ComDim-ICA
Multiblock Independent Components Analysis

D. N. Rutledge
UMR GENIAL, INRA, AgroParisTech, Université Paris-Saclay Paris/ France
rutledge@agroparistech.fr

L. Schmidtke
National Wine and Grape Industry Centre, Charles Sturt University, Wagga Wagga / Australia
LSchmidtke@csu.edu.au
ComDim

• ComDim or Common Components and Specific Weights Analysis (CCSWA) is an exploratory multi-block data analysis method
• Simultaneous analysis of several data tables with different variables describing the same samples
• Determines a common space for all blocks
• Each block has a specific contribution (salience) to the definition of each dimension of the common space
• Originally developed in sensometrics
• Has been applied to the fusion of analytical data


V. Cariou, D. Jouan-Rimbaud Bouveresse E.M. Qannari, D.N. Rutledge "ComDim methods for the analysis of multiblock data in a data fusion perspective" in Data Fusion Methodology and Applications (Data Handling in Science and Technology), (ed. Marina Cocchi) Elsevier Science Publishers, Amsterdam, 2018
Start with $p$ matrices $X_i$ of size $n \times k_i$ ($i = 1$ to $p$)

Each $X_i$ column-centered and scaled by dividing by matrix norm : $X_{s_i}$

For each $X_{s_i}$, an $n \times n$ scalar product matrix $W_i$ can be computed as :

$$W_i = X_{s_i} \cdot X_{s_i}^T$$

$W_i$ reflect the dispersion of the *samples* in the space of that table

Each $W_i$ is multiplied by a scalar, $\lambda_i$ (initially all set to 1)

At each iteration, a sum of the $p$ weighted $W_i$ matrices is computed, resulting in a global $W_G$ matrix
Original ComDim algorithm

\[ W = X.X^T \]

\[ \lambda_1 \cdot W_1 \]

\[ \lambda_2 \cdot W_2 \]

\[ \text{Dif} = \sum [(X_1.X_1^T - \lambda_1 q.q^T) + (X_2.X_2^T - \lambda_2 q.q^T)] \]

\[ \text{Aux} = I - q \cdot q^T \]

\[ X_1 = \text{Aux} \cdot X_1 \]

\[ X_2 = \text{Aux} \cdot X_2 \]

\[ \text{Dif}^2 - \text{Dif}_{n-1}^2 < \text{limit} \]
Original ComDim algorithm

\[ W_k = \sum_{dim=1}^{r} \lambda^{(k)}_{dim} q_{dim} q'_{dim} + E_k \]

• Sequential determination of:
  – Global scores of individuals on each CC: \( q_{dim} \)
  – Saliences of tables: \( \lambda^k_{dim} \)
  – Loadings of variables: \( u^k_{dim} \)
  – Local scores of individuals for each table: \( t^k_{dim} \)
  – Sum of saliences of all tables for each CC
  – Sum of saliences of all CCs for each table
  – Variance extracted by each CC
  – ...
New ComDim algorithm
(proposed by M. Hanafi)

Start with $p$ matrices $X_i$ of size $n \times k_i \ (i = 1 \text{ to } p)$

Each $X_i$ column-centered and scaled by dividing by matrix norm to give $Xs_i$

Each $Xs_i$ is multiplied by a scalar, $\sqrt{\lambda_i}$ (initially all set to 1)

At each iteration, the $p$ weighted $Xs_i$ matrices are concatenated column-wise, resulting in a global $X_G$ matrix

$$X_G = [\sqrt{\lambda_1}Xs_1, \sqrt{\lambda_2}Xs_2, \sqrt{\lambda_3}Xs_3, \ldots, \sqrt{\lambda_p}Xs_p]$$

M. Hanafi, Personal communication
7th International Meeting on Chemometrics and Quality, 23-25 October 2018, Fès, Morocco
New ComDim algorithm

\[X_G = U_G \cdot S V_G \cdot V_G\]

\[\lambda_1 = q^T \cdot X_1 \cdot X_1^T \cdot q\]

\[\lambda_2 = q^T \cdot X_2 \cdot X_2^T \cdot q\]

\[\text{Dif} = \sum \left[ (X_1 \cdot X_1^T \cdot \lambda_1 \cdot q \cdot q^T) + (X_2 \cdot X_2^T \cdot \lambda_2 \cdot q \cdot q^T) \right]\]

\[\text{Dif} = \text{Dif}_{n-2} - \text{Dif}_{n-1} < \text{limit}\]

\[\text{aux} = I - q \cdot q^T\]

\[X_1 = \text{aux} \cdot X_1\]

\[X_2 = \text{aux} \cdot X_2\]
Multi-Block ICA!

\[ \lambda_1 = q^T X_1 X_1^T q \]
\[ \lambda_2 = q^T X_2 X_2^T q \]

\[ \text{Dif} = \Sigma [(X_1 X_1^T - \lambda_1 q q^T) + (X_2 X_2^T - \lambda_2 q q^T)] \]

\[ \text{Dif}_{n^2} - \text{Dif}_{n-1^2} < \text{limit} \]

\[ X_1 = \text{aux} X_1 \]
\[ X_2 = \text{aux} X_2 \]

\[ \text{aux} = I - q q^T \]

\[ X_G = \Lambda G \cdot S_G \]

\[ X_1 = \sqrt{\lambda_1} \cdot X_1 \]
\[ X_2 = \sqrt{\lambda_2} \cdot X_2 \]

\[ X_G = \sqrt{\lambda_1} \cdot X_1 \]
\[ X_G = \sqrt{\lambda_2} \cdot X_2 \]

\[ X_G = \Lambda G \cdot S_G \]
Independent Components Analysis

Aims to extract the unknown source signals mixed together in unknown proportions in the observed signals that form the rows of the data matrix.

ICs or Source Signals : analogous to PCA Loadings

Proportions : analogous to PCA Scores

D. Jouan-Rimbaud Bouveresse, D.N.Rutledge
"Independent Components Analysis: Theory And Applications" in
Resolving Spectral Mixtures, (ed. C. Ruckebusch)
Independent Components Analysis (ICA)

Data matrix $\rightarrow$ a set of observed signals, where:

- each observed sensor signal, $x_i$, is the weighted sum of pure source signals, $s_j$
- the weighting coefficients, $a_{ij}$, are proportions of the source signals, $s_j$

In matrix notation:

$$x_1 = a_{11} \ast s_1 + a_{12} \ast s_2$$
$$x_2 = a_{21} \ast s_1 + a_{22} \ast s_2$$
$$\vdots$$
$$x_n = a_{n1} \ast s_1 + a_{n2} \ast s_2$$

$$X = A \ast S$$
Independent Components Analysis
ICA looks for “meaningful" vectors

Hypotheses:

1) No reason for the variations in one pure signal to depend in any way on the variations in another pure signal

   Pure source signals should therefore be « independent »

2) The measured signals being combinations of several independent sources, they should be more gaussian than the sources

   (Central Limit Theorem)
JADE
(Joint Approximate Diagonalization of Eigenmatrices)

- Developed by Cardoso and Souloumiac in 1993

- A blind source separation method to extract independent non-Gaussian sources from signal mixtures with Gaussian noise

- Based on the construction of a fourth-order cumulant array from the data

- Matlab function freely downloadable from
  
  http://perso.telecom-paristech.fr/~cardoso/Algo/Jade/jadeR.m

Cardoso, J-F. and Souloumiac, A.
Blind beamforming for non-Gaussian signals.

D.N. Rutledge, D. Jouan-Rimbaud Bouveresse,
Independent Components Analysis with the JADE algorithm

D.N. Rutledge, D. Jouan-Rimbaud Bouveresse,
Corrigendum to “Independent Components Analysis with the JADE algorithm”
The JADE algorithm: a multi-step procedure

1. Row-centered $X_{rc}$
   - $X_{rc}$: $r \times c$
   - $X_{rc}$ undergoes Singular Value Decomposition (SVD) to obtain the "normed" scores $U$ and singular values matrix $S$.
   - SVD: $X_{rc} = U S V^T = U S$
   - $U$: $r \times n$
   - $S$: $n \times n$
   - $V$: $c \times n$

2. Whitening matrix
   - $B = \sqrt{c} \times S^{-1} \times U^T$
   - $B$: $n \times r$
   - The whitening matrix $P_w^T = \sqrt{c} \times P^T$ produces the whitened matrix $P_w^T = \sqrt{c} \times P^T$.

3. 4th order tensor of loadings cumulants
   - $K$: $n \times n \times n \times n$
   - $K$: whitened matrix
   - Decompose $K$ to obtain orthogonal $M_i$ and rotation $V$ matrices.

4. Whitening matrix
   - $B^T \times V^T$
   - The whitening matrix $W = B^T \times V^T$ results in a demixing matrix.

5. Calculate signals
   - $W^T \times X$
   - The matrix of pure source signals $S$ is calculated as $S^{-1}$.
   - Calculate proportions $A$ = Mixing matrix

Diagram:
- $X_{rc}$: $r \times c$
- $U$: $r \times n$
- $S$: $n \times n$
- $V$: $c \times n$
- $B$: $n \times r$
- $P_w^T$: $n \times r$
- $K$: $n \times n \times n \times n$
- $M_i$: $n \times n$
- $V$: $n \times n$
- $W$: $r \times n$
- $X$: $r \times c$
- $S$: $n \times c$
- $A$: $r \times n$
Comparison of ICA and ComDim with ComDim-ICA

Application to TD-NMR Lignin-Starch data

20 samples in triplicate, with different characteristics:

- 2 Shapes: Films / Cylinders
- 2 Moisture levels: stabilized in atmospheres at 33% / 75% H₂O
- 5 Lignin concentrations: 0%, 5%, 10%, 15%, 30%

8 types of Time Domain-NMR signals
Comparison of ComDim with ComDim-ICA

ComDim Saliences for CC1, CC2 & CC3
Convergence in 124 mS

ComDim-ICA Saliences for CC1, CC2 & CC3
Convergence in 104 mS
Comparison of ICA and ComDim with ComDim-ICA

Moisture

ICA

ComDim

ComDim-ICA

SNV-scaled Loadings Factor 4

ICA

ComDim

ComDim-ICA
Comparison of ICA and ComDim with ComDim-ICA
Comparison of ICA and ComDim with ComDim-ICA
Conclusion

ComDim-ICA

A non-supervised multi-block method

ICA on iteratively re-weighted concatenated data tables

• Better than ICA on unweighted concatenated data tables

• Better than ComDim
  (PCA on iteratively re-weighted concatenated data tables)
Thank you