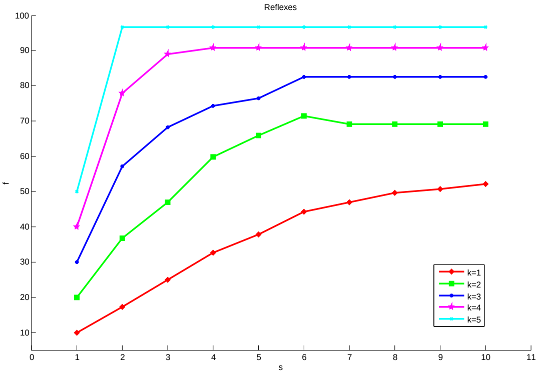
Case $r = 1$

1. Random allocation heuristic (`rnd`)
2. Transportation heuristic (`transp`)
3. Exchange
4. VNS

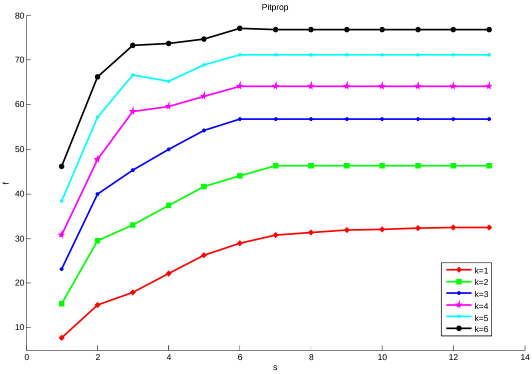
s-SPCA: f varying s and k for `Hearingloss`



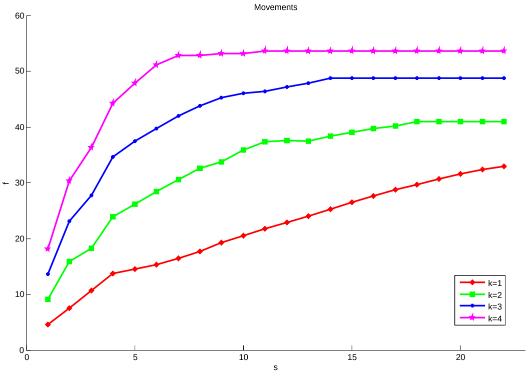
s-SPCA: f varying s and k for `Reflexes`



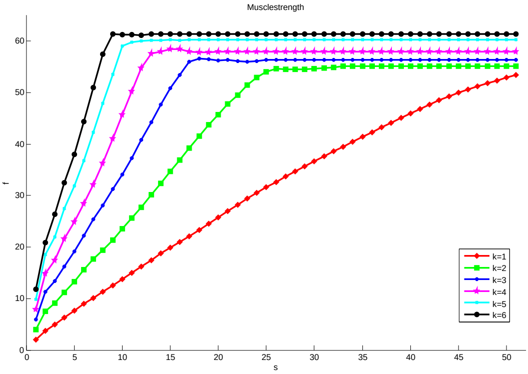
s-SPCA: f varying s and k for Pitprops



s-SPCA: f varying s and k for `Movements`



s-SPCA: f varying s and k for Musclestrength



s-SPCA against classical approaches

Data set	Sparsity	PCA	VARIMAX	SCA	SPCA	s-SPCA
Hearingloss	%zeros	0	0	12.50	75.00	75.00
	f	87.37	68.40	85.40	84.08	85.37
Reflexes	%zeros	0	16.00	52.00	80.00	80.00
	f	97.05	72.20	91.50	96.17	96.70
Pitprops	%zeros	1.28	7.69	80.77	83.33	83.33
	f	87.00	78.90	74.80	71.99	76.85
Movements	%zeros	0	2.27	20.45	75.00	75.00
	f	55.00	43.20	53.80	49.84	53.60
Musclestregth	%zeros	1.63	8.50	34.64	80.07	80.07
	f	70.40	70.39	68.10	60.00	61.39

s-SPCA against classical approaches

Data set	Sparsity	PCA	VARIMAX	SCA	SPCA	s-SPCA
Hearingloss	%zeros	0	0	12.50	75.00	75.00
	f	87.37	68.40	85.40	84.08	85.37
Reflexes	%zeros	0	16.00	52.00	80.00	80.00
	f	97.05	72.20	91.50	96.17	96.70
Pitprops	%zeros	1.28	7.69	80.77	83.33	83.33
	f	87.00	78.90	74.80	71.99	76.85
Movements	%zeros	0	2.27	20.45	75.00	75.00
	f	55.00	43.20	53.80	49.84	53.60
Musclestregth	%zeros	1.63	8.50	34.64	80.07	80.07
	f	70.40	70.39	68.10	60.00	61.39

s-SPCA against classical approaches

Data set	Sparsity	PCA	VARIMAX	SCA	SPCA	s-SPCA
Hearingloss	%zeros	0	0	12.50	75.00	75.00
	f	87.37	68.40	85.40	84.08	85.37
Reflexes	%zeros	0	16.00	52.00	80.00	80.00
	f	97.05	72.20	91.50	96.17	96.70
Pitprops	%zeros	1.28	7.69	80.77	83.33	83.33
	f	87.00	78.90	74.80	71.99	76.85
Movements	%zeros	0	2.27	20.45	75.00	75.00
	f	55.00	43.20	53.80	49.84	53.60
Musclestregth	%zeros	1.63	8.50	34.64	80.07	80.07
	f	70.40	70.39	68.10	60.00	61.39

s-SPCA against classical approaches

Data set	Sparsity	PCA	VARIMAX	SCA	SPCA	s-SPCA
`Hearingloss`	%zeros	0	0	12.50	75.00	75.00
	f	87.37	68.40	85.40	84.08	85.37
`Reflexes`	%zeros	0	16.00	52.00	80.00	80.00
	f	97.05	72.20	91.50	96.17	96.70
`Pitprops`	%zeros	1.28	7.69	80.77	83.33	83.33
	f	87.00	78.90	74.80	71.99	76.85
`Movements`	%zeros	0	2.27	20.45	75.00	75.00
	f	55.00	43.20	53.80	49.84	53.60
`Musclestregth`	%zeros	1.63	8.50	34.64	80.07	80.07
	f	70.40	70.39	68.10	60.00	61.39

s-SPCA against classical approaches

Data set	Sparsity	PCA	VARIMAX	SCA	SPCA	s-SPCA
`Hearingloss`	%zeros	0	0	12.50	75.00	75.00
	f	87.37	68.40	85.40	84.08	85.37
`Reflexes`	%zeros	0	16.00	52.00	80.00	80.00
	f	97.05	72.20	91.50	96.17	96.70
`Pitprops`	%zeros	1.28	7.69	80.77	83.33	83.33
	f	87.00	78.90	74.80	71.99	76.85
`Movements`	%zeros	0	2.27	20.45	75.00	75.00
	f	55.00	43.20	53.80	49.84	53.60
`Musclestregth`	%zeros	1.63	8.50	34.64	80.07	80.07
	f	70.40	70.39	68.10	60.00	61.39

s-SPCA against classical approaches

Data set	Sparsity	PCA	VARIMAX	SCA	SPCA	s-SPCA
Hearingloss	%zeros	0	0	12.50	75.00	75.00
	f	87.37	68.40	85.40	84.08	85.37
Reflexes	%zeros	0	16.00	52.00	80.00	80.00
	f	97.05	72.20	91.50	96.17	96.70
Pitprops	%zeros	1.28	7.69	80.77	83.33	83.33
	f	87.00	78.90	74.80	71.99	76.85
Movements	%zeros	0	2.27	20.45	75.00	75.00
	f	55.00	43.20	53.80	49.84	53.60
Musclestregth	%zeros	1.63	8.50	34.64	80.07	80.07
	f	70.40	70.39	68.10	60.00	61.39

s-SPCA against classical approaches

Data set	Sparsity	PCA	VARIMAX	SCA	SPCA	s-SPCA
Hearingloss	%zeros	0	0	12.50	75.00	75.00
	f	87.37	68.40	85.40	84.08	85.37
Reflexes	%zeros	0	16.00	52.00	80.00	80.00
	f	97.05	72.20	91.50	96.17	96.70
Pitprops	%zeros	1.28	7.69	80.77	83.33	83.33
	f	87.00	78.90	74.80	71.99	76.85
Movements	%zeros	0	2.27	20.45	75.00	75.00
	f	55.00	43.20	53.80	49.84	53.60
Musclestregth	%zeros	1.63	8.50	34.64	80.07	80.07
	f	70.40	70.39	68.10	60.00	61.39

Algorithms tested

Case $r \geq 1$

1. Transportation heuristic (`transp`)
2. Plain NLP:
   1. Find c^* : principal components
   2. Find z^*, transportation solution from c^*
   3. Write the NLP with $z_{ij} \in \{0, 1\}$ as $z_{ij}(1 - z_{ij}) = 0$
   4. Solve the NLP with (c^*, z^*) as starting point
3. Exchange algorithm, starting from (c^*, z^*)

Algorithms tested

Case $r \geq 1$

1. Transportation heuristic (`transp`)
2. Plain NLP:
    1. Find c^* : principal components
    2. Find z^*, transportation solution from c^*
    3. Write the NLP with $z_{il} \in \{0, 1\}$ as $z_{il}(1 - z_{il}) = 0$
    4. Solve the NLP with (c^*, z^*) as starting point
3. Exchange algorithm, starting from (c^*, z^*)

Algorithms tested

Case $r \geq 1$

1. Transportation heuristic (`transp`)
2. Plain NLP:
   1. Find c^* : principal components
   2. Find z^*, transportation solution from c^*
   3. Write the NLP with $z_{il} \in \{0, 1\}$ as $z_{il}(1 - z_{il}) = 0$
   4. Solve the NLP with (c^*, z^*) as starting point
3. Exchange algorithm, starting from (c^*, z^*)

Hearing loss

s/r	1	2	3	4
1	50.00	50.00	50.00	50.00
2	75.66	67.64	67.64	67.64
3	75.66	75.66	75.66	75.66
4	75.66	86.01	86.02	86.02
5	75.66	86.10	86.12	86.81
6	75.66	86.10	87.20	87.19
7	75.66	86.10	86.91	87.37
8	75.66	86.10	86.91	87.37

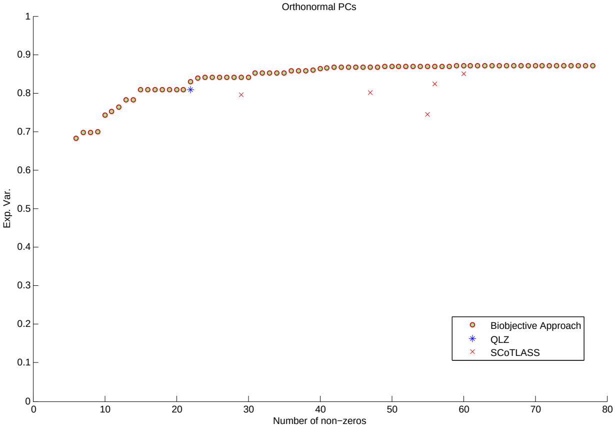
Reflexes

s/r	1	2	3	4	5
1	50.00	50.00	50.00	50.00	50.00
2	96.70	78.10	78.10	78.10	78.10
3	91.22	93.02	93.02	93.02	93.02
4	89.35	96.80	96.84	96.84	96.84
5	89.35	96.80	96.80	96.84	96.84
6	89.35	96.90	97.02	96.92	96.91
7	89.35	96.90	96.94	97.05	97.02
8	89.35	96.90	96.94	97.05	97.05
9	89.35	96.90	96.97	97.04	97.05
10	89.35	96.90	96.99	97.03	97.05

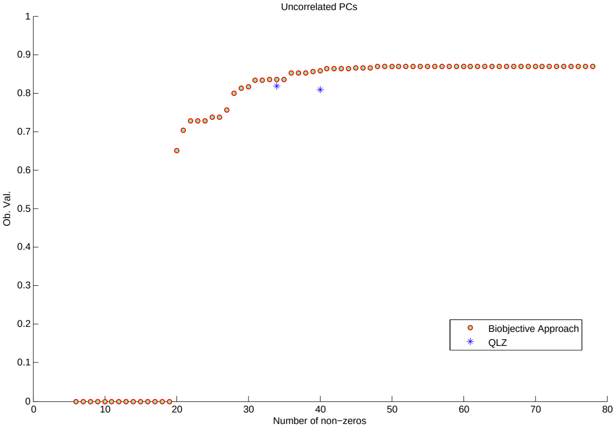
Pitprop

s/r	1	2	3	4	5	6
1	46.15	46.15	46.15	46.15	46.15	46.15
2	57.60	60.28	60.28	60.28	60.28	60.28
3	71.05	56.38	-	-	-	-
4	67.75	75.30	70.97	70.96	70.96	70.96
5	71.92	76.19	76.89	79.96	79.96	79.96
6	74.25	78.76	81.86	82.09	82.09	82.09
7	74.25	78.76	82.30	84.39	84.39	84.39
8	74.25	78.76	83.00	85.46	85.46	85.46
9	74.25	78.76	84.31	86.65	86.50	86.65
10	74.25	78.76	84.31	86.27	86.98	86.98
11	74.25	78.76	84.31	86.15	87.00	87.00
12	74.25	78.76	84.31	86.15	87.00	87.00
13	74.25	78.76	84.31	86.15	87.00	87.00

Biobjective Approach



Biobjective Approach



Interpretable Factor Analysis

Emilio Carrizosa, Vanesa Guerrero, Dolores Romero Morales & Albert Satorra
Enhancing Interpretability in Factor Analysis by Means of Mathematical Optimization
Multivariate Behavioral Research, 2020.

The model

$$y = \Lambda f + \varepsilon$$

- $y \in \mathbb{R}^p$: observed
- $f \in \mathbb{R}^r$ $(r << p)$: factors
- Λ : loading matrix
- ε : error term, $cov(f, \varepsilon) = 0$

Emilio Carrizosa, Vanesa Guerrero, Dolores Romero Morales & Albert Satorra
Enhancing Interpretability in Factor Analysis by Means of Mathematical Optimization
Multivariate Behavioral Research, 2020.

The model

$$y = \Lambda f + \varepsilon$$

- $y \in \mathbb{R}^p$: observed
- $f \in \mathbb{R}^r$ $(r << p)$: factors
- Λ : loading matrix
- ε : error term, $cov(f, \varepsilon) = 0$

$$y = \Lambda f + \varepsilon$$

Rotational invariance

- For an orthogonal matrix M (i.e., $MM^\top = M^\top M = I$),

$$\Lambda f = \Lambda M^\top M f$$

- Aim: find some orthogonal M such that $\Lambda M^\top$ is sparse (e.g. by maximizing the variance of $\Lambda M^\top$, or its ℓ_4 norm)
- Doing this, it is expected that factors are easier to interpret, since they are linked to a few original features
- We can always assume we have explanatory variables, which can assist interpretation (take, for instance, the y)
- Even more, we assume given groups $C_1, \ldots, C_q$ of explanatory variables
- Define h_{ij}

$$h_{ij} = \begin{cases} 1, & \text{if cluster } C_i \text{ is matched with factor } j \\ 0, & \text{else} \end{cases}$$

- Goodness of fit: $S(M, H) = \min\{R^2_{ij}(M) : h_{ij} = 1\}$
- $R^2_{ij}(M)$: coefficient of determination if rotation M is used

$$y = \Lambda f + \varepsilon$$

Rotational invariance

- For an orthogonal matrix M (i.e., $MM^\top = M^\top M = I$),

$$\Lambda f = \Lambda M^\top M f$$

- Aim: find some orthogonal M such that $\Lambda M^\top$ is sparse (e.g. by maximizing the variance of $\Lambda M^\top$, or its ℓ_4 norm)
- Doing this, it is expected that factors are easier to interpret, since they are linked to a few original features
- We can always assume we have explanatory variables, which can assist interpretation (take, for instance, the y)
- Even more, we assume given groups $C_1, \ldots, C_q$ of explanatory variables
- Define h_{ij}

$$h_{ij} = \begin{cases} 1, & \text{if cluster } C_i \text{ is matched with factor } j \\ 0, & \text{else} \end{cases}$$

- Goodness of fit: $S(M, H) = \min\{R^2_{ij}(M) : h_{ij} = 1\}$
- $R^2_{ij}(M)$: coefficient of determination if rotation M is used

$$y = \Lambda f + \varepsilon$$

Rotational invariance

- For an orthogonal matrix M (i.e., $MM^\top = M^\top M = I$),

$$\Lambda f = \Lambda M^\top M f$$

- Aim: find some orthogonal M such that $\Lambda M^\top$ is sparse (e.g. by maximizing the variance of $\Lambda M^\top$, or its ℓ_4 norm)
- Doing this, it is expected that factors are easier to interpret, since they are linked to a few original features
- We can always assume we have explanatory variables, which can assist interpretation (take, for instance, the y)
- Even more, we assume given groups $C_1, \ldots, C_q$ of explanatory variables
- Define h_{ij}

$$h_{ij} = \begin{cases} 1, & \text{if cluster } C_i \text{ is matched with factor } j \\ 0, & \text{else} \end{cases}$$

- Goodness of fit: $S(M, H) = \min\{R^2_{ij}(M) : h_{ij} = 1\}$
- $R^2_{ij}(M)$: coefficient of determination if rotation M is used

$$y = \Lambda f + \varepsilon$$

Rotational invariance

- For an orthogonal matrix M (i.e., $MM^\top = M^\top M = I$),

$$\Lambda f = \Lambda M^\top M f$$

- Aim: find some orthogonal M such that $\Lambda M^\top$ is sparse (e.g. by maximizing the variance of $\Lambda M^\top$, or its ℓ_4 norm)
- Doing this, it is expected that factors are easier to interpret, since they are linked to a few original features
- We can always assume we have explanatory variables, which can assist interpretation (take, for instance, the y)
- Even more, we assume given groups $C_1, \ldots, C_q$ of explanatory variables
- Define h_{ij}

$$h_{ij} = \begin{cases} 1, & \text{if cluster } C_i \text{ is matched with factor } j \\ 0, & \text{else} \end{cases}$$

- Goodness of fit: $S(M, H) = \min\{R^2_{ij}(M) : h_{ij} = 1\}$
- $R^2_{ij}(M)$: coefficient of determination if rotation M is used

$$y = \Lambda f + \varepsilon$$

Rotational invariance

- For an orthogonal matrix M (i.e., $MM^\top = M^\top M = I$),

$$\Lambda f = \Lambda M^\top M f$$

- Aim: find some orthogonal M such that $\Lambda M^\top$ is sparse (e.g. by maximizing the variance of $\Lambda M^\top$, or its ℓ_4 norm)
- Doing this, it is expected that factors are easier to interpret, since they are linked to a few original features
- We can always assume we have explanatory variables, which can assist interpretation (take, for instance, the y)
- Even more, we assume given groups $C_1, \ldots, C_q$ of explanatory variables
- Define h_{ij}

$$h_{ij} = \begin{cases} 1, & \text{if cluster } C_i \text{ is matched with factor } j \\ 0, & \text{else} \end{cases}$$

- Goodness of fit: $S(M, H) = \min\{R^2_{ij}(M) : h_{ij} = 1\}$
- $R^2_{ij}(M)$: coefficient of determination if rotation M is used

$$y = \Lambda f + \varepsilon$$

Rotational invariance

- For an orthogonal matrix M (i.e., $MM^\top = M^\top M = I$),

$$\Lambda f = \Lambda M^\top M f$$

- Aim: find some orthogonal M such that $\Lambda M^\top$ is sparse (e.g. by maximizing the variance of $\Lambda M^\top$, or its ℓ_4 norm)
- Doing this, it is expected that factors are easier to interpret, since they are linked to a few original features
- We can always assume we have explanatory variables, which can assist interpretation (take, for instance, the y)
- Even more, we assume given groups $C_1, \ldots, C_q$ of explanatory variables
- Define h_{ij}

$$h_{ij} = \begin{cases} 1, & \text{if cluster } C_i \text{ is matched with factor } j \\ 0, & \text{else} \end{cases}$$

- Goodness of fit: $S(M, H) = \min\{R^2_{ij}(M) : h_{ij} = 1\}$
- $R^2_{ij}(M)$: coefficient of determination if rotation M is used

$$y = \Lambda f + \varepsilon$$

Rotational invariance

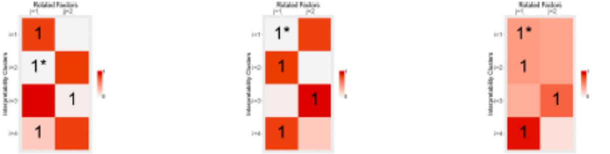
- For an orthogonal matrix M (i.e., $MM^\top = M^\top M = I$),

$$\Lambda f = \Lambda M^\top M f$$

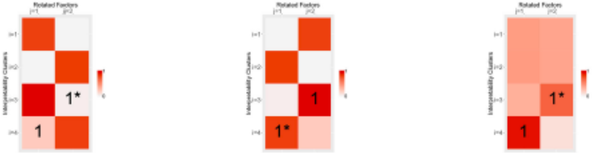
- Aim: find some orthogonal M such that $\Lambda M^\top$ is sparse (e.g. by maximizing the variance of $\Lambda M^\top$, or its ℓ_4 norm)
- Doing this, it is expected that factors are easier to interpret, since they are linked to a few original features
- We can always assume we have explanatory variables, which can assist interpretation (take, for instance, the y)
- Even more, we assume given groups $C_1, \ldots, C_q$ of explanatory variables
- Define h_{ij}

$$h_{ij} = \begin{cases} 1, & \text{if cluster } C_i \text{ is matched with factor } j \\ 0, & \text{else} \end{cases}$$

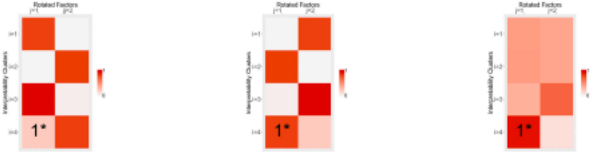
- Goodness of fit: $S(M, H) = \min\{R^2_{ij}(M) : h_{ij} = 1\}$
- $R^2_{ij}(M)$: coefficient of determination if rotation M is used



(a) $\mathcal{S}(\boldsymbol{M}_1, \boldsymbol{H}_1) = 0.00$ (b) $\mathcal{S}(\boldsymbol{M}_2, \boldsymbol{H}_1) = 0.00$ (c) $\mathcal{S}(\boldsymbol{M}_3, \boldsymbol{H}_1) = 0.41$



(d) $\mathcal{S}(\boldsymbol{M}_1, \boldsymbol{H}_2) = 0.03$ (e) $\mathcal{S}(\boldsymbol{M}_2, \boldsymbol{H}_2) = 0.80$ (f) $\mathcal{S}(\boldsymbol{M}_3, \boldsymbol{H}_2) = 0.67$



(g) $\mathcal{S}(\boldsymbol{M}_1, \boldsymbol{H}_3) = 0.20$ (h) $\mathcal{S}(\boldsymbol{M}_2, \boldsymbol{H}_3) = 0.80$ (i) $\mathcal{S}(\boldsymbol{M}_3, \boldsymbol{H}_4) = 0.90$

$$S(M,H) = \min\{R_{ij}^2(M) : \ h_{ij} = 1\}$$

$$\begin{aligned} \max \quad & S(M,H) \\ \text{s.t.} \quad & M^\top M = I \\ & H \in \mathcal{H} \end{aligned}$$

$$\begin{aligned} \max \quad & z \\ \text{s.t.} \quad & z \leq R_{ij}^2(M) h_{ij} + (1 - h_{ij}) \qquad \forall i,j \\ & M^\top M = I \\ & H \in \mathcal{H} \end{aligned}$$

- Highly nonconvex (in M) and linear integer (in H)
- Easily addressed via an alternating approach

$$S(M,H) = \min\{R_{ij}^2(M) : \; h_{ij} = 1\}$$

$$\begin{array}{ll} \max & S(M,H) \\ \text{s.t.} & M^\top M = I \\ & H \in \mathcal{H} \end{array}$$

$$\begin{array}{lll} \max & z & \\ \text{s.t.} & z \leq R_{ij}^2(M) h_{ij} + (1 - h_{ij}) & \forall i,j \\ & M^\top M = I & \\ & H \in \mathcal{H} & \end{array}$$

- Highly nonconvex (in M) and linear integer (in H)
- Easily addressed via an alternating approach

$$S(M,H) = \min\{R_{ij}^2(M) : \; h_{ij} = 1\}$$

$$\begin{array}{ll} \max & S(M,H) \\ \text{s.t.} & M^\top M = I \\ & H \in \mathcal{H} \end{array}$$

$$\begin{array}{lll} \max & z & \\ \text{s.t.} & z \leq R_{ij}^2(M) h_{ij} + (1 - h_{ij}) & \forall i,j \\ & M^\top M = I & \\ & H \in \mathcal{H} & \end{array}$$

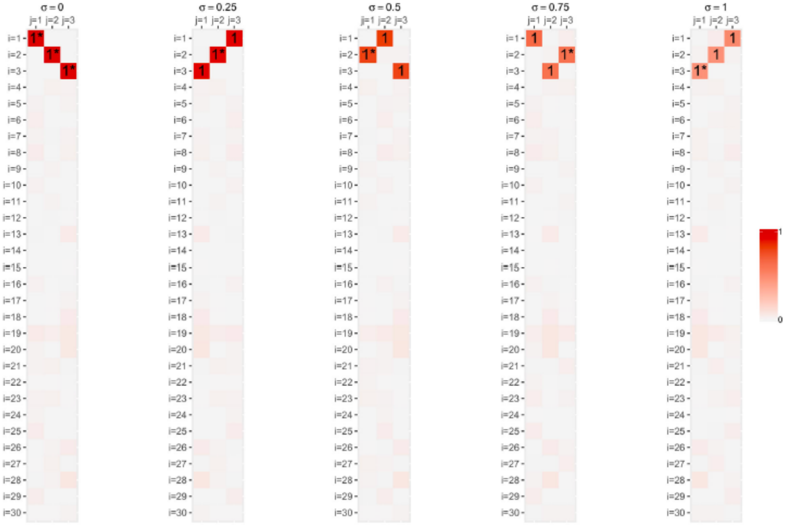
- Highly nonconvex (in M) and linear integer (in H)
- Easily addressed via an alternating approach

$$S(M,H) = \min\{R_{ij}^2(M) : \; h_{ij} = 1\}$$

$$\begin{array}{ll} \max & S(M,H) \\ \text{s.t.} & M^\top M = I \\ & H \in \mathcal{H} \end{array}$$

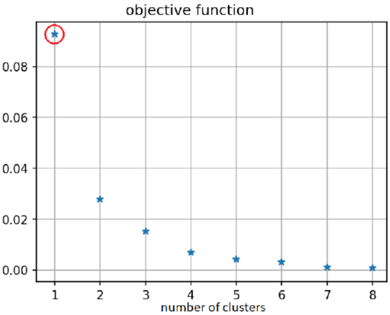
$$\begin{array}{lll} \max & z & \\ \text{s.t.} & z \leq R_{ij}^2(M) h_{ij} + (1 - h_{ij}) & \forall i,j \\ & M^\top M = I & \\ & H \in \mathcal{H} & \end{array}$$

- Highly nonconvex (in M) and linear integer (in H)
- Easily addressed via an alternating approach



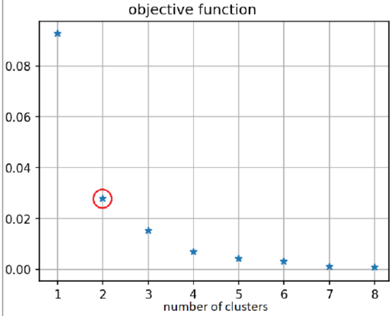
(a) $\mathcal{S}(\boldsymbol{H}, \boldsymbol{M}) = 1$ (b) $\mathcal{S}(\boldsymbol{H}, \boldsymbol{M}) = 0.939$ (c) $\mathcal{S}(\boldsymbol{H}, \boldsymbol{M}) = 0.787$ (d) $\mathcal{S}(\boldsymbol{H}, \boldsymbol{M}) = 0.612$ (e) $\mathcal{S}(\boldsymbol{H}, \boldsymbol{M}) = 0.460$

Seeking Interpretability in Clustering



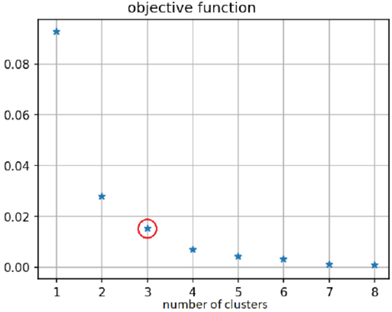
Chosen explanation

'AGE $\leq$ 100'



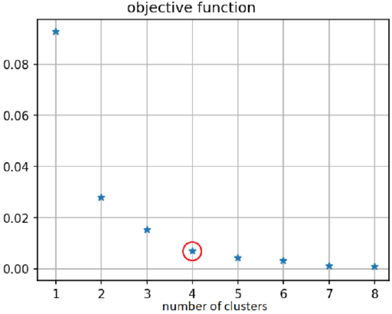
Chosen explanation

'(NOX $\leq$ 0.61) AND (RAD $\leq$ 8)',
'(INDUS $>$ 12.83) AND (TAX $>$ 403)'



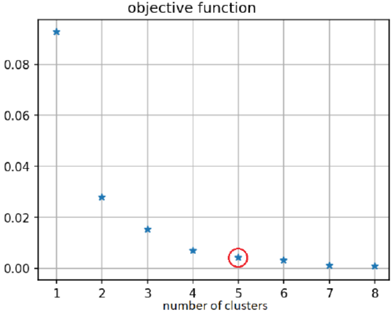
Chosen explanation

'(INDUS $>$ 12.83) AND (TAX $>$ 403)',
'(INDUS $\leq$ 12.83) AND (CHAS $=$ 0)',
'(CHAS $=$ 1) AND (DIS $\leq$ 6.27)'



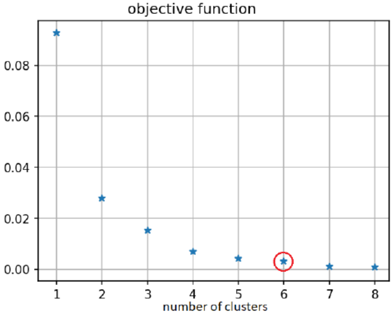
Chosen explanation

'(RAD $>$ 8) AND (CHAS $=$ 0)',
'(AGE $>$ 65.4) AND (RAD $\leq$ 6)',
'(AGE $\leq$ 52.4) AND (NOX $\leq$ 0.52)',
'(CHAS $=$ 1) AND (DIS $\leq$ 6.27)'



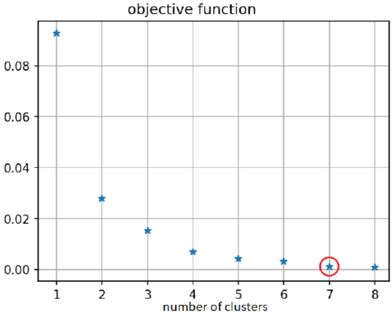
Chosen explanation

'(INDUS $>$ 18.1) AND (CHAS $=$ 0)',
'(ZN $\leq$ 20) AND (INDUS $\leq$ 10.01)',
'(RAD $>$ 8) AND (CHAS $=$ 0)',
'(ZN $>$ 20) AND (AGE $\leq$ 52.4)',
'(CHAS $=$ 1) AND (DIS $\leq$ 6.27)'



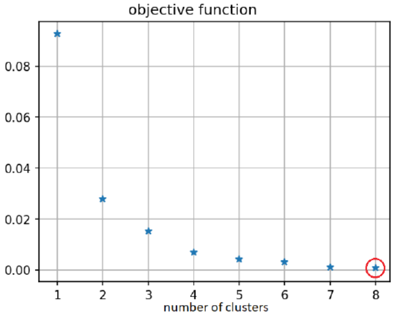
Chosen explanation

'(CHAS $= 1$) AND (DIS ≤ 6.27)',
'(RAD > 8) AND (B > 290.27)',
'(ZN ≤ 20) AND (INDUS ≤ 10.01)',
'(INDUS > 18.1) AND (CHAS $= 0$)',
'(RAD > 8) AND (B ≤ 290.27)',
'(ZN > 20) AND (AGE ≤ 52.4)'



Chosen explanation

'(RAD $>$ 8) AND (B $\leq$ 290.27)',
'(AGE $\leq$ 52.4) AND (ZN $\leq$ 28)',
'(RAD $>$ 8) AND (B $>$ 290.27)',
'(ZN $>$ 42.5) AND (CHAS $=$ 0)',
'(CHAS $=$ 1) AND (DIS $\leq$ 6.27)',
'(AGE $>$ 65.4) AND (INDUS $\leq$ 12.83)',
'(INDUS $>$ 18.1) AND (CHAS $=$ 0)'



Chosen explanation

'(CHAS $= 1$) AND (PTRATIO $\leq$ 18.6)',
'(INDUS > 18.1) AND (CHAS $= 0$)',
'(RAD > 8) AND (B > 290.27)',
'(AGE ≤ 52.4) AND (ZN ≤ 28)',
'(RAD > 8) AND (B ≤ 290.27)',
'(RAD > 8) AND (CHAS $= 1$)',
'(ZN > 42.5) AND (CHAS $= 0$)',
'(AGE > 65.4) AND (INDUS $\leq$ 12.83)'