CANONICAL REPRESENTATION OF EXPLORATORY MULTIBLOCK METHODS

Mohamed Hanafi

Statistics,Sensometrics and Chemometrics (StatSC), ONIRIS, Nantes, France

École Nationale Vétérinaire, Agroalimentaire et de l'Alimentation Nantes Atlantique

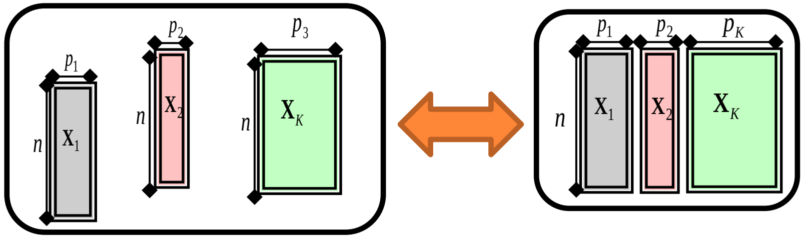
Chimiometrie 2019

31/01-01/02 Montpellier

SUMMARY

- A comprehensive picture of Multiblock methods
  - Two main ideas.
  - Polyomorphism and heterogeneity.
- Canonical representation of multiblock methods
  - Motivation and aim.
  - Definition and properties.
  - Some interests (comparaison of methods)
- Conclusions

Multiblock methods in chemometrics



Acronyms	Names	References
HPCA	**Hierarchical Component Analysis**	**1**
MCOA	**Multiple Co-inertia Analysis**	**2**
STATIS	**Structuration de Tableaux A Trois Indices de la Statistique**	**3**
CCSWA (comdim)	**Common Components and Specific Weights Analysis**	**4**
CPCA	**Consensus Principal Component Analysis**	**5**
PCA	**Principal Component Analysis**	**6**

(1) Wold, S., Kettaneh, N. and Tjessem, K., (1996). Hierarchical multi-block PLS and PC models for easier interpretation and as an alternative to variable selection. Journal of Chemometrics. (10), 463-482

(2) Chessel, D. & Hanafi, M. (1996). Revue de Statistique Appliquée, 44, 35-60.

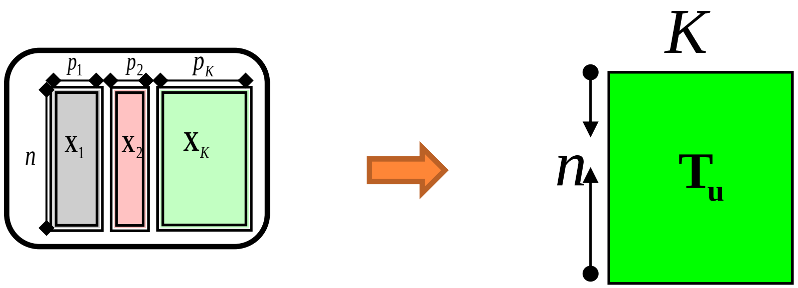
(3) Lavit C. (1976). Analyse conjointe de tableaux quantitatifs. Masson, Paris.

(4) Qannari E.M., Wakeling I., Courcoux Ph., MacFie J.H. (2000). Defining the underlying sensory dimensions. Food Quality and Preference 11, 151-154.

(5) Westerhuis, J. A, Kourti, T., Macgregor J. F. (1998). Analysis of Multiblock and Hierarchical PCA and PLS Models. Journal of Chemometrics. (12), 301-321.

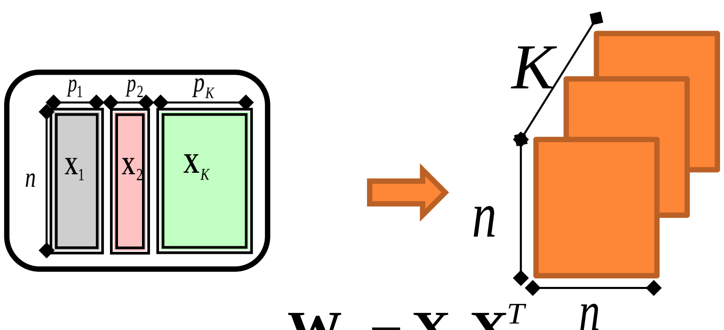
(6) Wold, S., Esbensen,K., Geladi,P. (1987). Chemometrics and Intelligent Laboratory Systems. Volume 2, Issues 1–3, August 1987, Pages 37-52

Two ideas behind methods



One block Analysis (synthetic block)

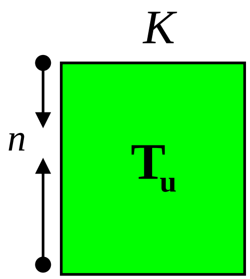
- $\mathbf{T_u} \leftarrow [\mathbf{t}_1 \quad \mathbf{t}_2 \quad \cdots \quad \mathbf{t}_K]$
- $\mathbf{t}_k \leftarrow \mathbf{X}_k \mathbf{u}_k \left(\|\mathbf{u}_k\| = 1 \right)$



Tensor Analysis (tensor of scalar products between individuals)

$$\mathbf{W}_k = \mathbf{X}_k \mathbf{X}_k^T$$

Two ideas behind methods

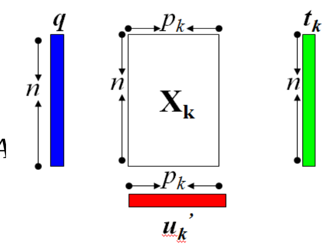


$\sum_{k=1}^{K} \text{cov}^2\left(\mathbf{X}_k \mathbf{u}_k, \mathbf{v}\right)$	$\sum_{k=1}^{K} \text{cov}^4\left(\mathbf{X}_k \mathbf{u}_k, \mathbf{v}\right)$
CPCA MCOA	HPCA

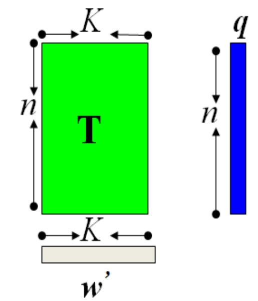
MCOA➔ (deflation on block loadings)

CPCA,HPCA➔ (deflation on global scores)

- $\mathbf{v} \leftarrow \mathbf{v} / \sqrt{\mathbf{v}^T \mathbf{v}}$
- $\mathbf{u}_k \leftarrow \begin{cases} \mathbf{X}_k^T \mathbf{v} & HPCA \\ \mathbf{X}_k^T \mathbf{v} / \left\| \mathbf{X}_k^T \mathbf{v} \right\| & CPCA \end{cases}$
- $\mathbf{t}_k \leftarrow \mathbf{X} \mathbf{u}_k$



- $\mathbf{T} \leftarrow \left[\mathbf{t}_1 \quad \mathbf{t}_2 \quad \cdots \quad \mathbf{t}_K\right]$
- $\mathbf{w} \leftarrow \mathbf{T}^T \mathbf{v}$
- $\mathbf{v} \leftarrow \mathbf{T}\mathbf{w}$



Hanafi, M., Kohler, A., Qannari E. M. (2011). Connections between Multiple Co-inertia Analysis and Consensus Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, 106 (1) : 37-40.

Hanafi, M., Kohler, A., Qannari E. M. (2010). Shedding new light on Hierarchical Principal Component Analysis. *Journal of Chemometrics*, 24, (1): 703-709.

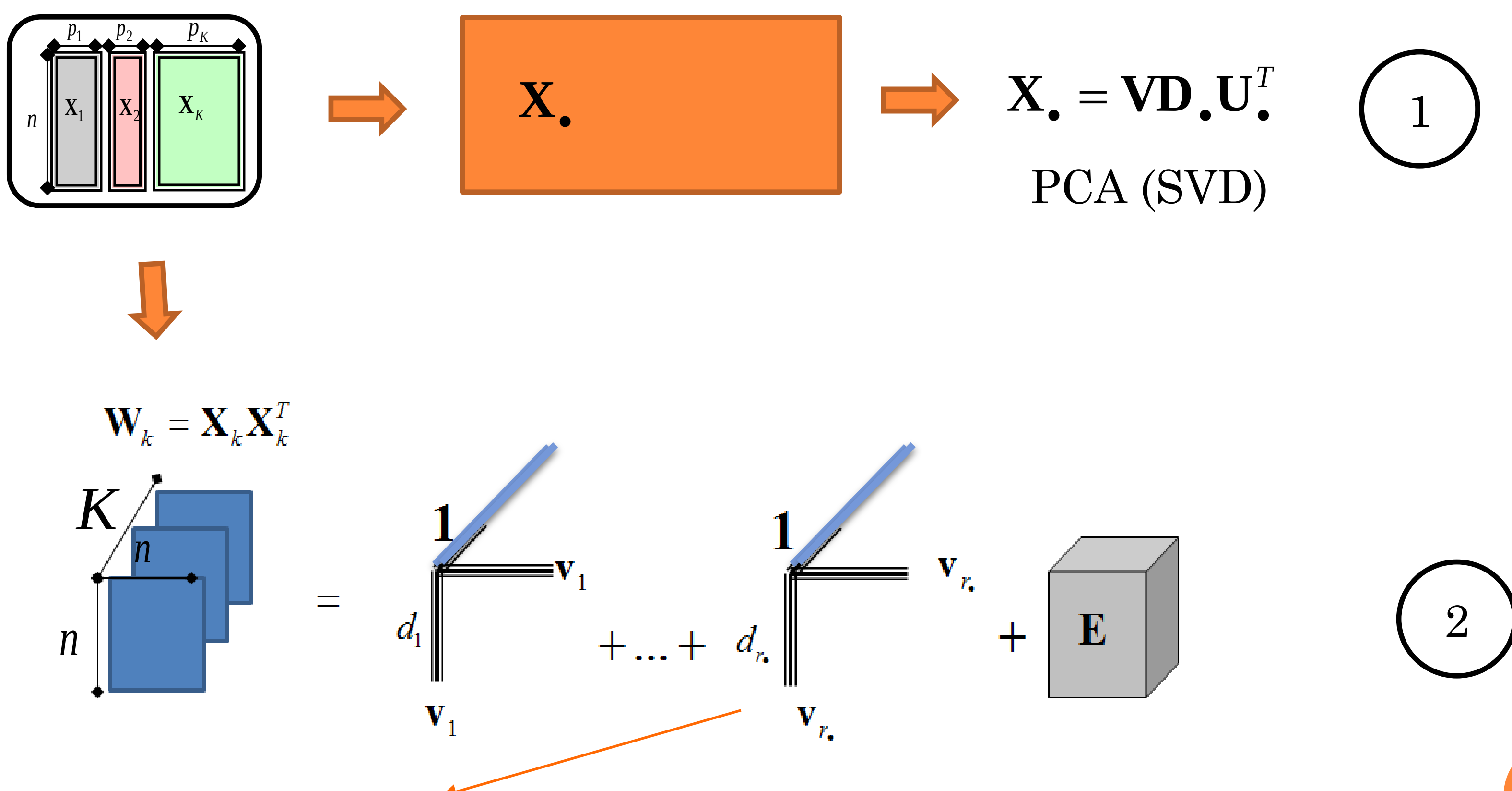
Hanafi, M., Qannari E.M. (2008). Nouvelles propriétés de l'Analyse en Composantes Communes et Poids Spécifiques. *Journal de la Société Française de Statistique*, 149(2): 75-97.

Wold, S., Kettaneh, N. and Tjessem, K., (1996). Hierarchical multi-block PLS and PC models for easier interpretation and as an alternative to variable selection. Journal of Chemometrics. (10), 463-482

Westerhuis, J. A, Kourti, T., Macgregor J. F., (1998). Analysis of Multiblock and Hierarchical PCA and PLS Models. Journal of Chemometrics, (12), pp. 301-321.

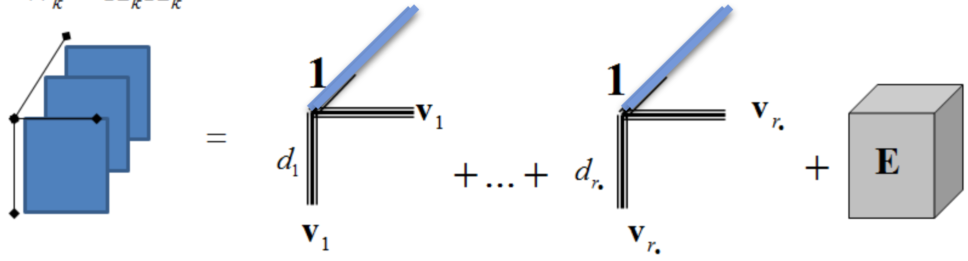
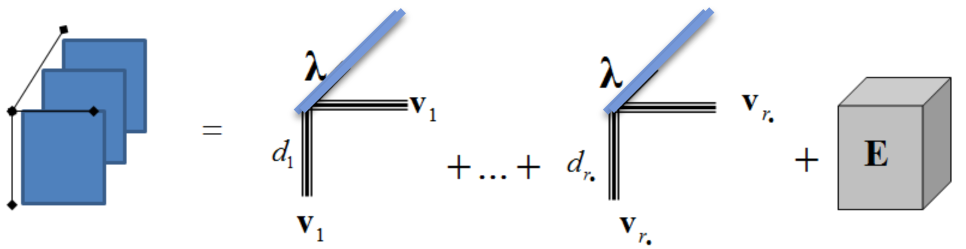
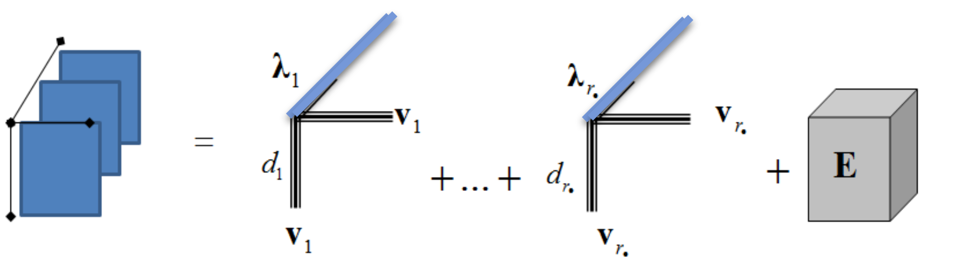
Tensor decomposition

PCA of the total block



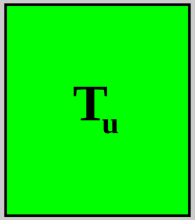
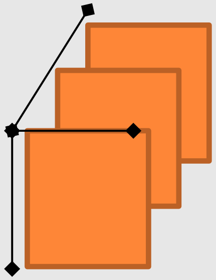
Tensor decomposition = a decomposition of a data by dimension with respect to the three ways.

Two ideas behind methods

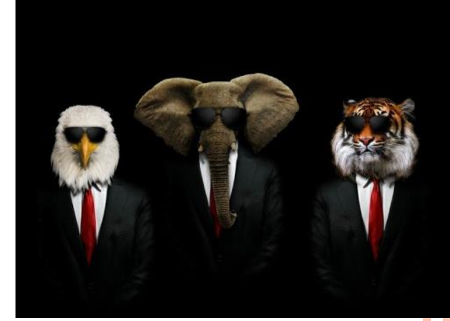
methods	representation	Criteria
PCA	$\mathbf{W}_k = \mathbf{X}_k \mathbf{X}_k^T$ 	$\sum_{k=1}^{K} \left\| \mathbf{W}_k - 1.\mathbf{W} \right\|^2$ $\mathbf{W} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^T$
STATIS		$\sum_{k=1}^{K} \left\| \mathbf{W}_k - \lambda_k \mathbf{W} \right\|^2$ $\mathbf{W} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^T$
CCSWA		$\sum_{k=1}^{K} \left\| \mathbf{W}_k - \mathbf{V} \Lambda_k \mathbf{V}^T \right\|^2$

POLYMORPHISM

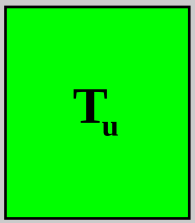
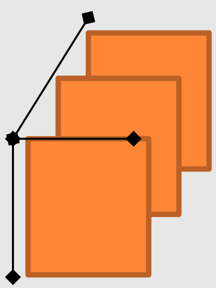
Two different presentations of the same multiblock method (polymorphism)

	CPCA ,MCOA, PCA	STATIS	CCSWA=HPCA
	$\sum_{k=1}^{K} \mathrm{cov}^2(\mathbf{X}_k \mathbf{u}_k, \mathbf{v})$	$\sum_{k=1}^{K} \lambda_k \mathrm{cov}^2(\mathbf{X}_k \mathbf{u}_k, \mathbf{v})$	$\sum_{k=1}^{K} \mathrm{cov}^4(\mathbf{X}_k \mathbf{u}_k, \mathbf{v})$
	$\sum_{k=1}^{K} \left\Vert \mathbf{W}_k - 1.\mathbf{W} \right\Vert^2$	$\sum_{k=1}^{K} \left\Vert \mathbf{W}_k - \lambda_k \mathbf{W} \right\Vert^2$	$\sum_{k=1}^{K} \left\Vert \mathbf{W}_k - \mathbf{V}\Lambda_k \mathbf{V}^T \right\Vert^2$

HETEROGENEITY



Outputs of methods have neither the same aspect nor the same form (heterogeneity)

	CPCA ,MCOA, PCA	STATIS	CCSWA=HPCA
	$\sum_{k=1}^{K} \mathrm{cov}^2\left(\mathbf{X}_k\mathbf{u}_k, \mathbf{v}\right)$	$\sum_{k=1}^{K} \lambda_k \mathrm{cov}^2\left(\mathbf{X}_k\mathbf{u}_k, \mathbf{v}\right)$	$\sum_{k=1}^{K} \mathrm{cov}^4\left(\mathbf{X}_k\mathbf{u}_k, \mathbf{v}\right)$
	$\sum_{k=1}^{K} \left\Vert \mathbf{W}_k - 1.\mathbf{W} \right\Vert^2$	$\sum_{k=1}^{K} \left\Vert \mathbf{W}_k - \lambda_k \mathbf{W} \right\Vert^2$	$\sum_{k=1}^{K} \left\Vert \mathbf{W}_k - \mathbf{V}\Lambda_k\mathbf{V}^T \right\Vert^2$

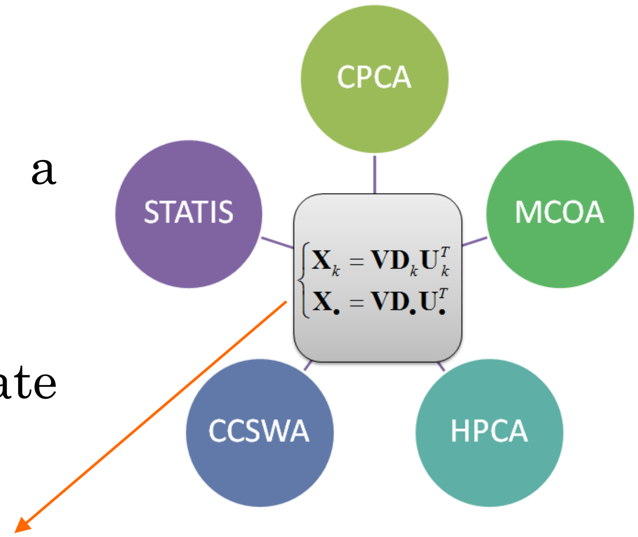
Users are often lost.
Evaluation of methods is difficult.
lack of software that brings together all the methods.

OBJECTIVES

- Bring multiblock methods closer to a simple PCA.

- Understand, implement and evaluate easily multiblock methods.

- Canonical representation of a multiblock method means a representation of its parameters in a way close to the representation of the parameters of a usual PCA (instinctive and natural)



Canonical representation

- Multibloc analysis = a double analysis
- Global analysis described by :

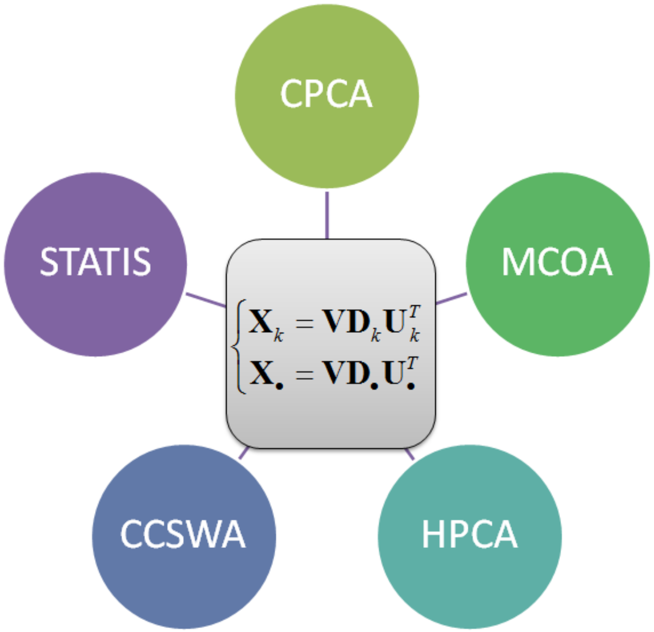
$$\mathbf{X}_{\bullet} = \mathbf{V}\mathbf{D}_{\bullet}\mathbf{U}_{\bullet}^{T}$$

- Block analysis described by :

$$\mathbf{X}_{k} = \mathbf{V}\mathbf{D}_{k}\mathbf{U}_{k}^{T}$$

- Properties

  - $\left\|\mathbf{X}_{\bullet}\right\|^{2} = \sum_{k=1}^{K}\left\|\mathbf{X}_{k}\right\|^{2}$
  - $\mathbf{D}_{\bullet}^{2} = \sum_{k=1}^{K}\mathbf{D}_{k}^{2}$
  - $d_{\bullet}^{(h)2} = \sum_{k=1}^{K} d_{k}^{(h)2}$
  - $d_{k}^{(h)2} = \sum_{j=1}^{p_k}\operatorname{cov}^2\left(\mathbf{X}_k[:, j], \mathbf{v}^{(h)}\right) = \operatorname{cov}^2\left(\mathbf{X}_k\mathbf{u}_k^{(h)}, \mathbf{v}^{(h)}\right) = \operatorname{cov}\left(\mathbf{X}_k\mathbf{X}_k^{T}\mathbf{v}^{(h)}, \mathbf{v}^{(h)}\right)$

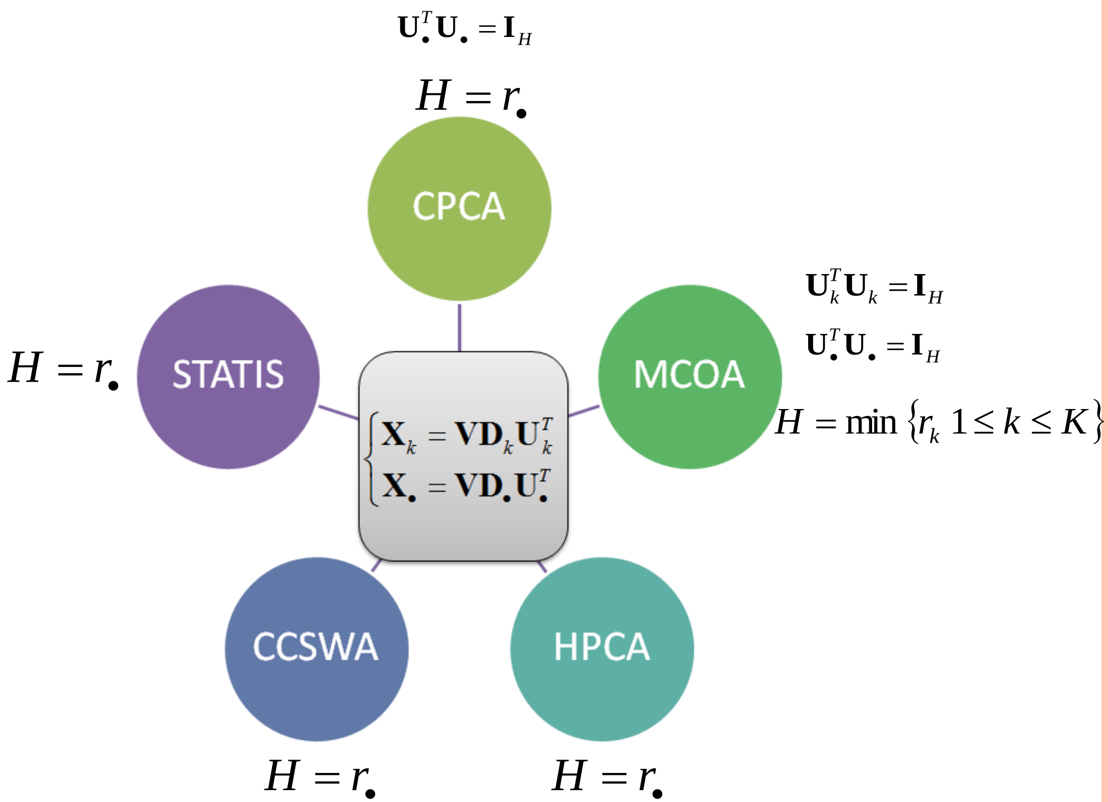


Parameters properties

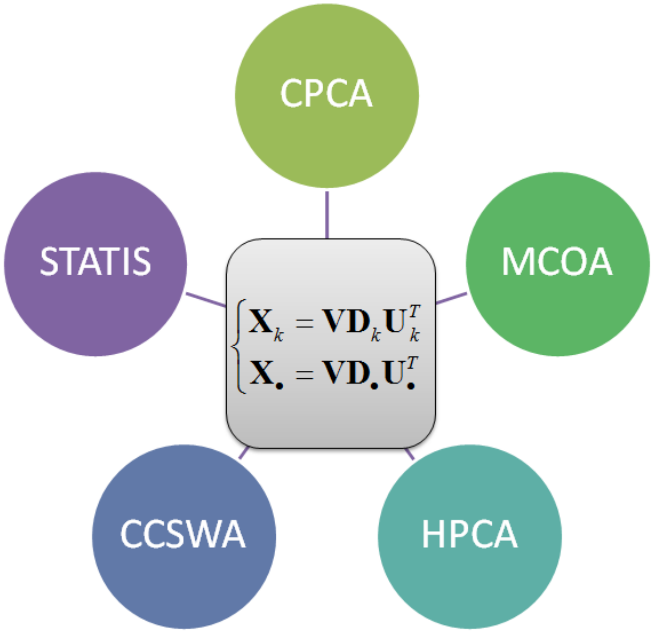
$$\mathbf{X}_k = \mathbf{V}\mathbf{D}_k\mathbf{U}_k^T$$

$$\mathbf{X}_\bullet = \mathbf{V}\mathbf{D}_\bullet\mathbf{U}_\bullet^T$$

- $\mathbf{V}^T\mathbf{V} = I_H$
- $diag(\mathbf{U}_\bullet^T\mathbf{U}_\bullet) = I_H$
- $diag(\mathbf{U}_k^T\mathbf{U}_k) = I_H$
- $\mathbf{D}_\bullet, \mathbf{D}_k \rightarrow diagonals$

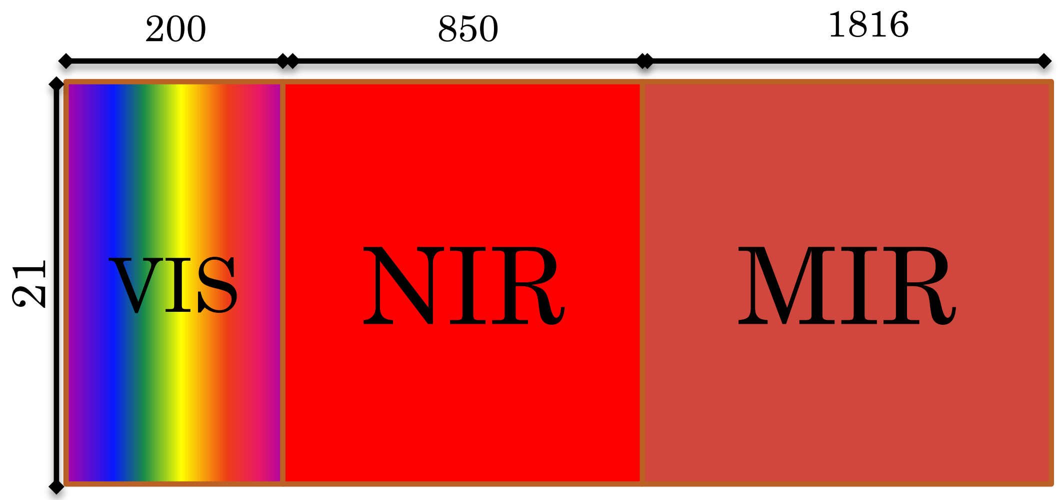


Proofs ?

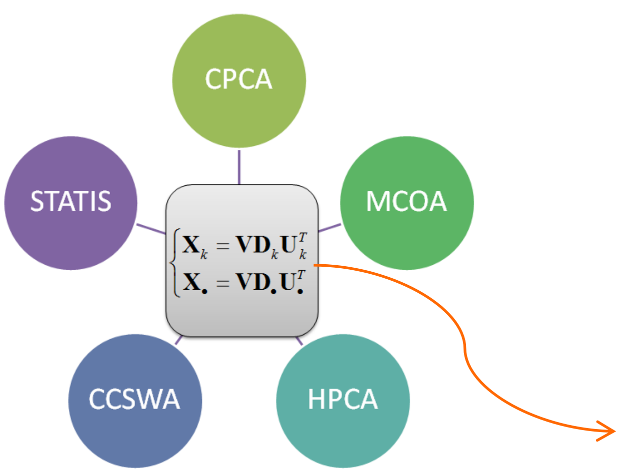


Deflation strategies of methods + matrix operations

Comparaison of methods



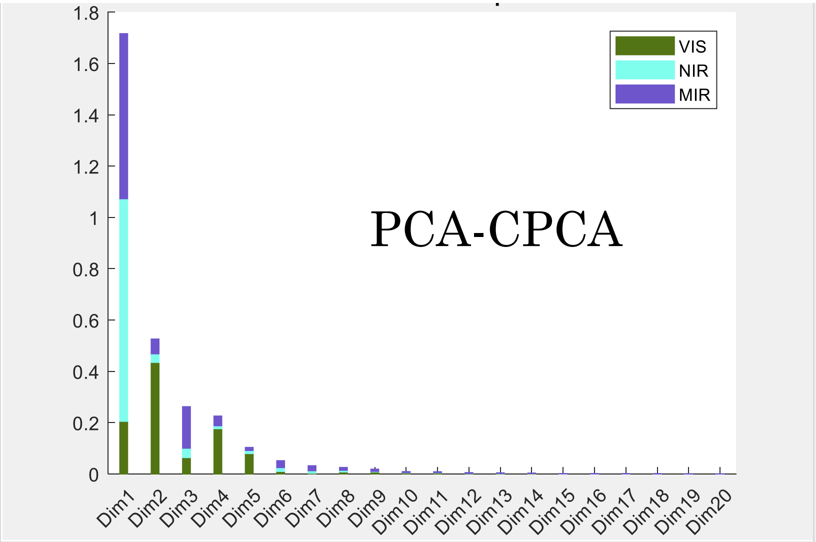
Example with 3 blocks

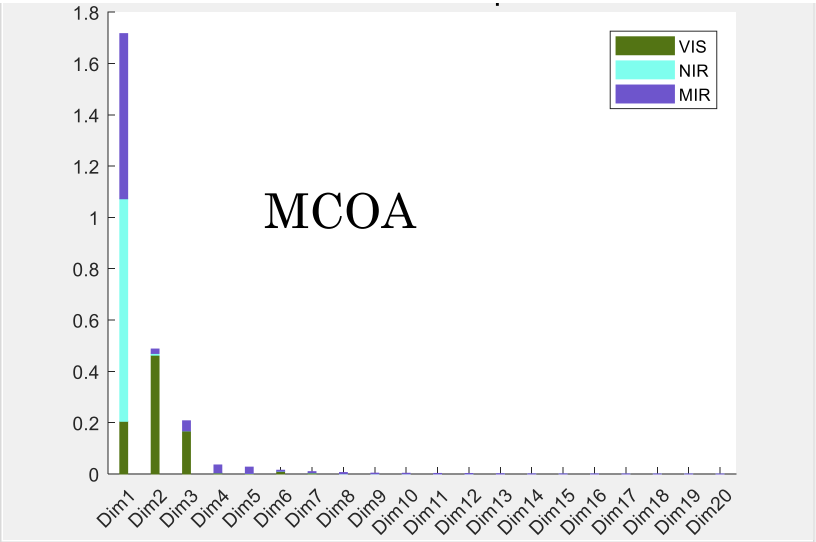


$$\mathbf{V}^{(method)} \quad method = \{PCA, MCOA, STATIS, CCSWA\}$$

$$\mathbf{D}^2_{block,method} \quad block = \{VIS, NIR, MIR\}$$

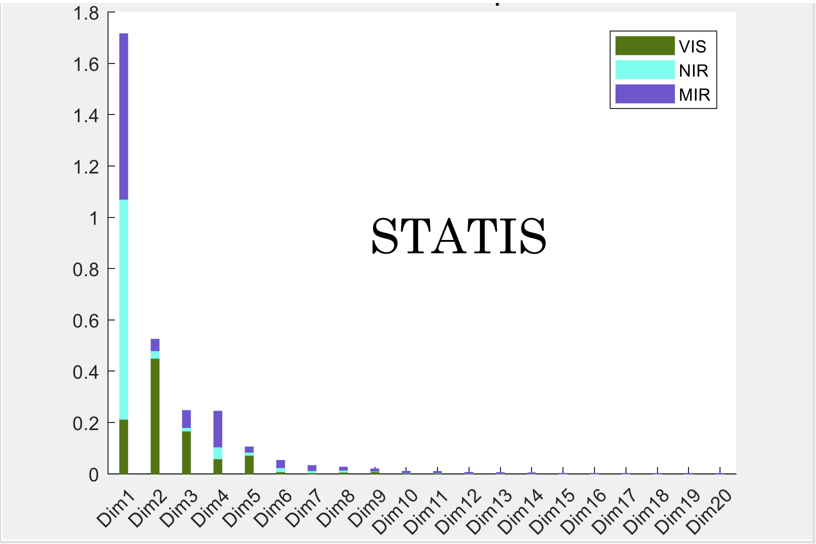
Explained variances by method

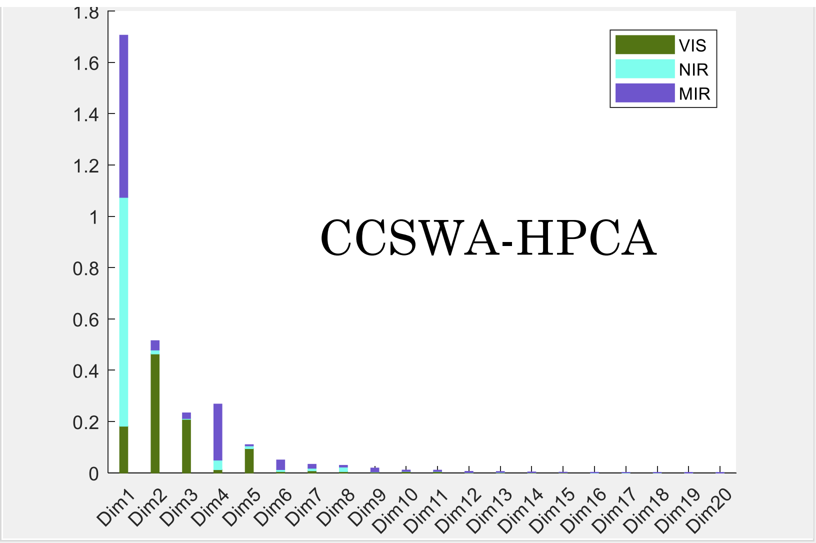




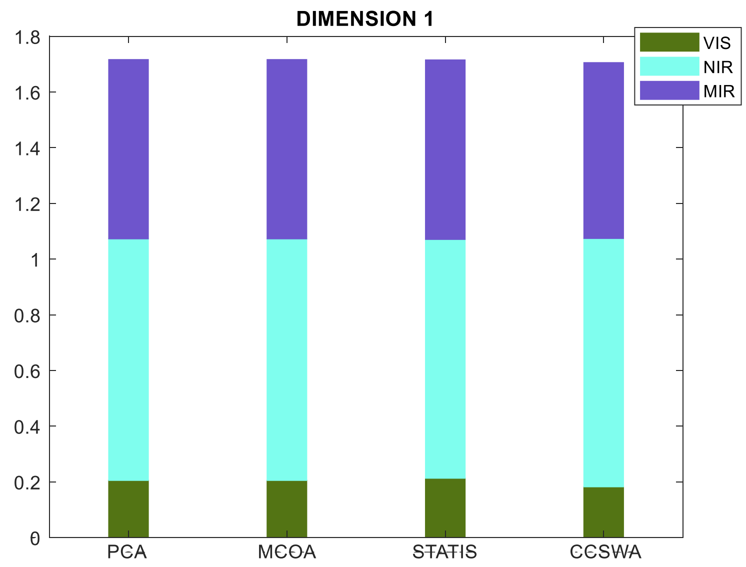
- $d_k^{(h)2} = \sum_{j=1}^{p_k} \mathrm{cov}^2\left(\mathbf{X}_k[:, j], \mathbf{v}^{(h)}\right) = \mathrm{cov}^2\left(\mathbf{X}_k \mathbf{u}_k^{(h)}, \mathbf{v}^{(h)}\right) = \mathrm{cov}\left(\mathbf{X}_k \mathbf{X}_k^T \mathbf{v}^{(h)}, \mathbf{v}^{(h)}\right)$

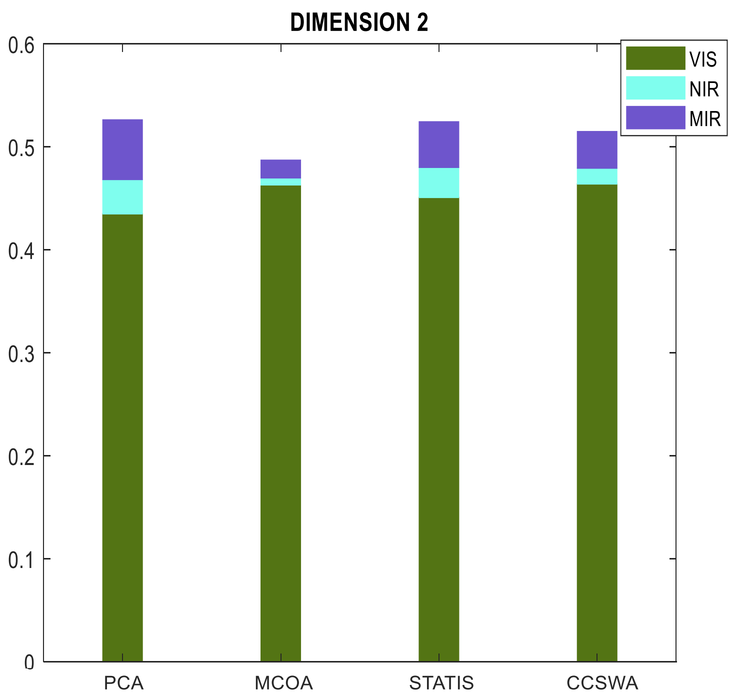
$$d_{\bullet}^{(h)2} = \sum_{k=1}^{K} d_k^{(h)2}$$





Explained variances by dimension



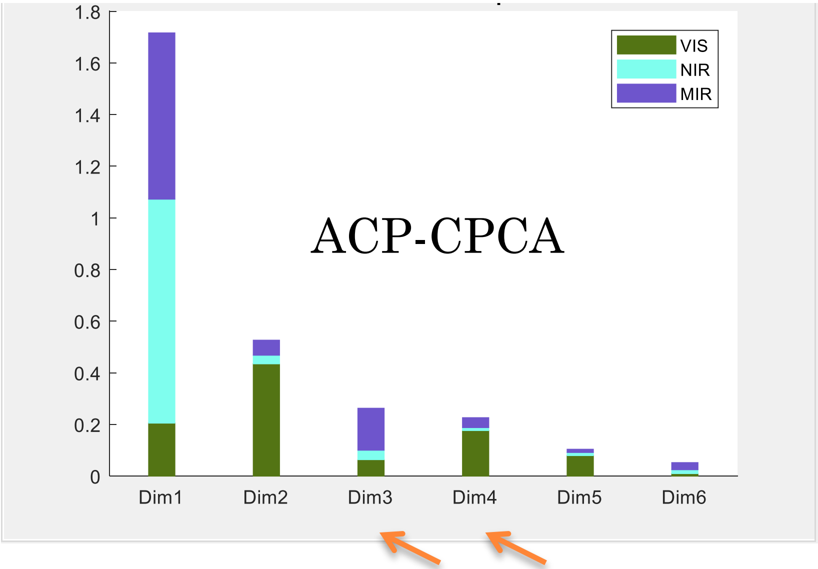


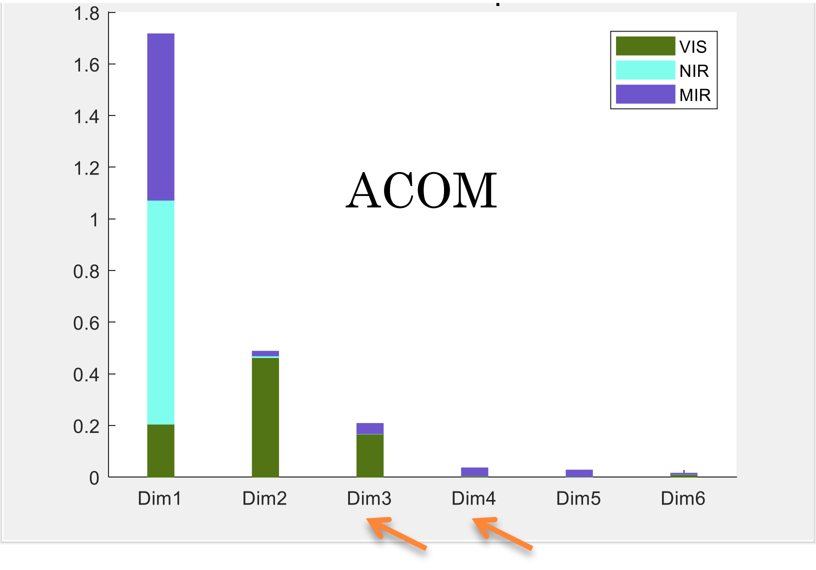
DIMENSION 1				
	ACP	ACOM	STATIS	ACCPS
ACP	1.000	**1.000**	0.996	0.996
ACOM		1.000	0.996	0.996
STATIS			1.000	0.993
ACCPS				1.000

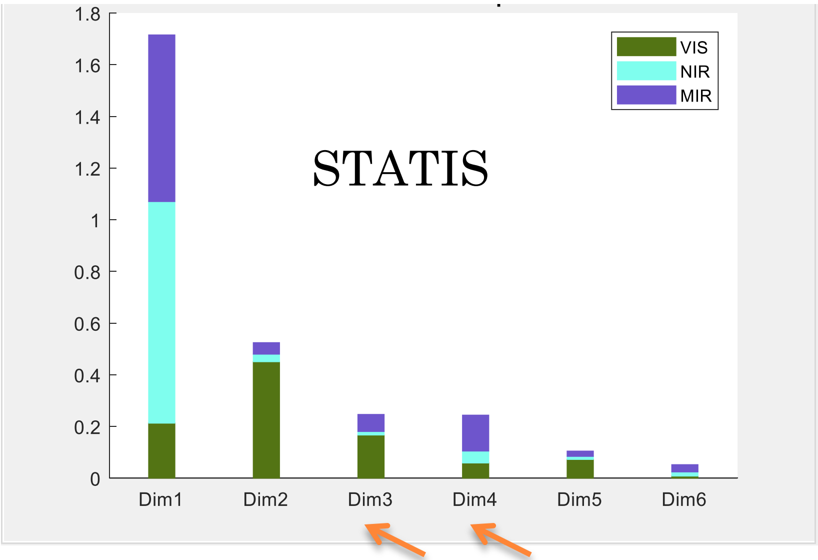
Correlation matrix between parameters

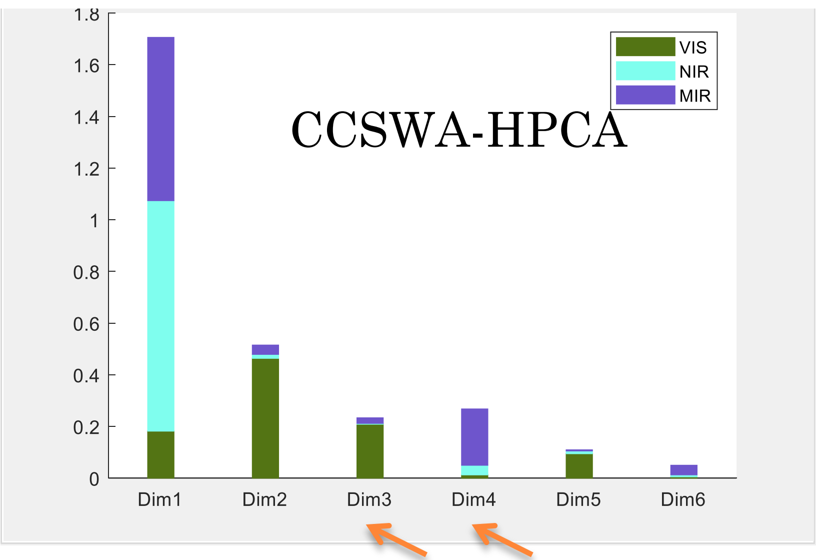
$$\mathbf{V}^{(method)} \quad method = \{PCA, MCOA, STATIS, CCSWA\}$$

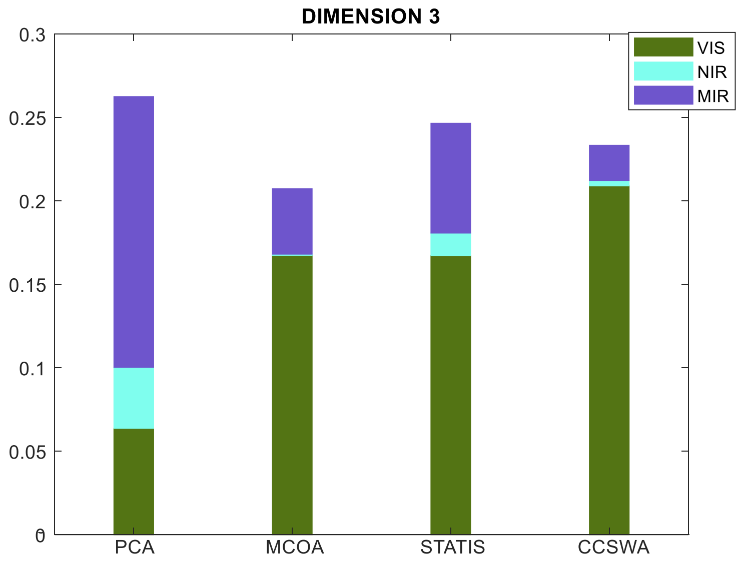
	DIMENSION 2			
	ACP	ACOM	STATIS	ACCPS
ACP	1.000	0.978	0.996	0.997
ACOM		1.000	0.991	0.991
STATIS			1.000	0.992
ACCPS				1.000


ACP-CPCA
VIS
NIR
MIR
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
Dim1
Dim2
Dim3
Dim4
Dim5
Dim6


ACOM
VIS
NIR
MIR
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
Dim1
Dim2
Dim3
Dim4
Dim5
Dim6


STATIS
VIS
NIR
MIR
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
Dim1
Dim2
Dim3
Dim4
Dim5
Dim6

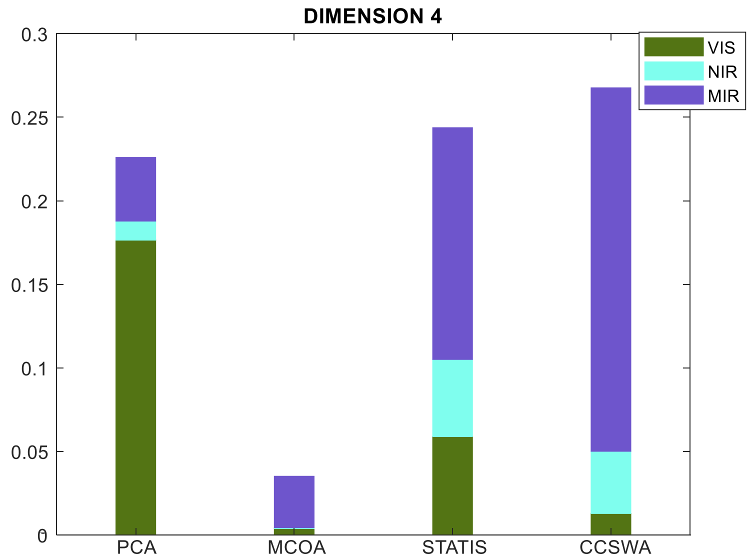

CCSWA-HPCA
VIS
NIR
MIR
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
Dim1
Dim2
Dim3
Dim4
Dim5
Dim6



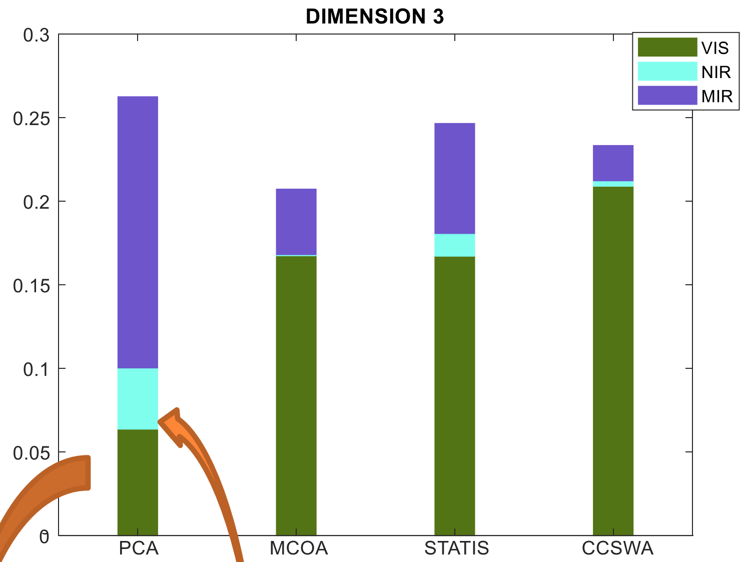
	DIMENSION 3			
	ACP	ACOM	STATIS	ACCPS
ACP	1.000	0.583	0.737	0.438
ACOM		1.000	0.993	0.925
STATIS			1.000	0.992
ACCPS				1.000

Correlation matrix between parameters

$$\mathbf{V}^{(method)} \quad method = \{PCA, MCOA, STATIS, CCSWA\}$$



	DIMENSION 4			
	ACP	ACOM	STATIS	ACCPS
ACP	1.000	0.064	0.735	0.427
ACOM		1.000	0.108	0.108
STATIS			1.000	0.913
ACCPS				1.000

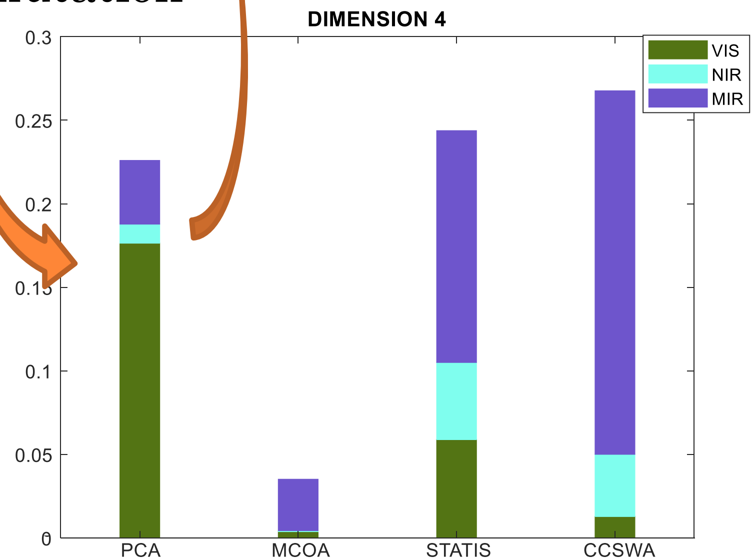


	DIMENSION 3			
	ACP	ACOM	STATIS	ACCPS
ACP	1.000	0.583	0.737	0.438
ACOM		1.000	0.993	0.925
STATIS			1.000	0.992
ACCPS				1.000

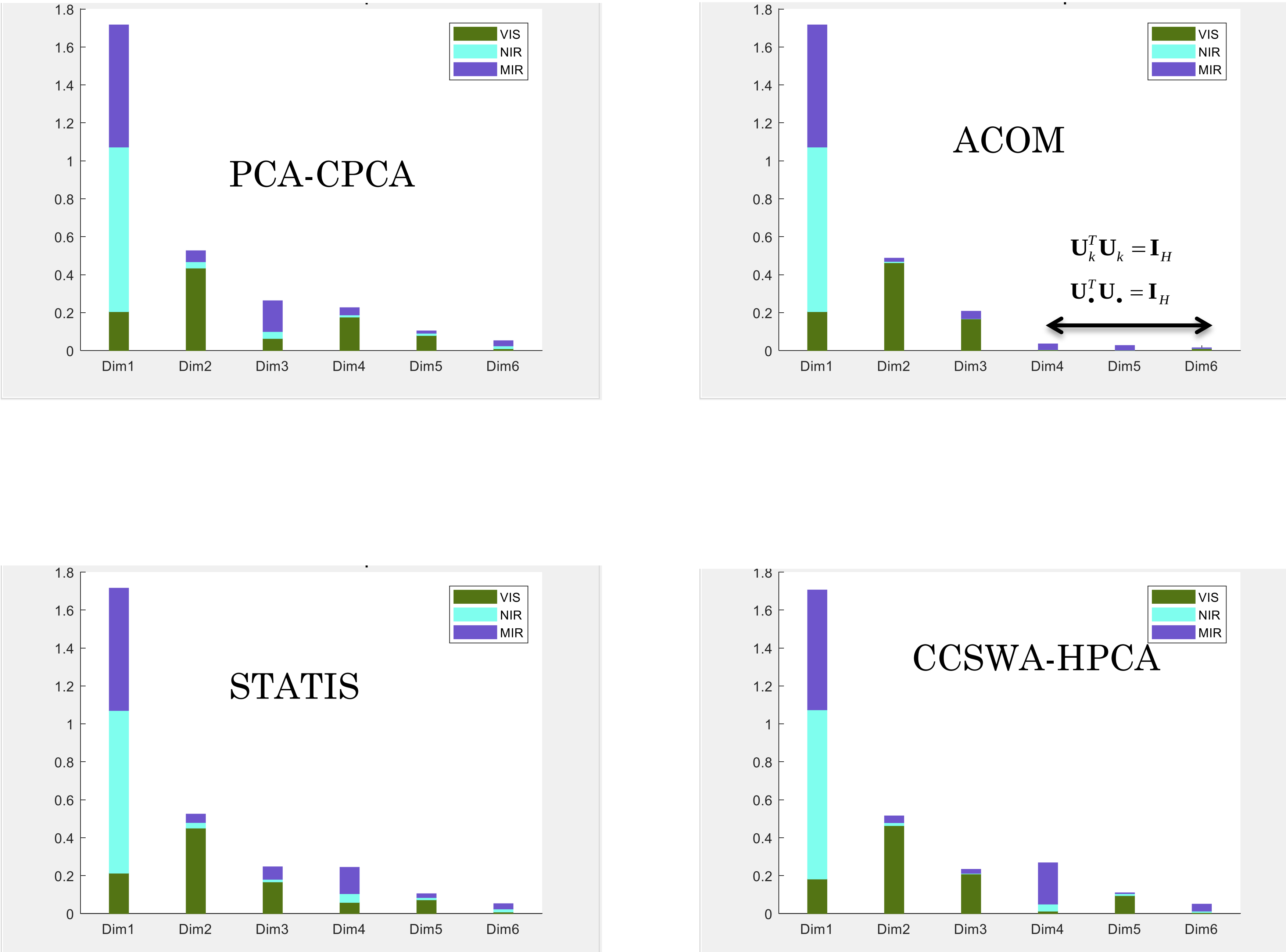
Correlation matrix between parameters

$$\mathbf{V}^{(method)} \qquad method = \{PCA, MCOA, STATIS, CCSWA\}$$

Permutation



	DIMENSION 4			
	ACP	ACOM	STATIS	ACCPS
ACP	1.000	0.064	0.735	0.427
ACOM		1.000	0.108	0.108
STATIS			1.000	0.913
ACCPS				1.000


PCA-CPCA
VIS
NIR
MIR
Dim1
Dim2
Dim3
Dim4
Dim5
Dim6
ACOM
$\mathbf{U}_k^T\mathbf{U}_k = \mathbf{I}_H$
$\mathbf{U}_\bullet^T\mathbf{U}_\bullet = \mathbf{I}_H$
STATIS
CCSWA-HPCA

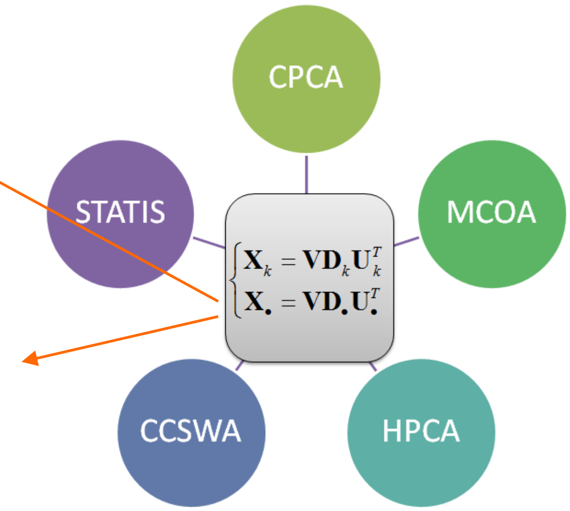
Remarks

- Results of methods « always » converge
  - Several comparaisons on real data sets give the same conclusions.
  - Several compraisons on random data sets give the same conclusions.
- Main perspective : Need a proof …

Conclusions

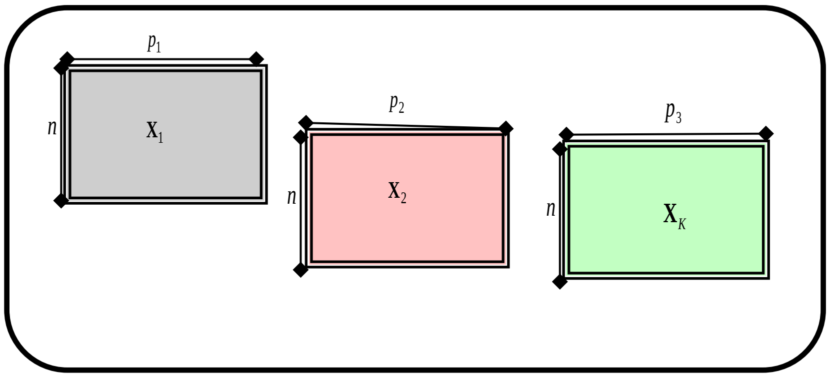
- Do multiblock method as a simple PCA.

- Implement and evaluate easily multiblock methods ➔

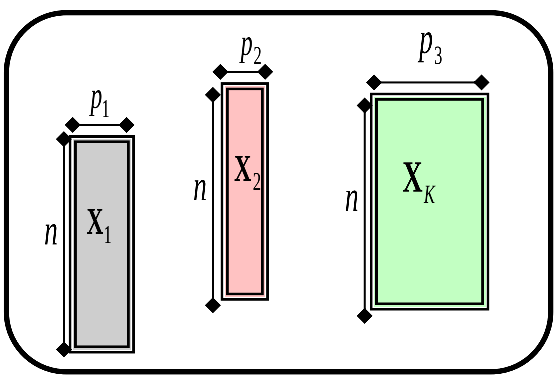


- Multiblock software (matlab and R) which brings together methods based on canonical representation of methods.

COMPUTATIONAL INTEREST OF POLYMORPHISM (CCSWA CASE)



$n < p_k$



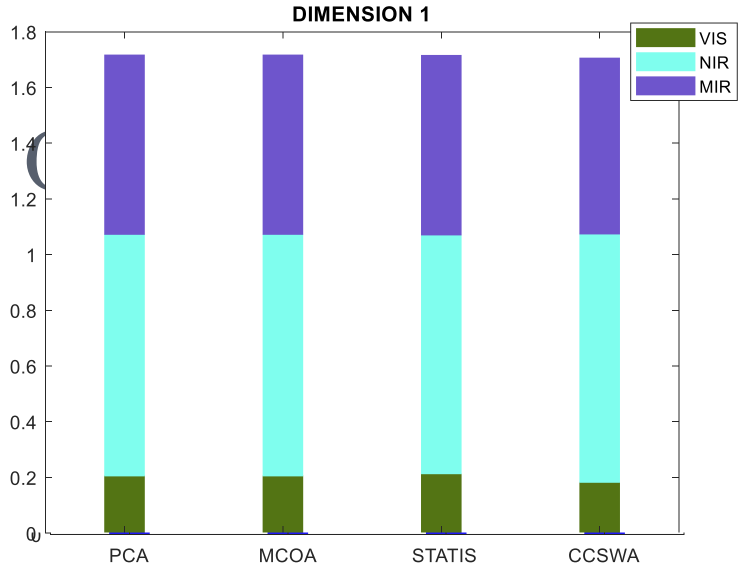
$n > p_k$

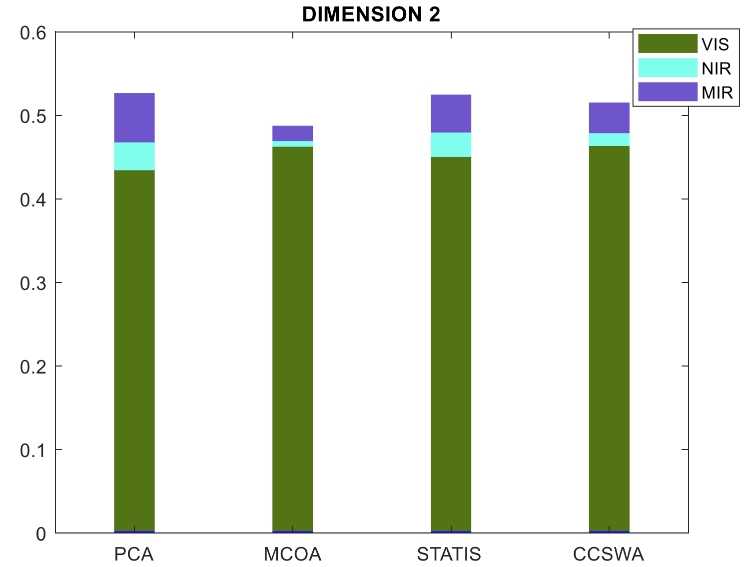
$$\sum_{k=1}^{K}\left\|\mathbf{W}_k-\lambda_k\mathbf{v}\mathbf{v}^T\right\|^2$$

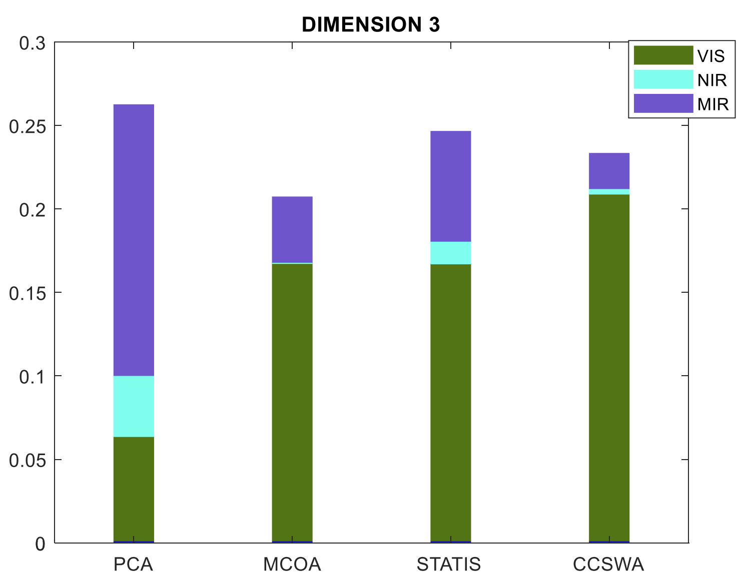
- $w_1, w_2, \ldots, w_m$
- $\mathbf{W}=\sum_{k=1}^{m} w_k\mathbf{X}_k\mathbf{X}_k'$
- *Extract* the first eigenvector $\mathbf{v}$ of $\underbrace{\mathbf{W}}_{(n,n)}$
- $w_k=\mathbf{v}_k'\mathbf{X}_k\mathbf{X}_k'\mathbf{v} \quad (1\leq k\leq m)$

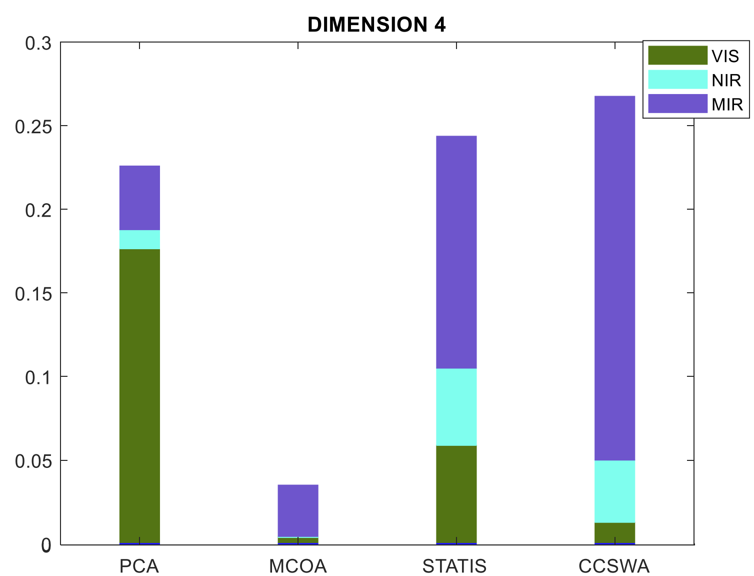
$$\sum_{k=1}^{K}\text{cov}^4\left(\mathbf{X}_k\mathbf{u}_k,\mathbf{v}\right)$$

- $\mathbf{v}=\mathbf{v}/\sqrt{\mathbf{v}^T\mathbf{v}}$
- $\mathbf{u}_k=\mathbf{X}_k^T\mathbf{v}$
- $\mathbf{t}_k\leftarrow\mathbf{X}_k\mathbf{u}_k$
- $\underbrace{\mathbf{T}_\mathbf{u}}_{(n,K)}\leftarrow\left[\mathbf{t}_1 \quad \mathbf{t}_2 \quad \cdots \quad \mathbf{t}_K\right]$
- *Extract* the first principal compoent $\mathbf{v}$ of $\mathbf{T}_\mathbf{u}$


DIMENSION 1
VIS
NIR
MIR
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
PCA
MCOA
STATIS
CCSWA


DIMENSION 2
VIS
NIR
MIR
0.6
0.5
0.4
0.3
0.2
0.1
0
PCA
MCOA
STATIS
CCSWA


DIMENSION 3
VIS
NIR
MIR
0.3
0.25
0.2
0.15
0.1
0.05
0
PCA
MCOA
STATIS
CCSWA


DIMENSION 4
VIS
NIR
MIR
0.3
0.25
0.2
0.15
0.1
0.05
0
PCA
MCOA
STATIS
CCSWA