Group-wise Principal Component Analysis

José Camacho, Edoardo Saccenti, Roberto Therón

SSC15
NAANTALI • FINLAND
19-22 June 2017
15th Scandinavian Symposium on Chemometrics

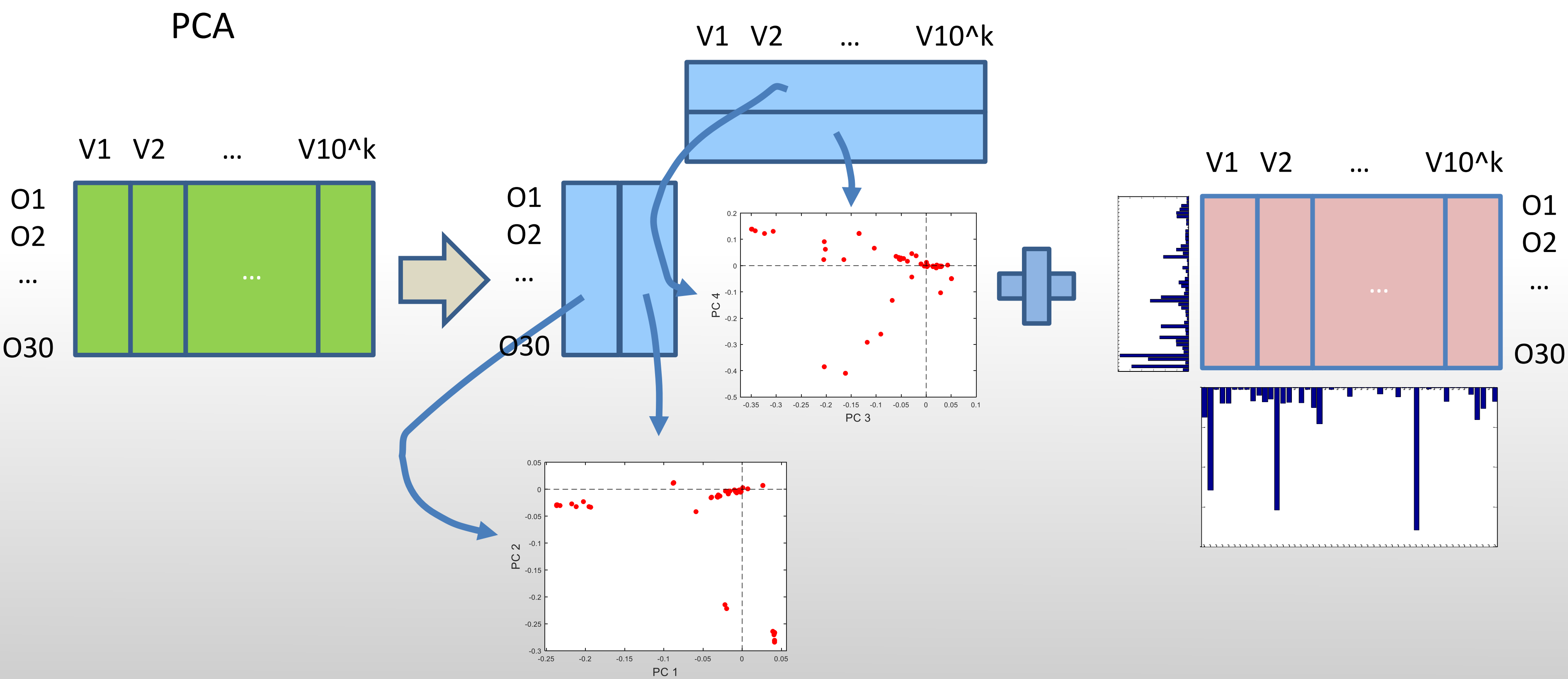
Network Engineering & Security Group
http://nesg.ugr.es
ugr Universidad de Granada

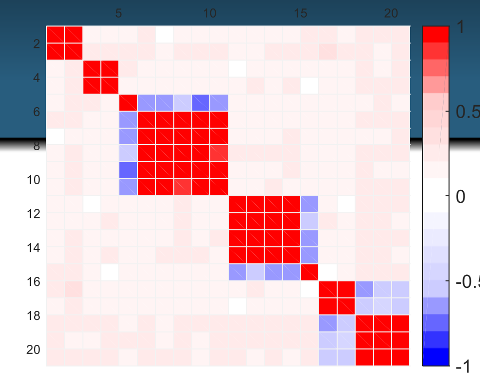
Exploratory Data Analysis (EDA)

- ***(Human) Learning from data****: to improve the understanding of a phenomenon of interest by analyzing data collected on a number of (hopefully) relevant variables.*
  - *Statistics*
  - *Visual Analytics*
  - *Machine learning*
  - *…*

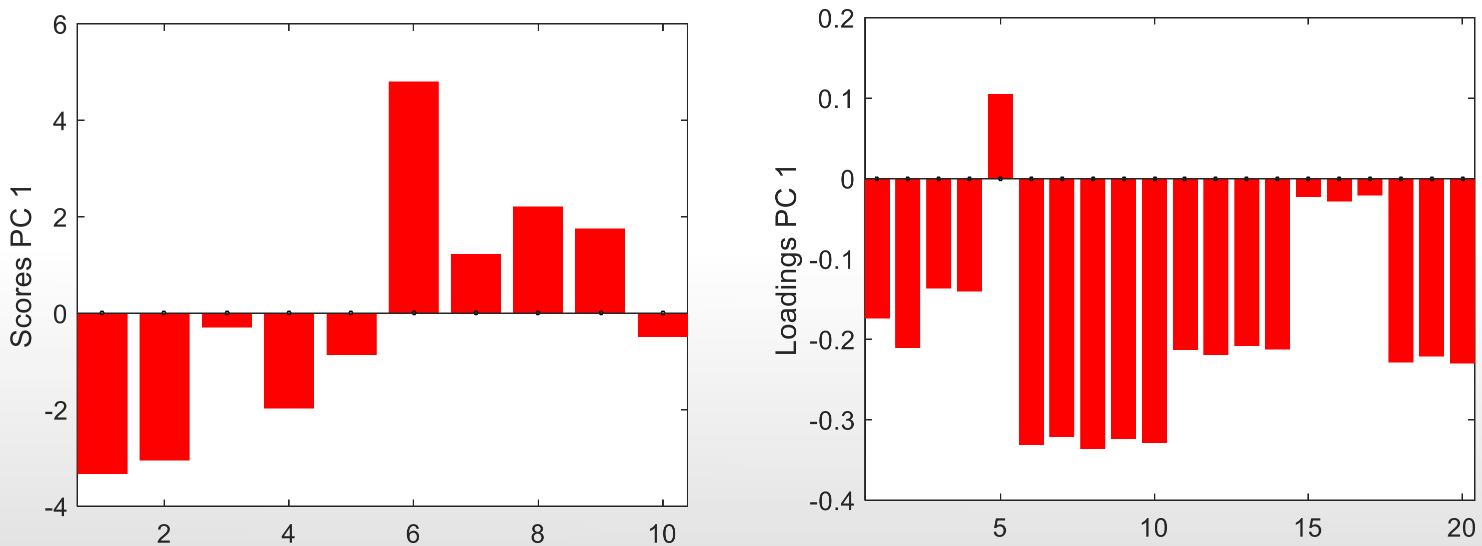
- Multivariate EDA approach: Matrix Factorization

$$\mathbf{X}=\mathbf{T}_A \cdot \mathbf{P}_A^T+\mathbf{E}_A$$

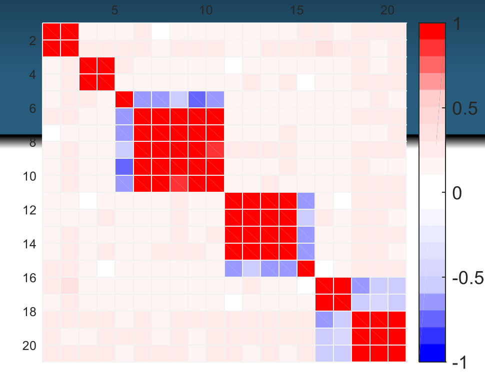




- PCA
  - Structure ≈ Maximum variance
  - PCA for EDA? ➔ X(20x10) = [[1:2], [3:4], [5:10], [11:15], [16:20]]

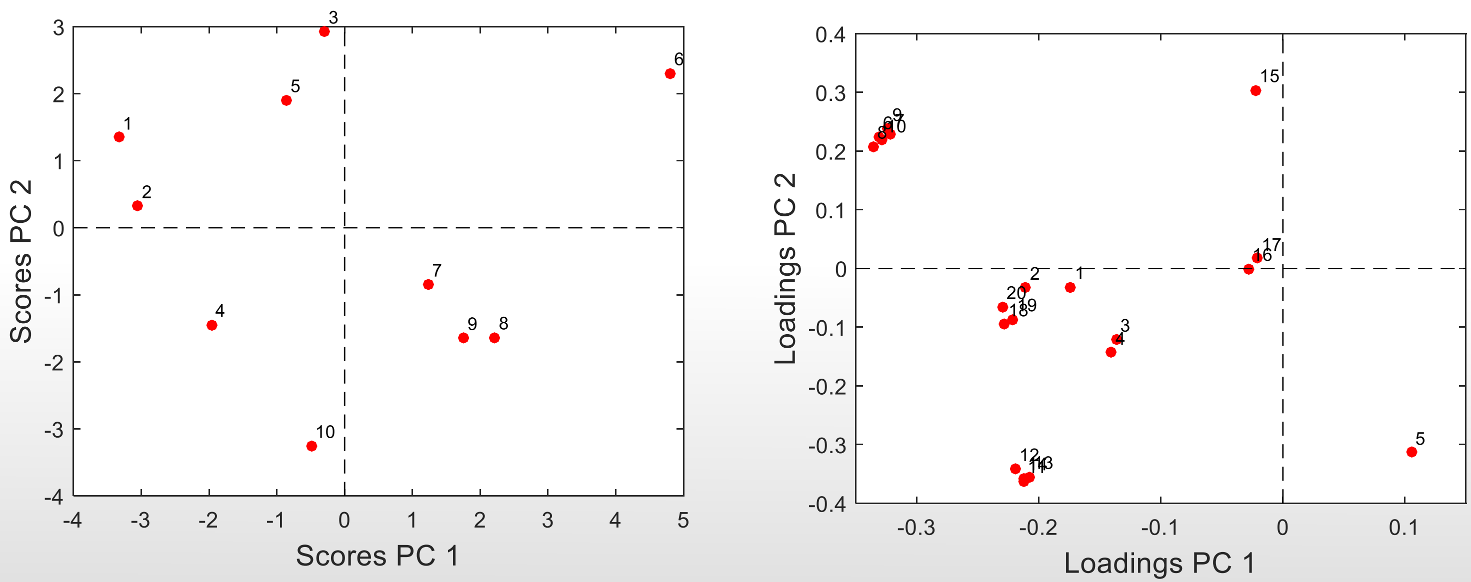


First PC



PCA

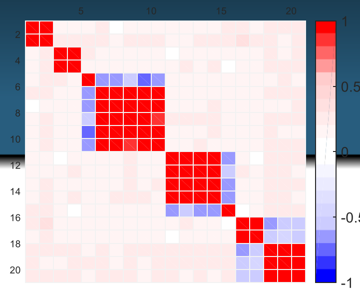
- Structure ≈ Maximum variance
- PCA for EDA? ➔ X(20x10) = [[1:2], [3:4], [5:10], [11:15], [16:20]]



First 2 PCs

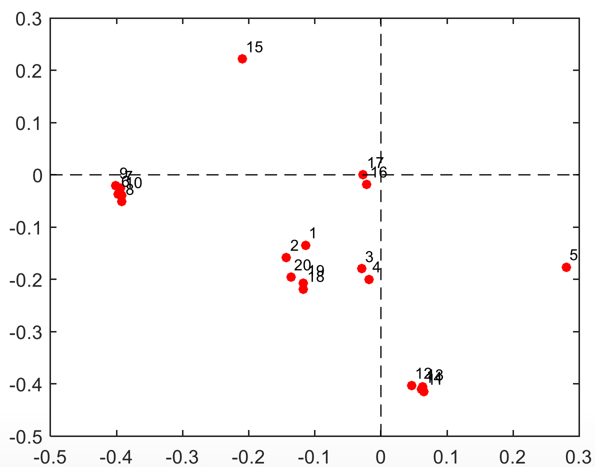
- PCA ← ? → Relationship among Variables
  - Jackson, Jolliffe → NO
  - PCA does not distinguish between unique variance and shared variance
    - ✓ Factor Analysis → Model Shared Variance
  - The PCA factorization is poorly interpretable because the principal components are linear combinations of all the variables
    - ✓ Rotation
    - ✓ Sparse Methods

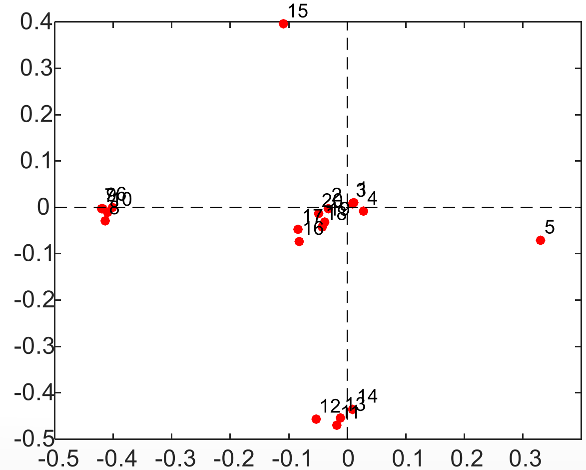
    } → Trade-off between variance and simplicity
  - 1 PC contains many SoV & 1 SoV in many PCs → GPCA (without biplots)



- Variance vs Simplicity:
  - PCA+ Varimax ➔ Rotation depends on #PCs (and scaling)

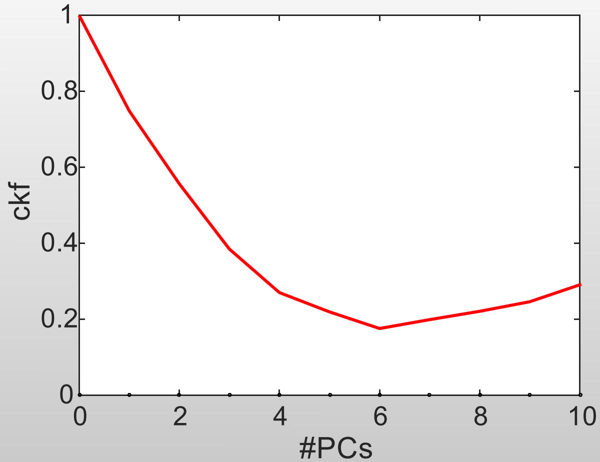
Rotate from 2 PCs

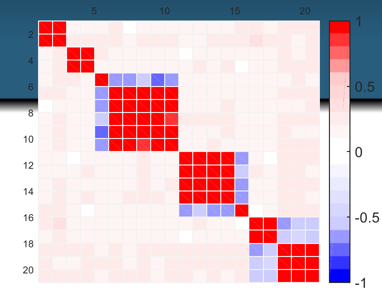




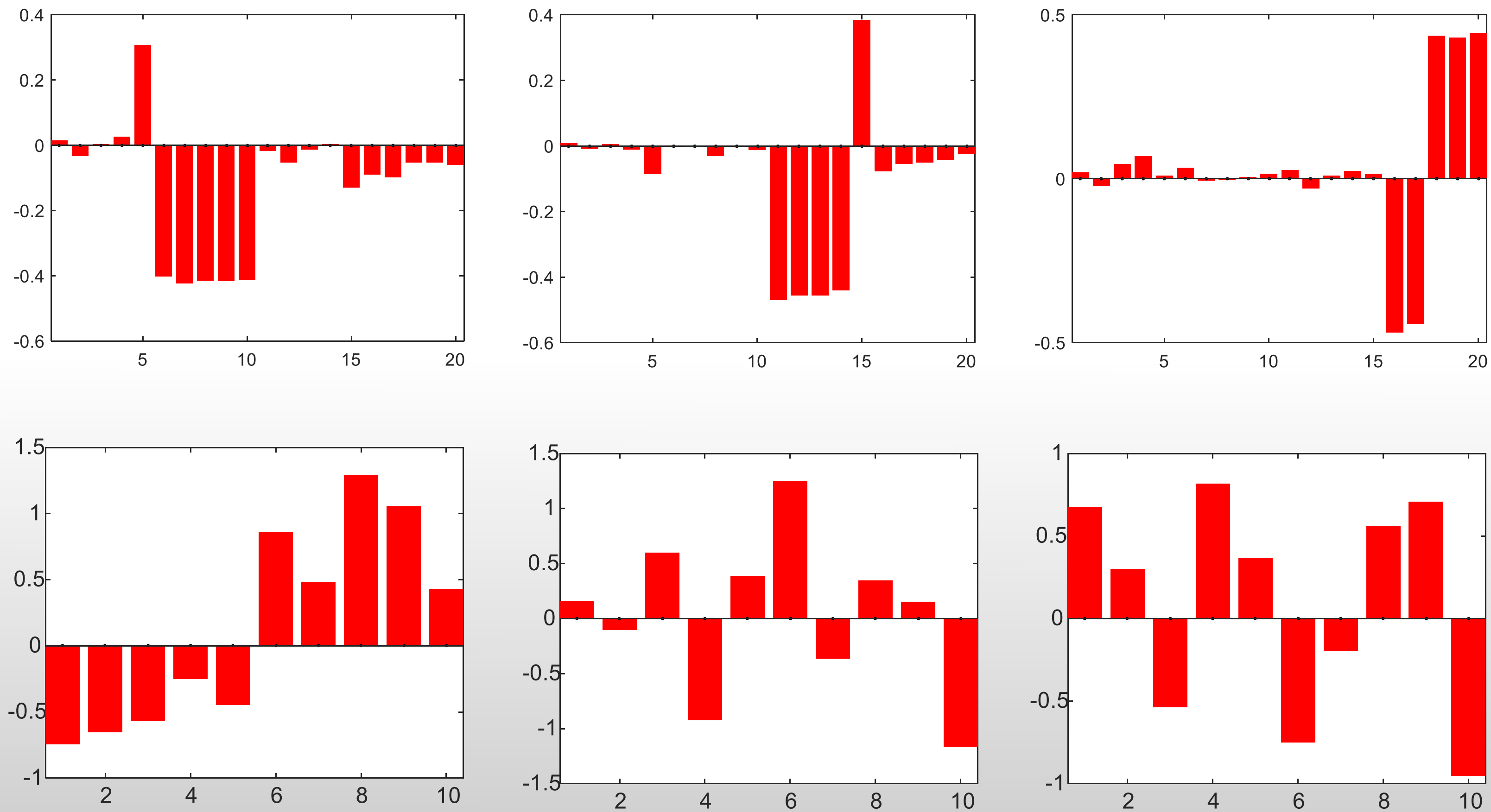
Rotate from 6 PCs

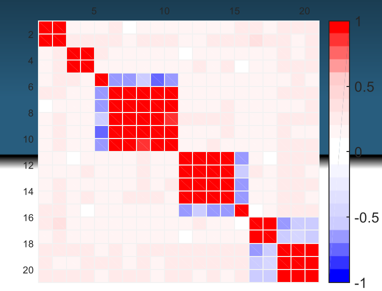
Fast PCA CV (ckf)
J. Chem. 29(2015):467–478



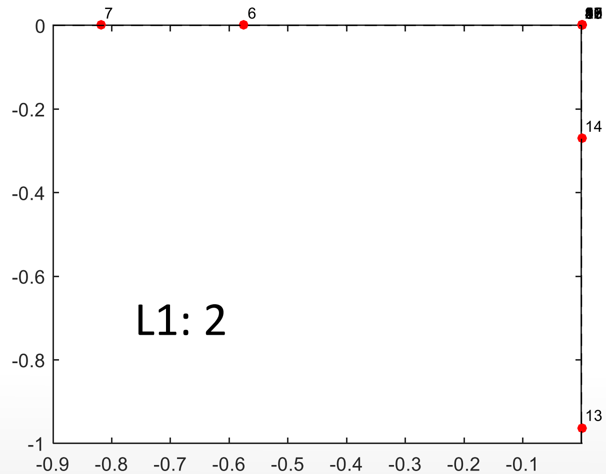


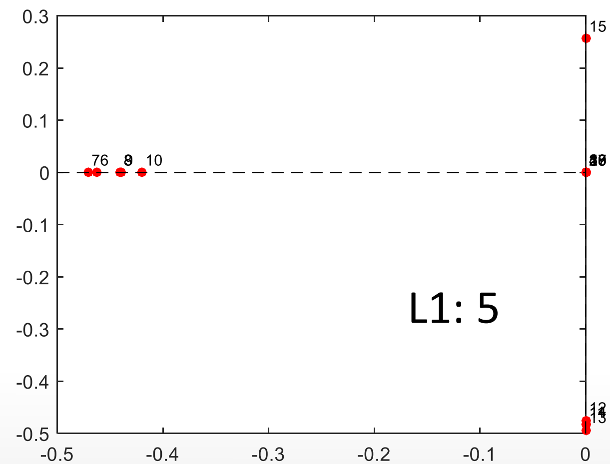
- Variance vs Simplicity:
  - PCA+ Varimax: Rotate from 6 PCs



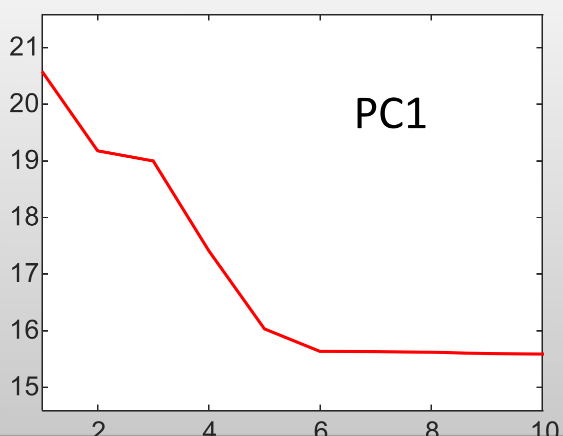


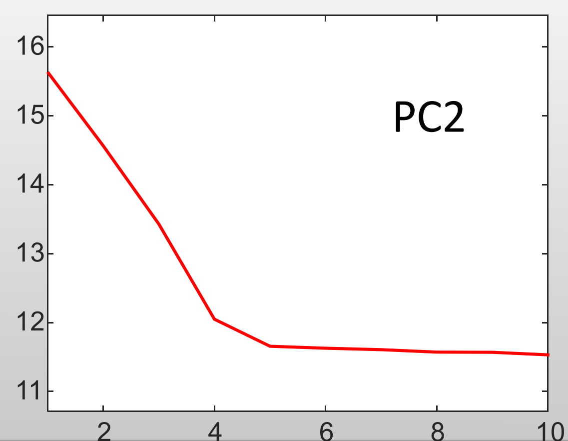
- Variance vs Simplicity:
  - SPCA: Depends on metaparameters

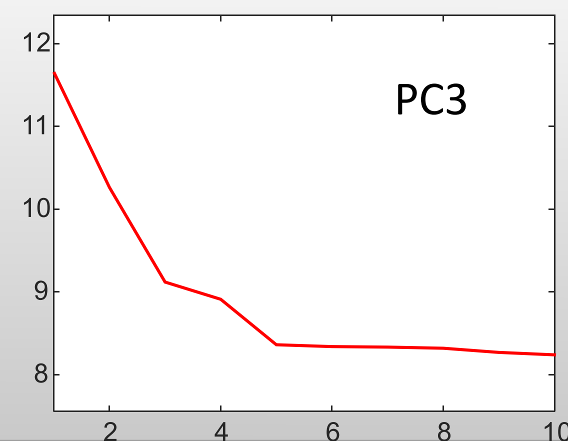




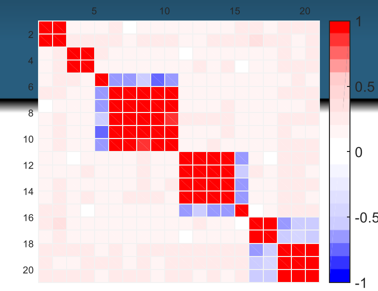
ckf





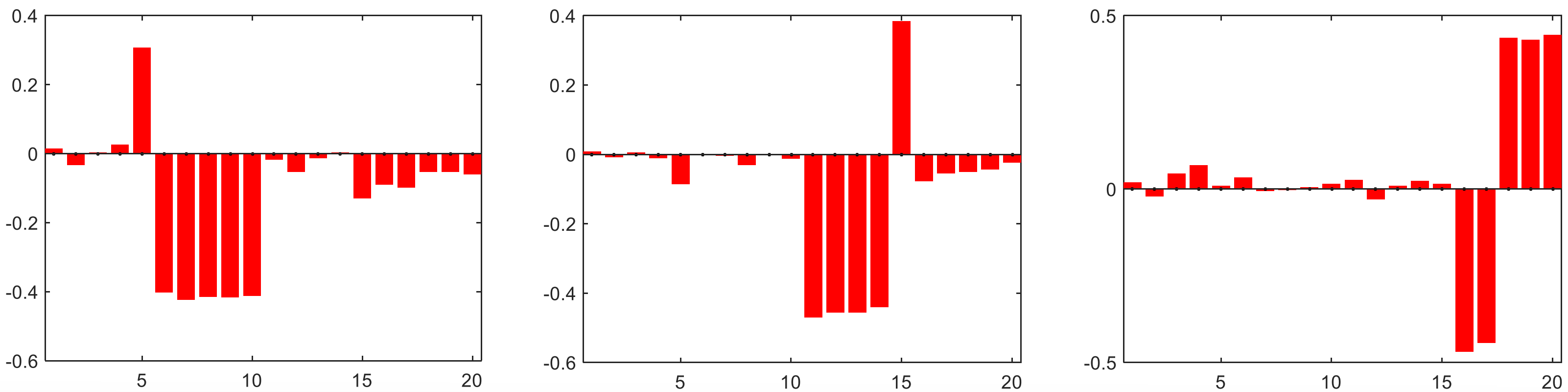


MULTIVARIATE EXPLORATORY DATA ANALYSIS

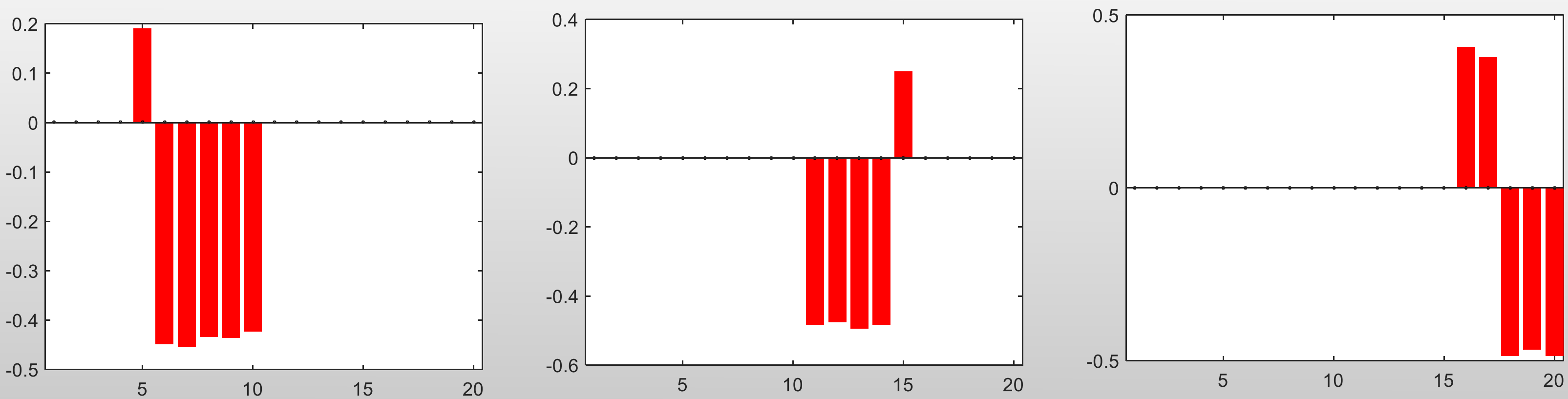


http://nesg.ugr.es

- Variance vs Simplicity:
  - PCA+ Varimax: Rotate from 6 PCs



  - SPCA: L1: [6,5,5]



- Variance vs Simplicity:
  - Result depends on a good choice of metaparameters
  - Typical approach (regression) inherited ➔ CV
  - Risks
    - Prediction ≠ Interpretation
      - Oversimplify / Overcomplicate
      - Application to non-sparse data
    - PCA CV ≠ PLS CV (problem with independent variables, see Journal of Chemometrics, 2012, 26 (7): 361-373.)

Missing-Data for EDA (MEDA)

Chemometrics and Intelligent Laboratory Systems 103 (2010) 8–18

ELSEVIER

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

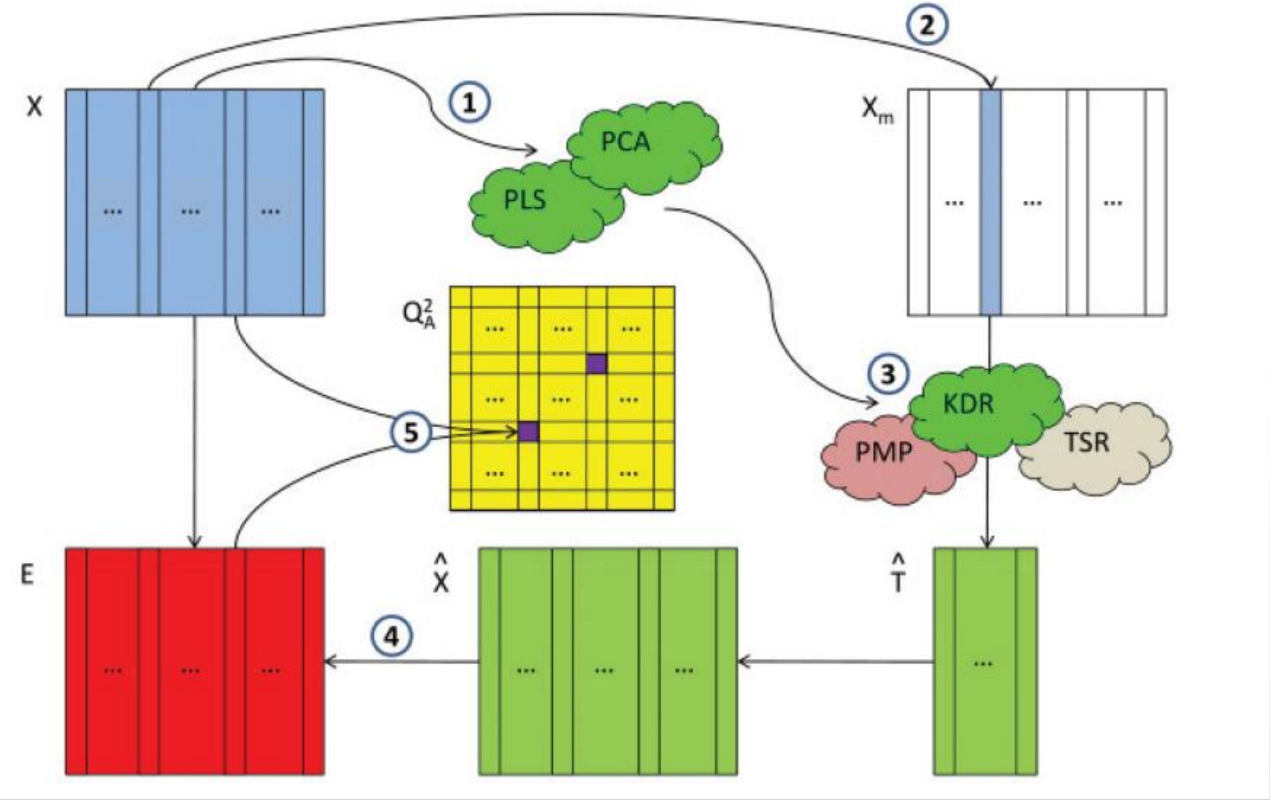
Missing-data theory in the context of exploratory data analysis

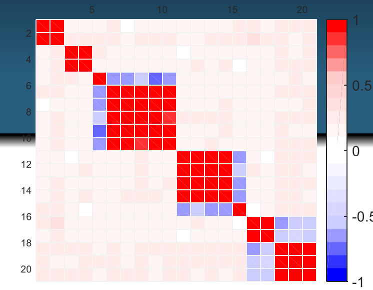
José Camacho

Departamento de Teoría de la Señal, Telemática y Comunicaciones, Universidad de Granada, 18071, Granada, Spain

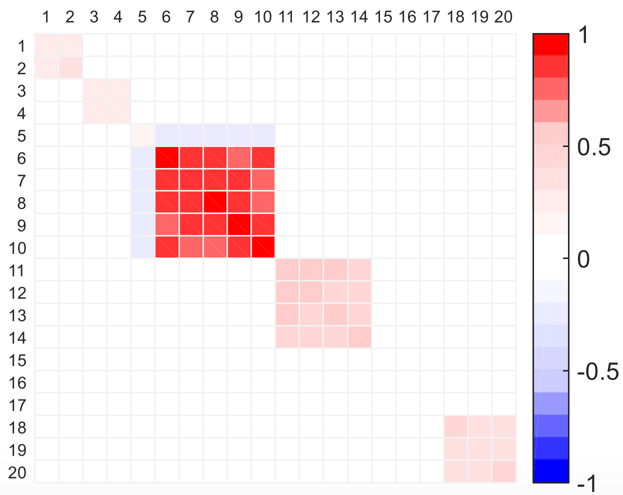
Instead of changing the model, change the visualization

- Missing-Data for EDA (MEDA)

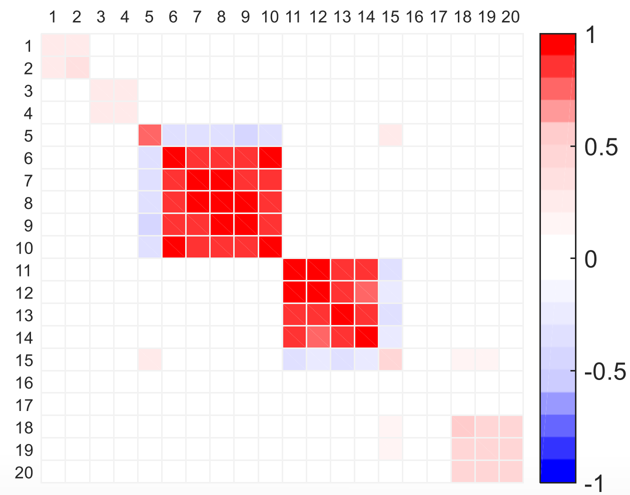




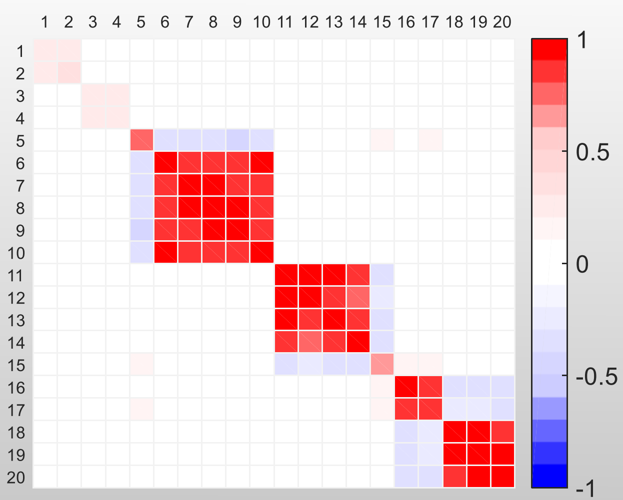
- MEDA:



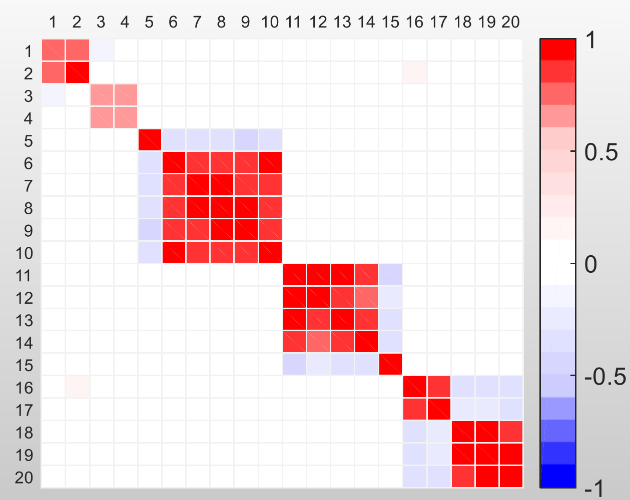
1 PC



1-2 PCs



1-3 PCs



1-4 PCs

Group-wise PCA:

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2017, VOL. 0, NO. 0, 1–12
https://doi.org/10.1080/10618600.2016.1265527

Taylor & Francis
Taylor & Francis Group

Check for updates

Group-Wise Principal Component Analysis for Exploratory Data Analysis

José Camacho [a], Rafael A. Rodríguez-Gómez [a], and Edoardo Saccenti [b]

[a]Department of Signal Theory, Networking and Communication, University of Granada, Granada, Spain; [b]Laboratory of Systems and Synthetic Biology, Wageningen University & Research Center, Wageningen, The Netherlands

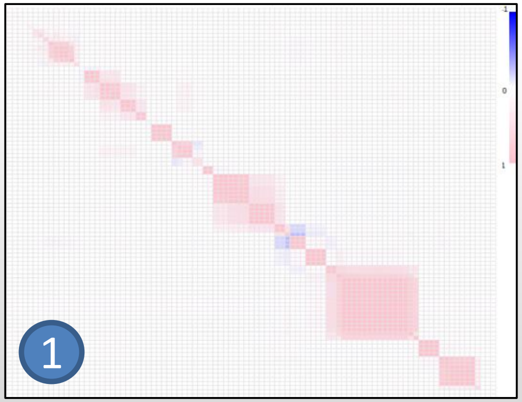
X Do not force a data-driven simple structure

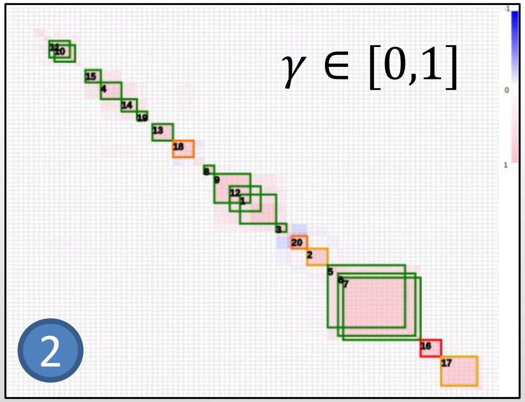
✓ Find the structure and force it in the model

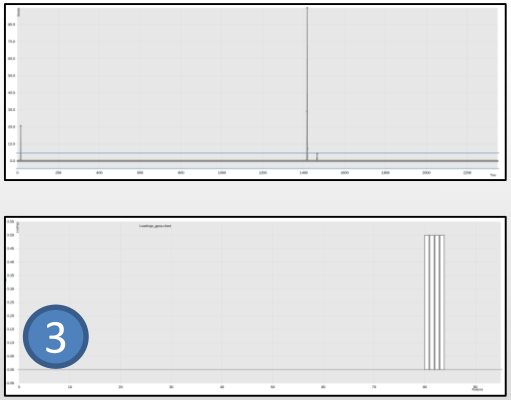
1 PC ←→ 1 SoV

Group-wise PCA:

- Three steps:
  1. Find structure (MEDA)
  2. Identify Groups of Variables (Group Identification Algorithm or GIA)
  3. Calibrate a group-wise PCA model (GPCA)



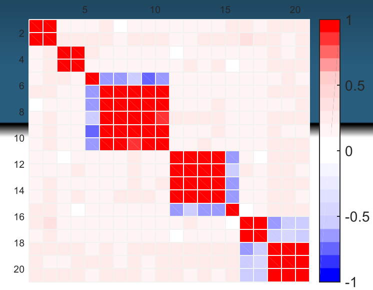




Group-wise PCA:

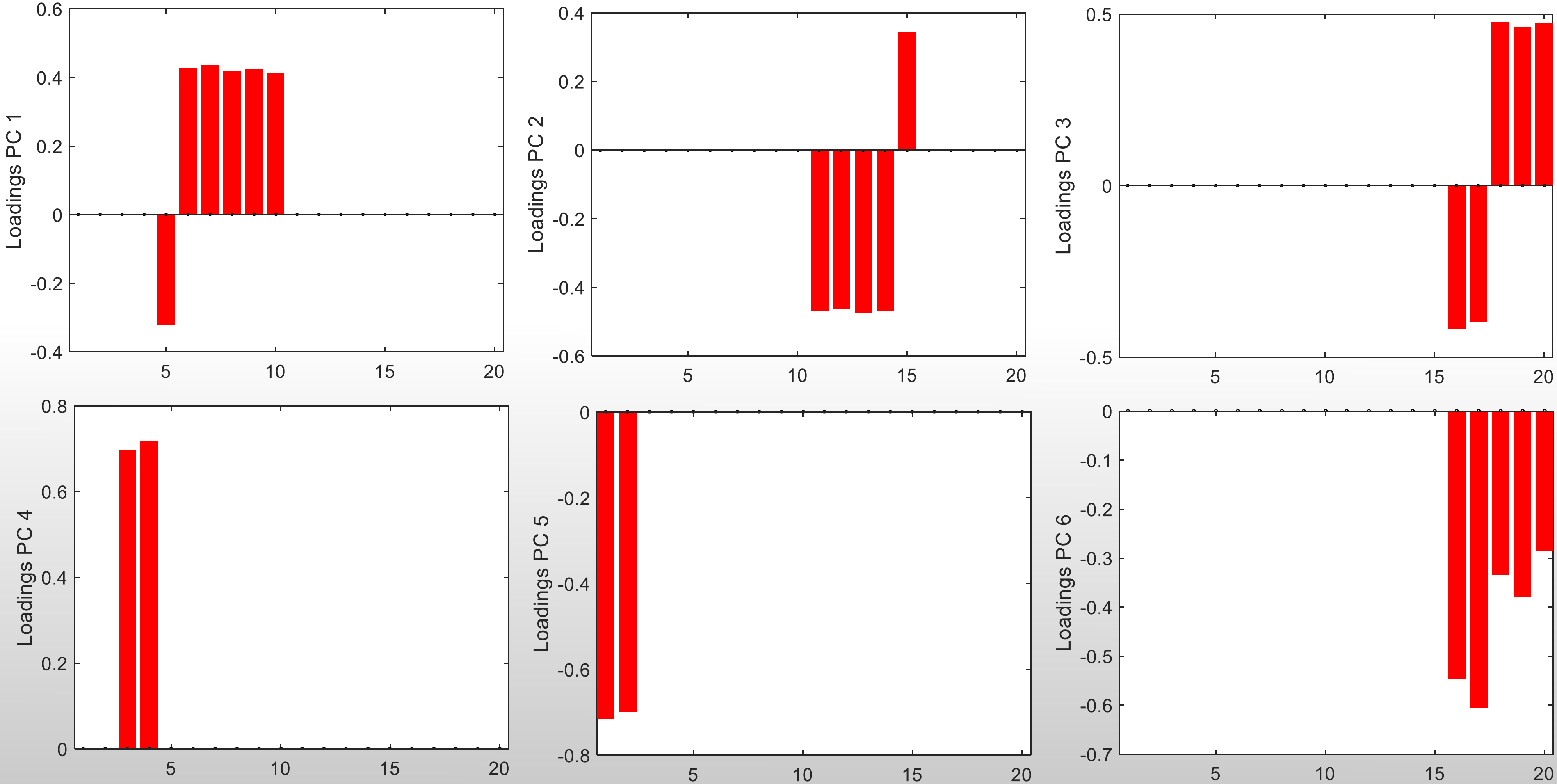
- Initialize: $$\mathbf{C} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$$ $$\mathbf{B} = \mathbf{I},$$
- For each PC
  - For each (k-th) group in GIA $$\mathbf{C}^k = \mathbf{C}$$ $$c_{lm}^k = 0,\ \forall l \notin S_k \ or\ \forall m \notin S_k.$$
  - Compute 1 PC: $$\mathbf{C}^k = \mathbf{p}^k(\sigma^k)^2(\mathbf{p}^k)^{\mathrm{T}} + \mathbf{E}^k.$$
- Choose PC with most variance: $$\mathbf{p}_a = \arg\min_{\mathbf{p}^k} \|\mathbf{E}^k\|_F$$ $$\mathbf{t}_a = \mathbf{X}\mathbf{p}_a.$$
- Deflate (Mackey‘08): $$\mathbf{q} = \mathbf{B}\mathbf{p}_a$$ $$\mathbf{C} = (\mathbf{I} - \mathbf{q}\mathbf{q}^{\mathrm{T}})\mathbf{C}(\mathbf{I} - \mathbf{q}\mathbf{q}^{\mathrm{T}})$$ $$\mathbf{B} = \mathbf{B}(\mathbf{I} - \mathbf{q}\mathbf{q}^{\mathrm{T}}).$$

- GPCA: X(20x10) = [1, 2, 3, [4:5], [6:9], 10]

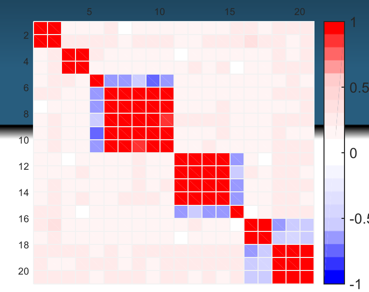


$\gamma = 0.2$

Visually selected

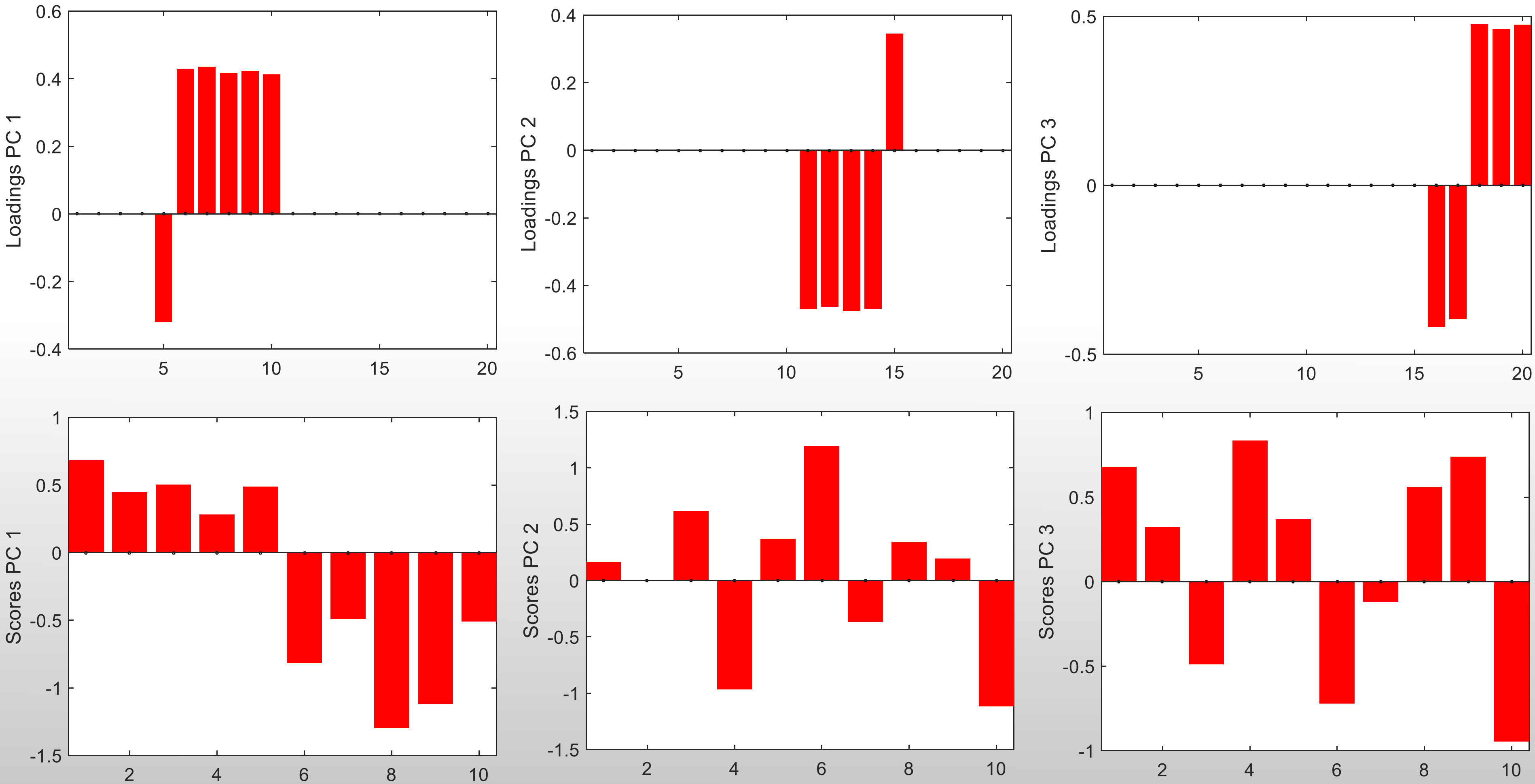


- GPCA: X(20x10) = [1, 2, 3, [4:5], [6:9], 10]

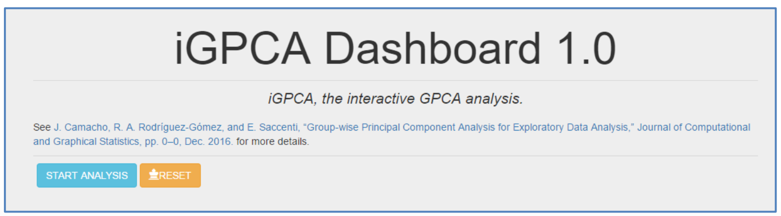


$\gamma = 0.2$

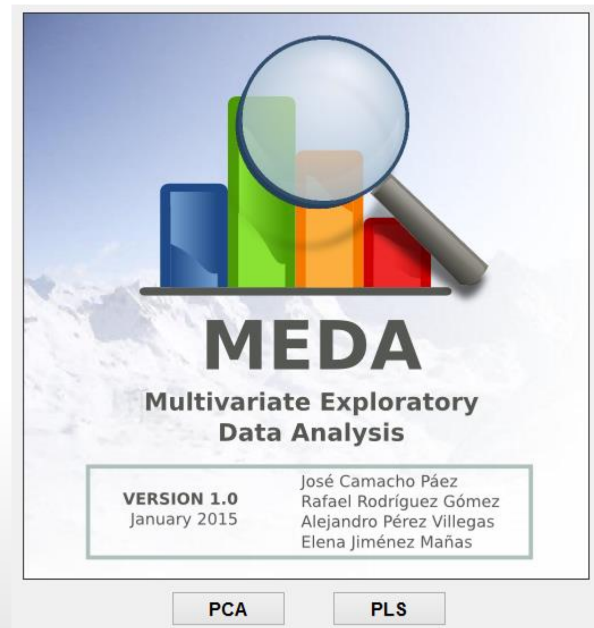
Visually selected

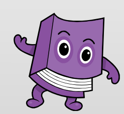


Want to play with GPCA?



http://nesg.ugr.es:5003





ChemoLab, (2015) 143: 49

https://github.com/josecamachop/MEDA-Toolbox

- Multivariate Exploratory Data Analysis can be tricky
  - Variance vs Simplicity selected by prediction (CV)
  - Rather: find structure in a EDA manner and impose it in model.

- GPCA:
  - It is sparse when data is group-wise (in the variables)
  - Only correlated variables (1 SoV) in a PC
  - Does not oversimplify/overcomplicate structure
  - Metaparameter selected from visualization (perfect for EDA)
  - Still, you can always set GPCA by CV

Group-wise Principal Component Analysis

José Camacho, Edoardo Saccenti, Roberto Therón

This work is partly supported by the Spanish Ministry of Economy and Competitiveness and FEDER funds through project TIN2014-60346-R

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Unión Europea
Fondo Europeo de Desarrollo Regional