



Méthodes récentes d'analyse de données -omiques

Douglas N. Rutledge

douglas.neil.rutledge@gmail.com

Muséum National d'Histoire Naturelle, 75005, Paris, France

Faculté de Pharmacie, Université Paris-Saclay, 91400 Orsay, France

université
PARIS-SACLAY



MUSÉUM
NATIONAL D'HISTOIRE NATURELLE

Deux méthodes particulières d'analyse de données multivariées

- **Analyse en Composantes Indépendantes**
- **ComDim**

Des Données Multivariées

Variables

Vecteur Colonne

Objets
Individus

$$\mathbf{X} = \begin{array}{c|ccc} & x_{11} & x_{1j} & \dots & x_{1m} \\ \hline & \dots & \dots & \dots & x_{im} \\ & \dots & \dots & x_{ij} & \dots \\ \hline & x_{n1} & x_{nj} & \dots & x_{nm} \end{array}$$

$$\mathbf{x}_1^T = \begin{array}{c} x_{11} \\ \dots \\ x_{i1} \\ x_{n1} \end{array}$$

Vecteur Ligne

$$\mathbf{x}_1 = x_{11} \quad \dots \quad x_{1j} \quad x_{1m}$$

Analyse en Composantes Principales (PCA)

Pearson, 1901

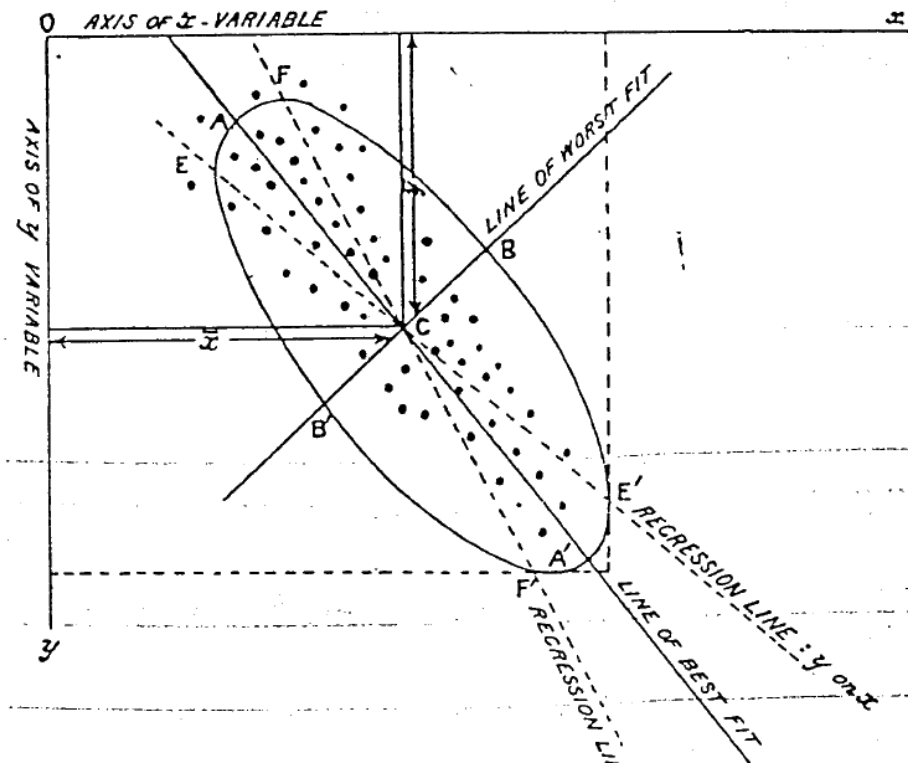
LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* By KARL PEARSON, F.R.S., University College, London*.

(1) IN many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane. Analytically this consists in taking

$$y = a_0 + a_1x, \text{ or } z = a_0 + a_1x + b_1y,$$

$$\text{or } z = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \dots + a_nx_n,$$

where $y, x, z, x_1, x_2, \dots, x_n$ are variables, and determining the "best" values for the constants $a_0, a_1, b_1, a_0, a_1, a_2, a_3, \dots, a_n$ in relation to the observed corresponding values of the variables. In nearly all the cases dealt with in the text-books of least squares, the variables on the right of our equations are treated as the independent, those on the left as the de-



Une matrice vue comme des points dans l'espace

$\mathbf{X} =$

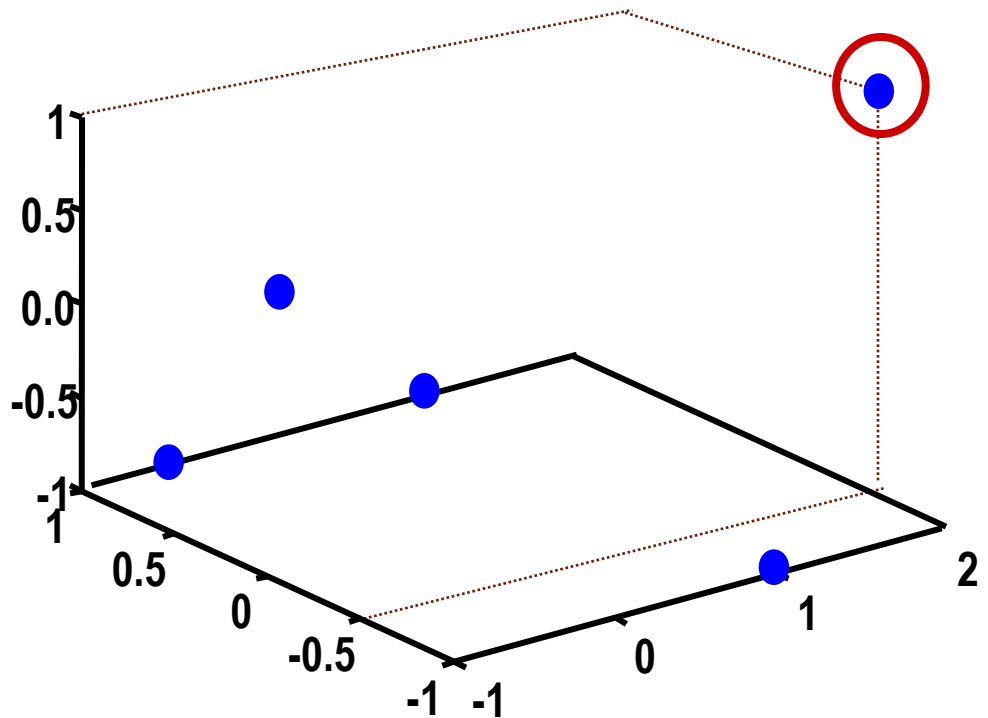
-1.00 0.00 0.50

1.00 -1.00 -1.00

-0.50 1.00 -1.00

1.00 1.00 -1.00

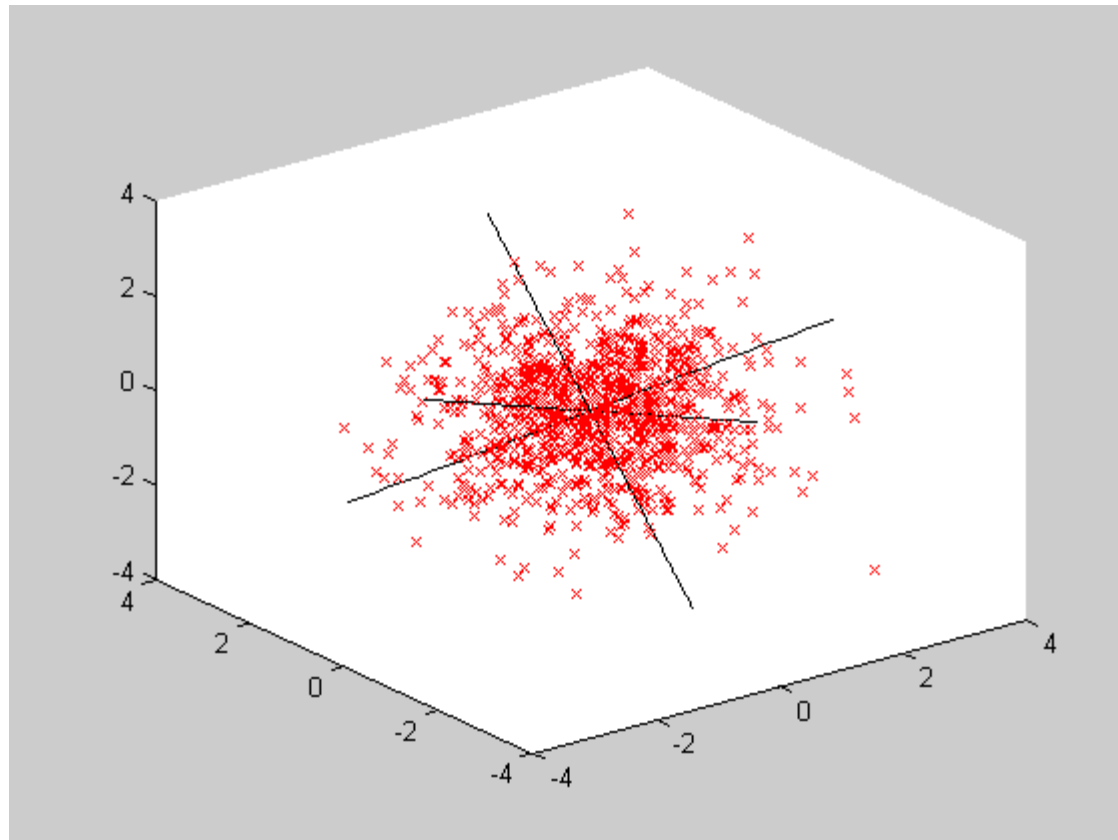
2.00 -0.50 1.00



Analyse en Composantes Principales

Objets dans une espace défini par les variables

- si **AUCUNE information** → distribution sphérique
- **AUCUN** axe *préfér*é



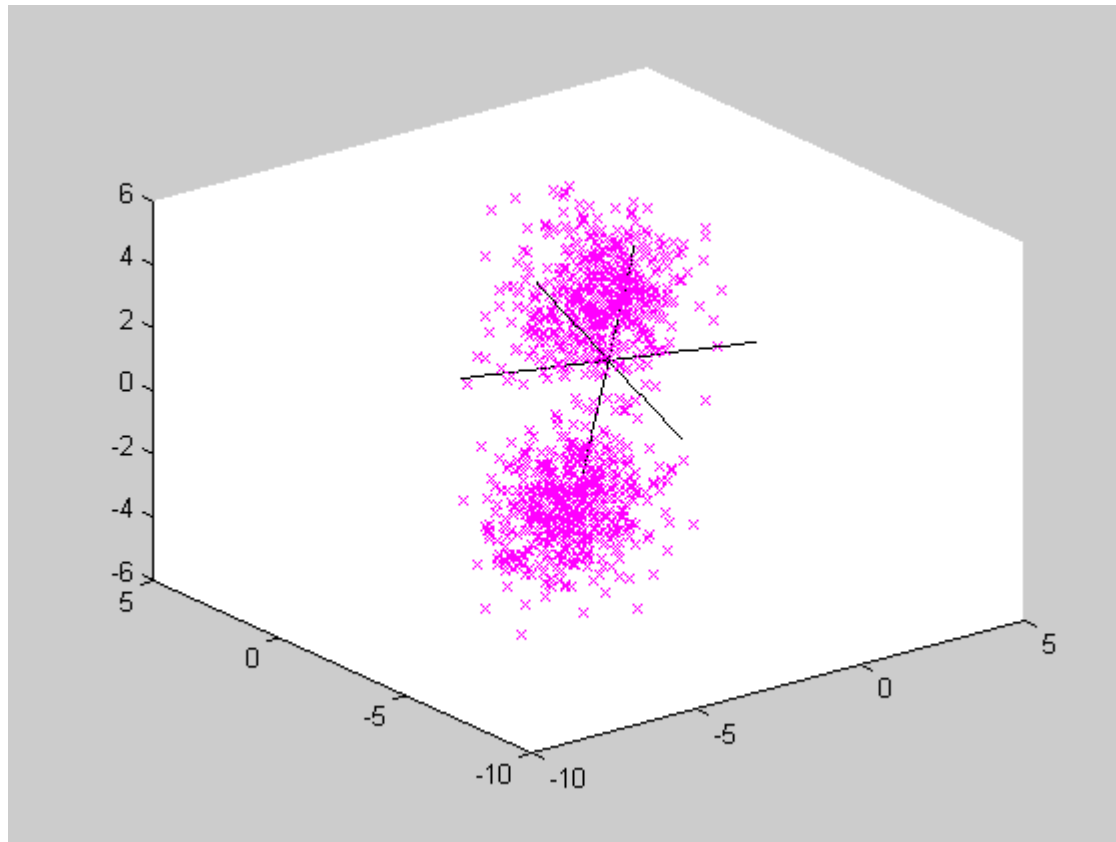
Analyse en Composantes Principales

Objets dans un espace défini par les variables

- si *information*

→ distribution non-sphérique

→ *des axes préférés*



Analyse en Composantes Indépendantes (ICA)

Variables

Vecteurs Ligne

Signaux
Observés

$\mathbf{X} =$	x_{11}	x_{1j}	$\dots$	x_{1m}
	$\dots$	$\dots$	$\dots$	x_{im}
	$\dots$	$\dots$	x_{ij}	$\dots$
	x_{n1}	x_{nj}	$\dots$	x_{nm}

$$\mathbf{X} = \begin{matrix} \mathbf{x}_1^T \\ \dots \\ \mathbf{x}_i^T \\ \dots \\ \mathbf{x}_n^T \end{matrix}$$

Analyse en Composantes Indépendantes

Vise à extraire des signaux sources inconnus
à partir des signaux observés.

Ces signaux observés étant des mélanges de sources
en proportions inconnues.

ICs ou Signaux Sources : analogues des “Loadings” de PCA

Proportions : analogues des “Scores” de PCA

Analyse en Composantes Principales (PCA)

matrice de données $\longrightarrow$ **points dans une espace multivariée**

Analyse en Composantes Indépendantes (ICA)

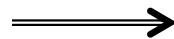
matrice de données $\longrightarrow$ **un jeu de signaux observés, où**

- Chaque **signal observé** est une somme pondérée de **signaux sources**
- Les coefficients de ponderation sont les proportions des **signaux sources**

$$\mathbf{x}_1 = \mathbf{a}_{11} * \mathbf{s}_1 + \mathbf{a}_{12} * \mathbf{s}_2$$

$$\mathbf{x}_2 = \mathbf{a}_{21} * \mathbf{s}_1 + \mathbf{a}_{22} * \mathbf{s}_2$$

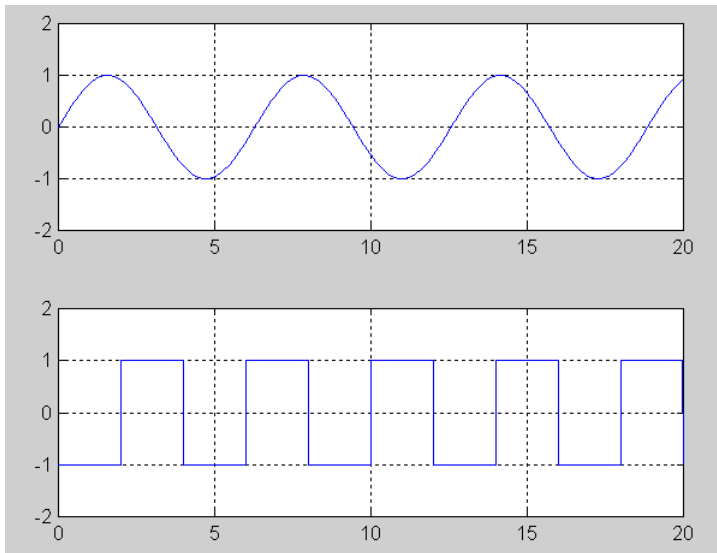
...



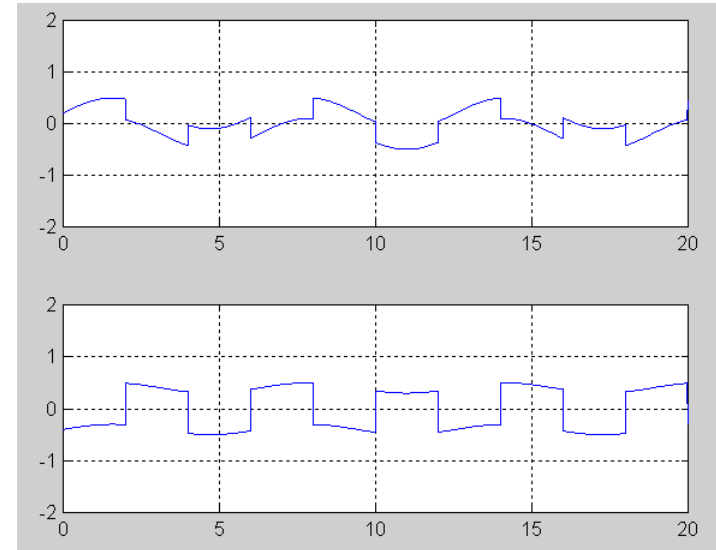
En notation matricielle :

$$\mathbf{X} = \mathbf{A} * \mathbf{S}$$

Exemple



Deux sources indépendantes



Deux mélanges

$$x_1 = a_{11}s_1 + a_{12}s_2$$

$$x_2 = a_{21}s_1 + a_{22}s_2$$

a_{ij} ... Proportion du signal source s_j dans le signal observé x_i

ICA vise à trouver des vecteurs avec “*une signification physique*”

Hypothèses :

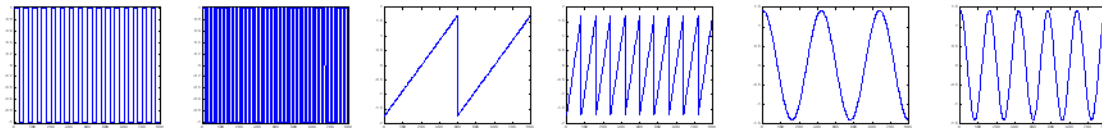
- 1) Aucune raison pour que les variations dans un **signal source** dépende des variations dans un autre **signal source**

Autrement dit, les **signaux sources** sont « indépendants »

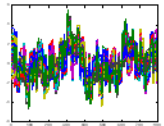
- 2) Les **signaux observés** étant des combinaisons de **sources indépendants**, ils devraient être plus gaussiens que les **sources**
(Théorème Central Limit)

Un exemple simulé

Signaux source

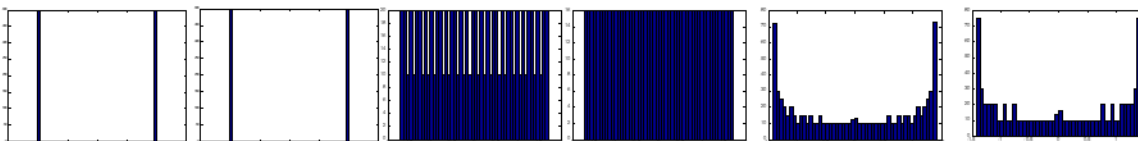


+ Bruit Gaussien

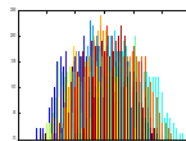


Signaux observés

Histogrammes des sources

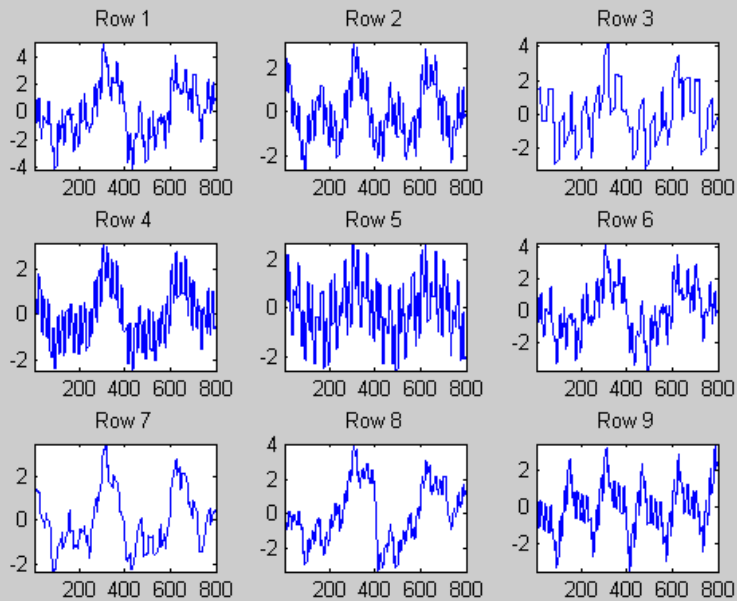


Histogrammes

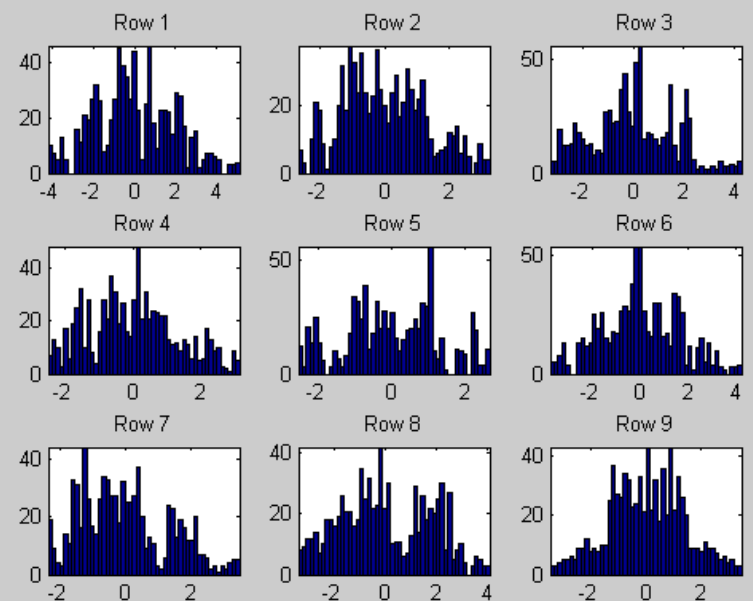


Matrice de départ

Lignes

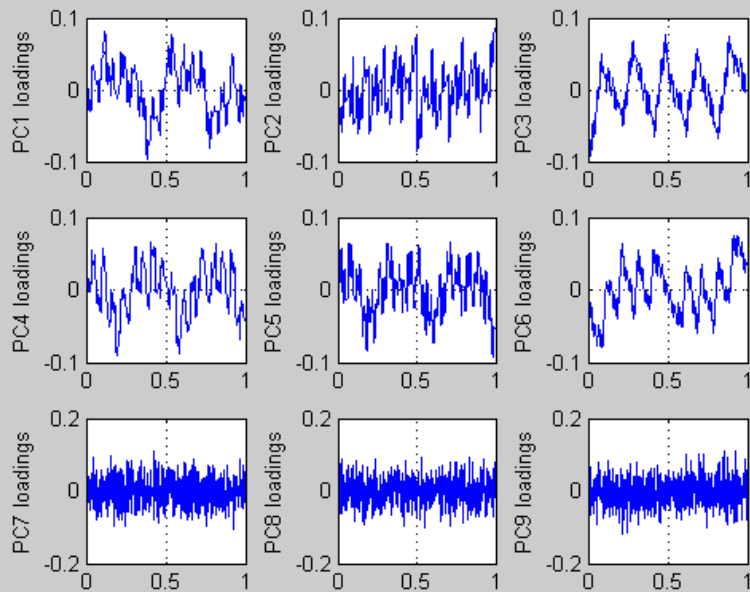


Histogrammes

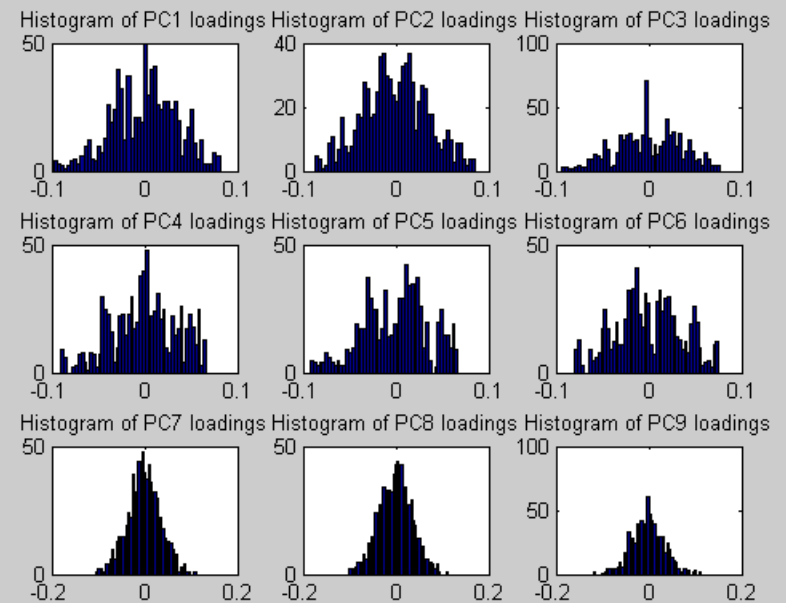


Résultats pour PCA

« Loadings »

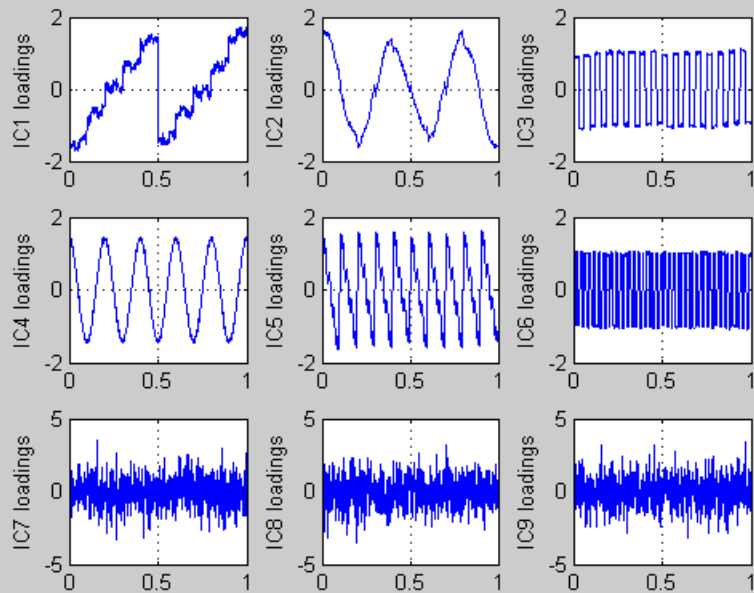


Histogrammes

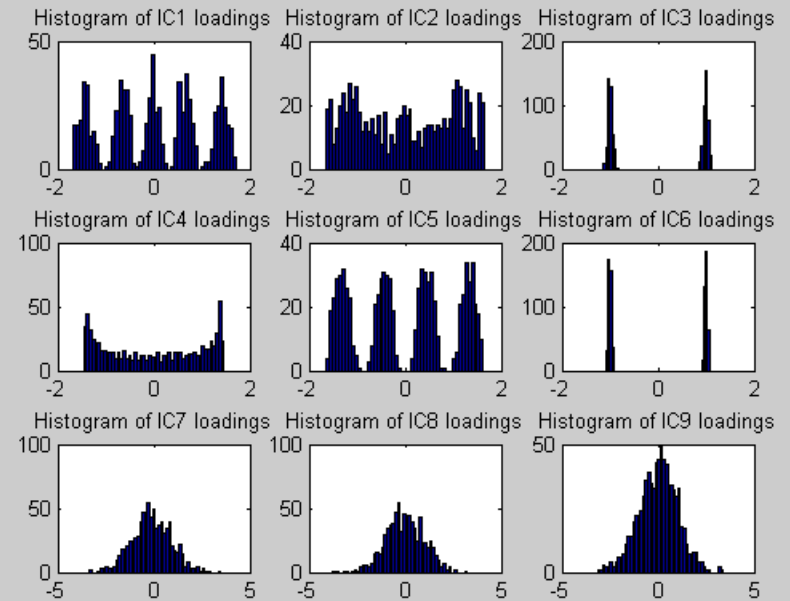


Résultats pour ICA

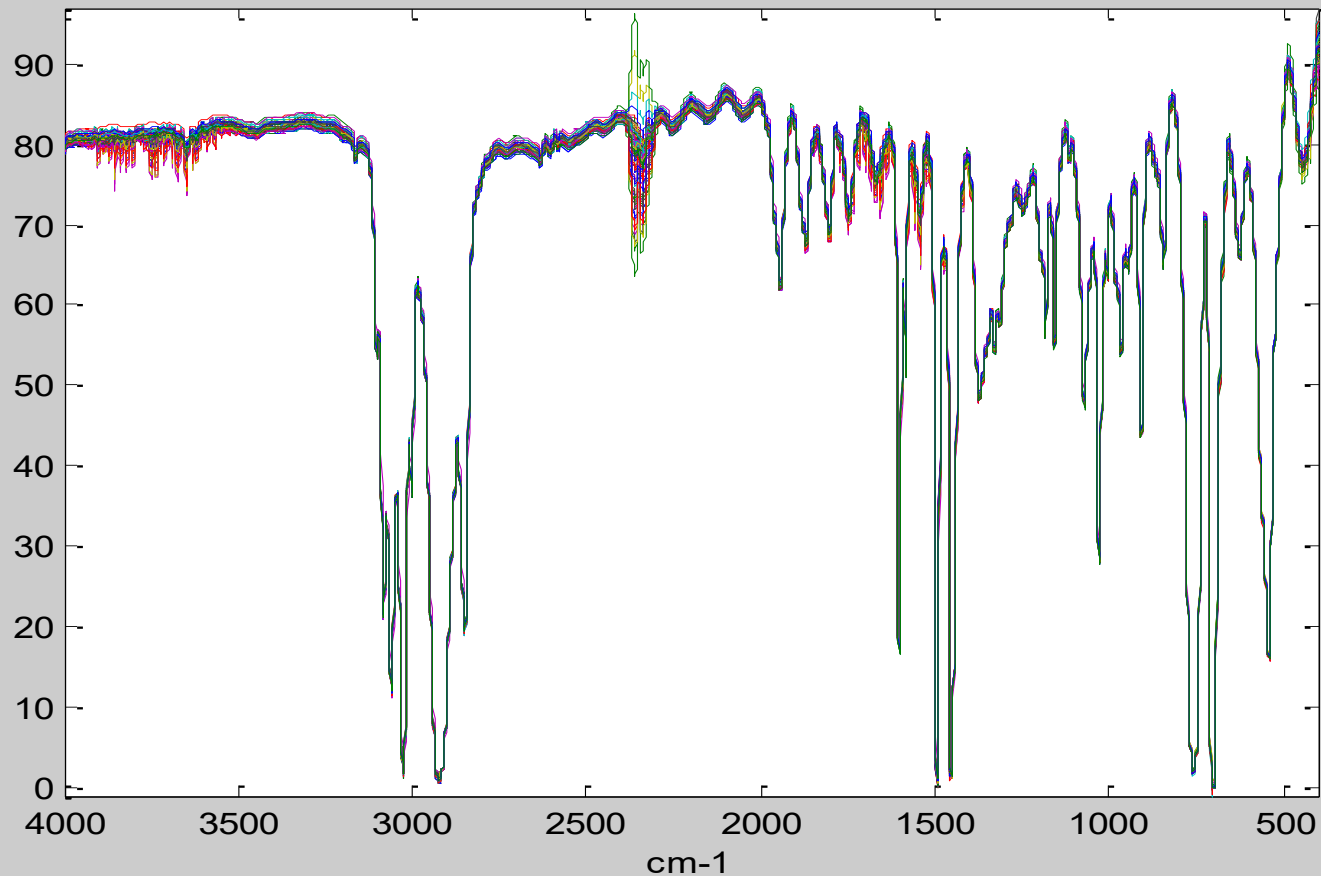
Signaux Sources



Histogrammes



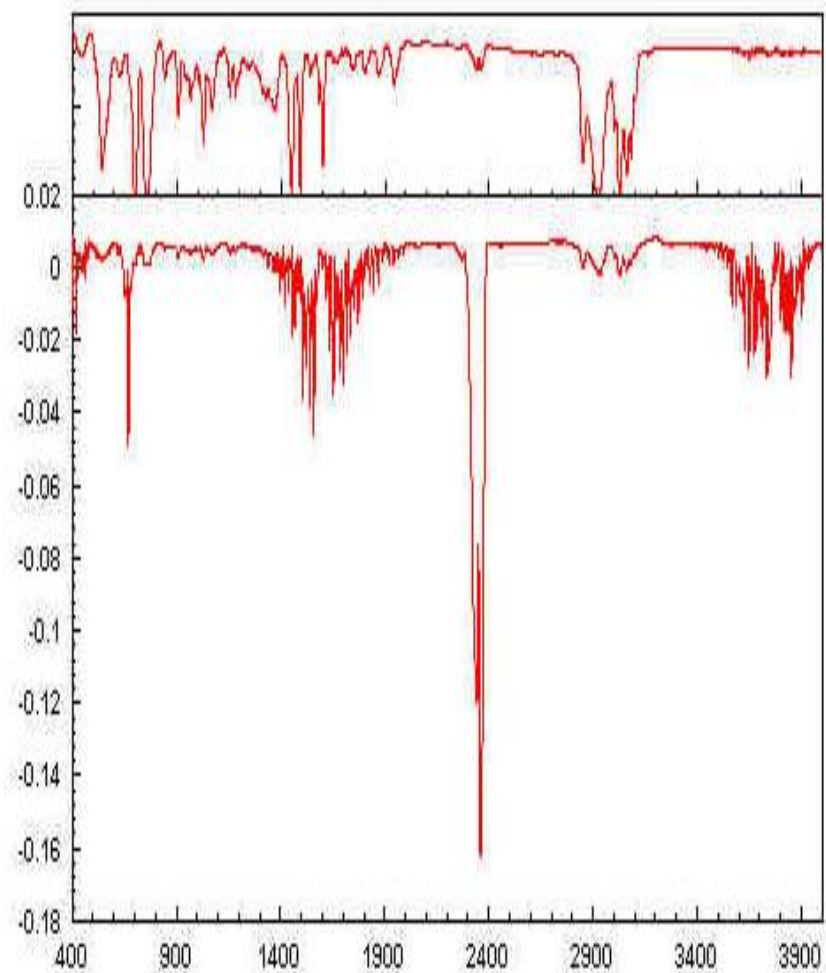
Information cachée dans des spectres



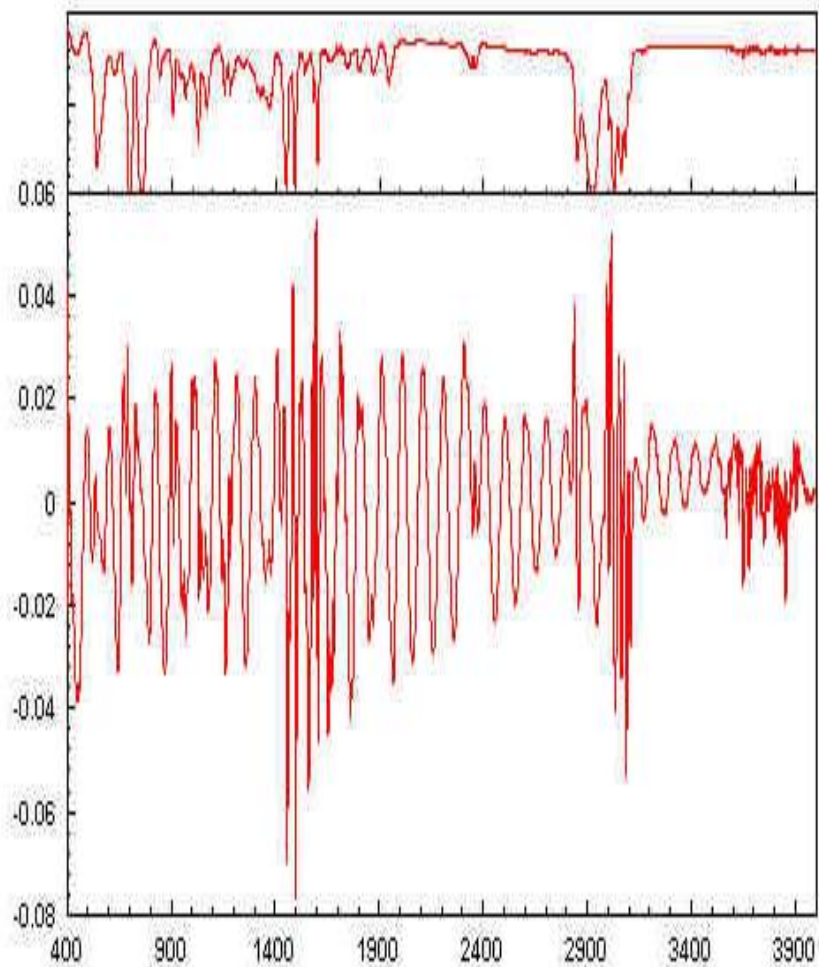
Source : R.Aries, D. Lidiard, R. Spragg, *Spectrosc. Internat.*, 2(3) 41-44

100 spectres Moyen-Infrarouge

- Objets : 100 spectromètres MIR
- Spectres acquis à la fin de la ligne de production
- Entre 4000cm^{-1} et 400cm^{-1} , résolution 1 cm^{-1}

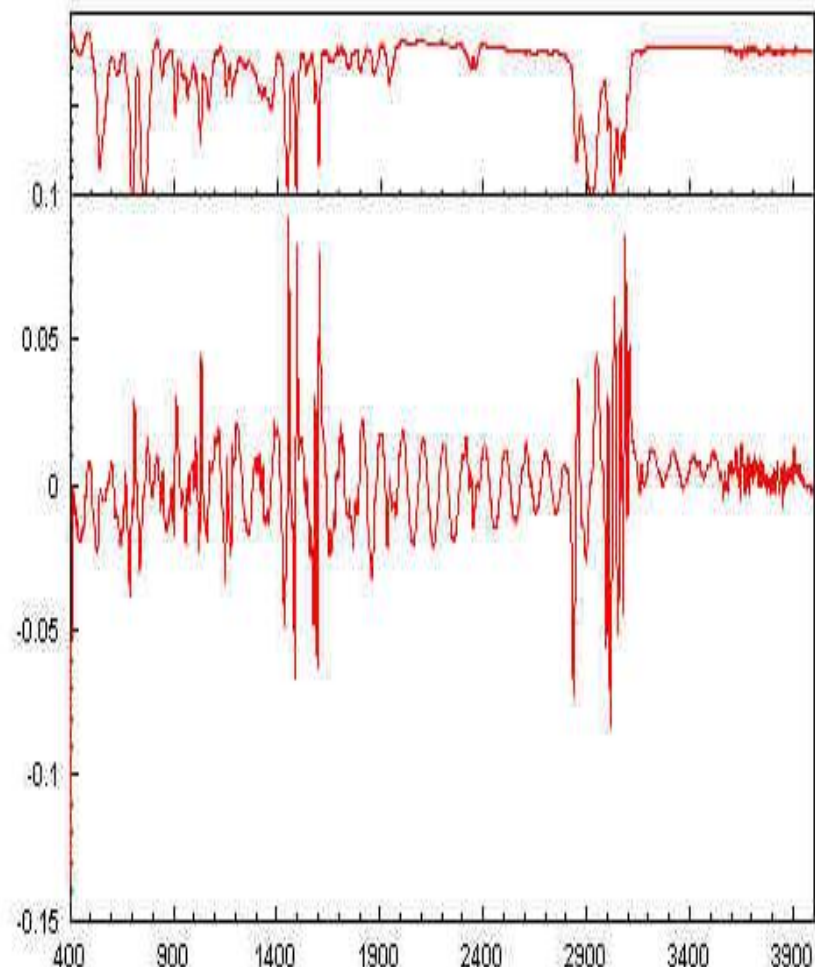


PC2 $\equiv$ CO₂ & H₂O



PC5

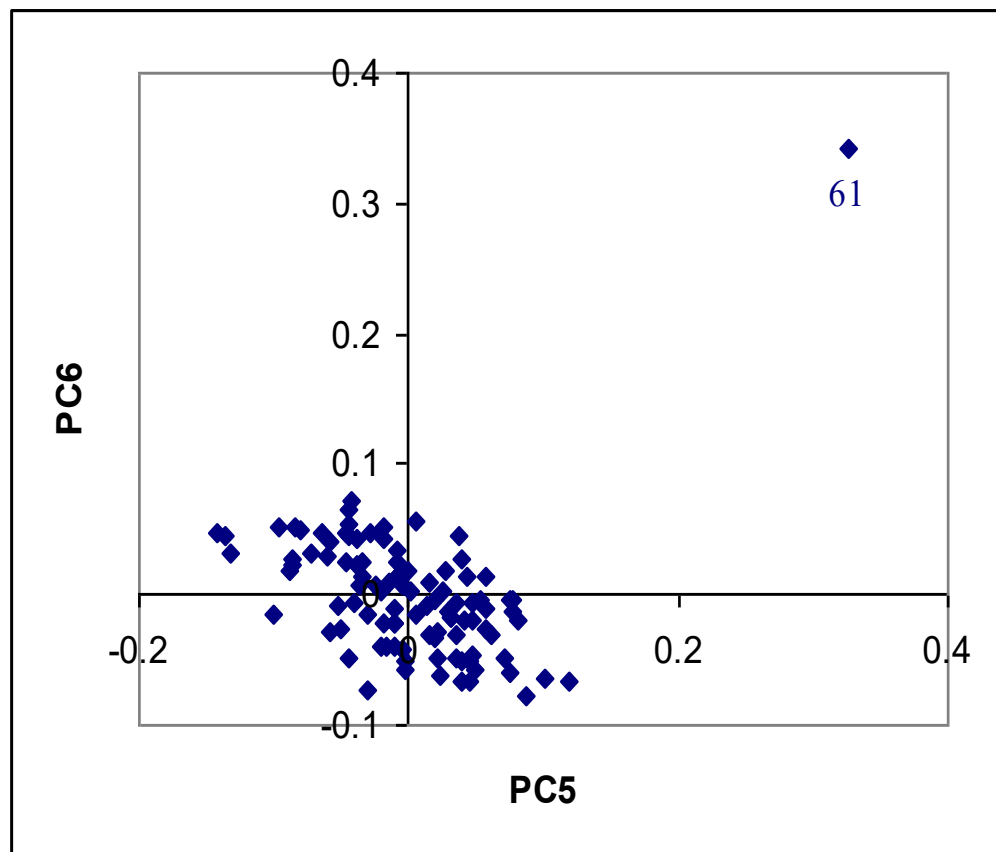
Interférogramme & dérivé

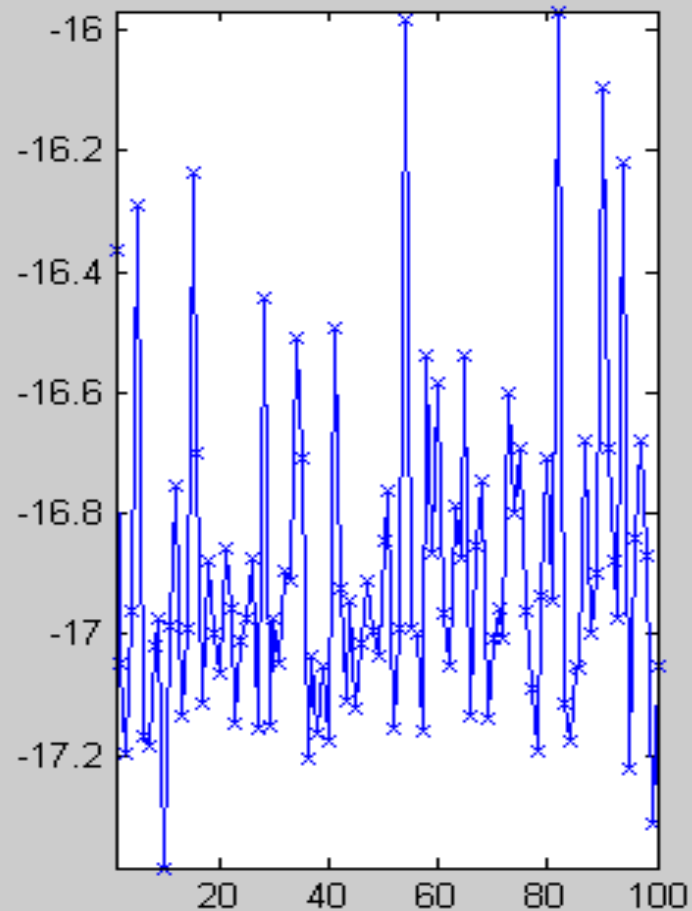
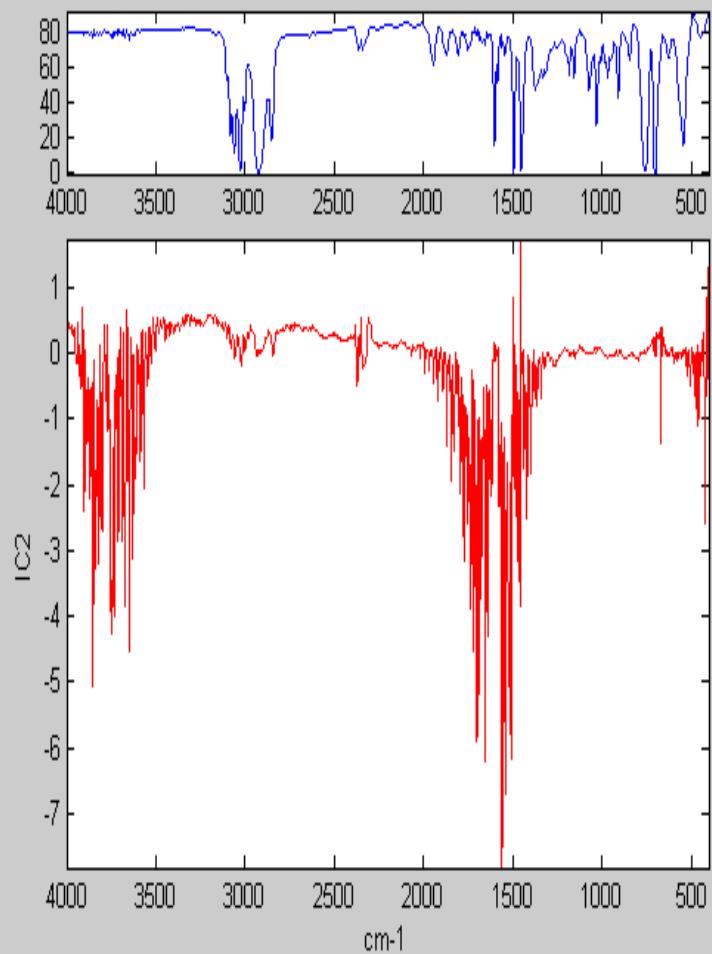


PC6

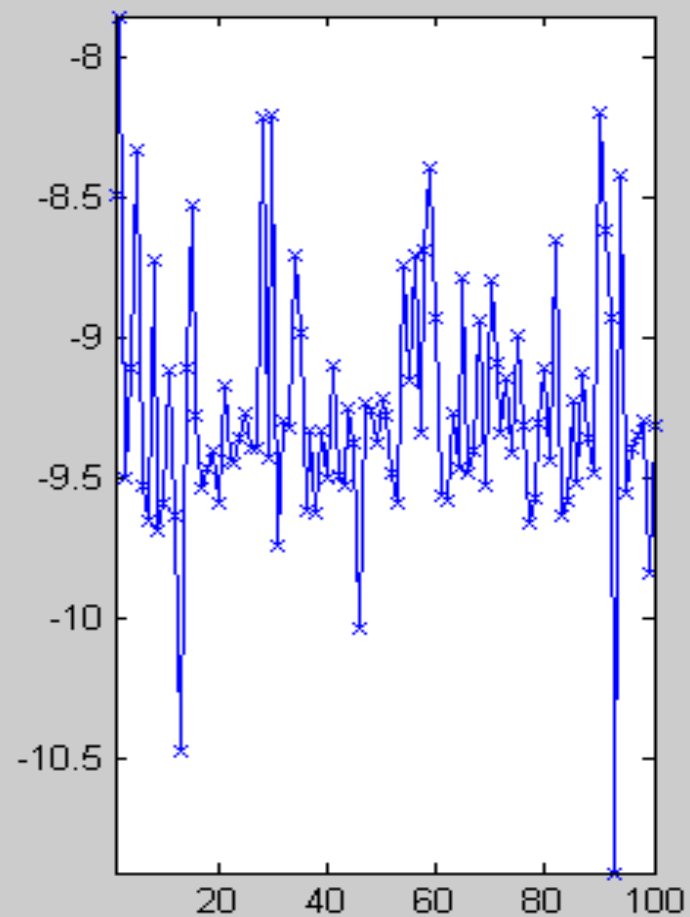
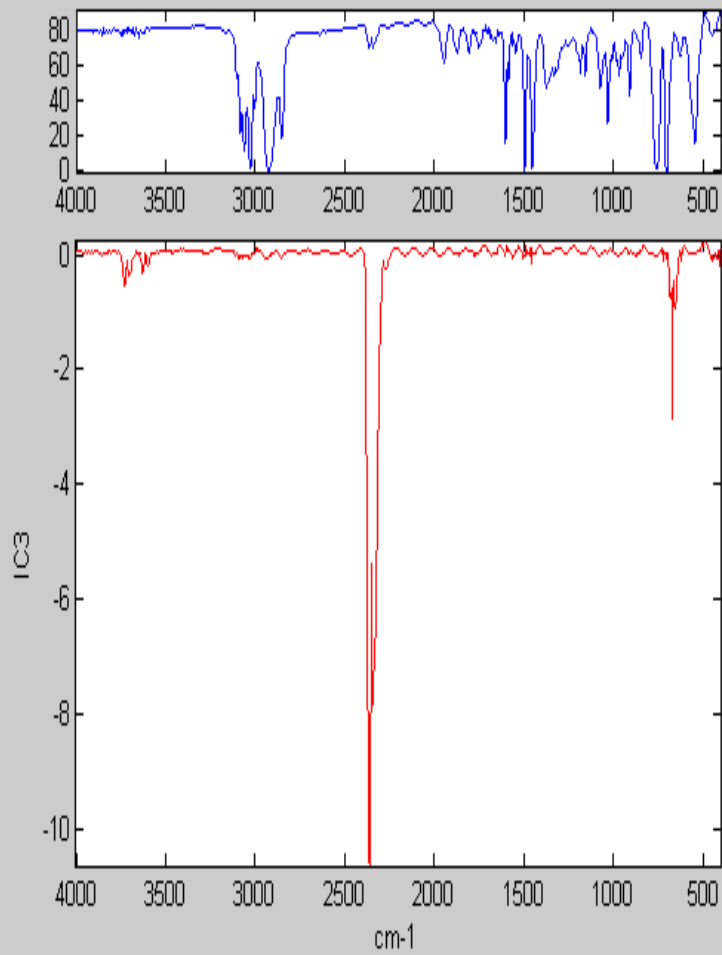
Interférogramme & dérivé

Scores pour PC5-PC6

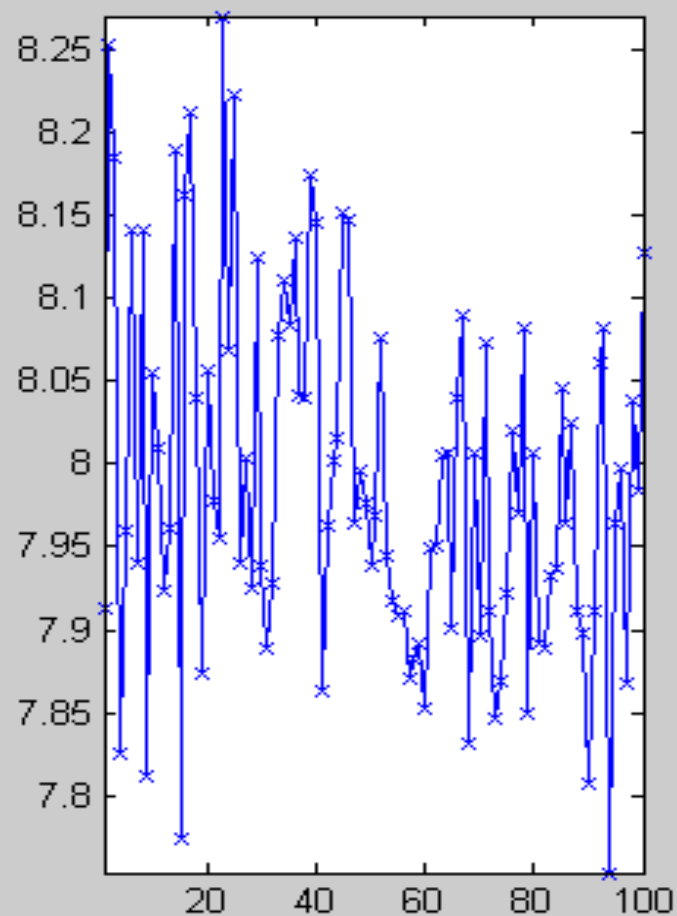
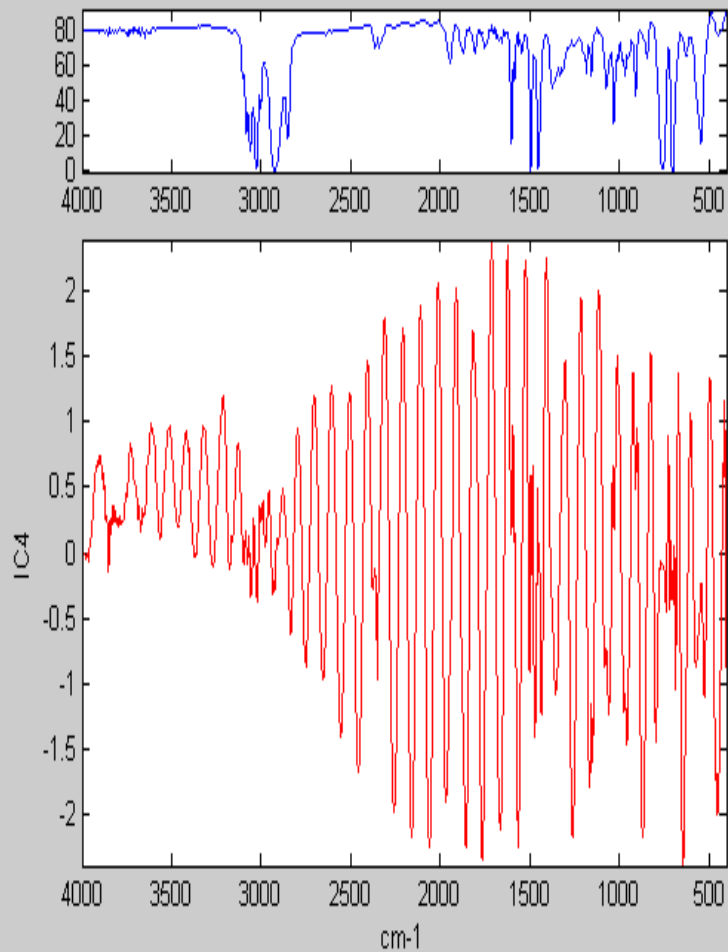




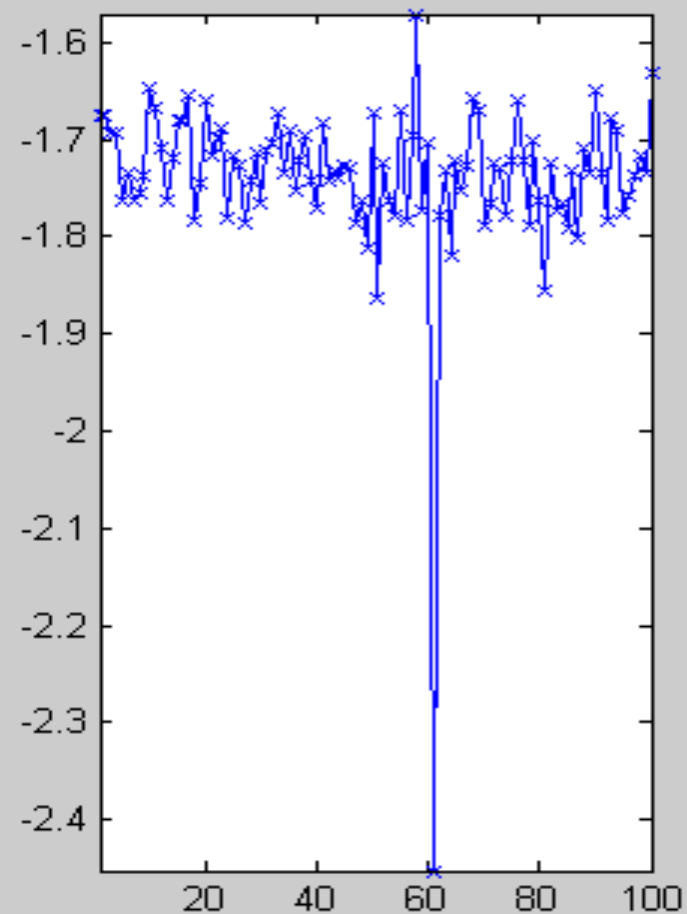
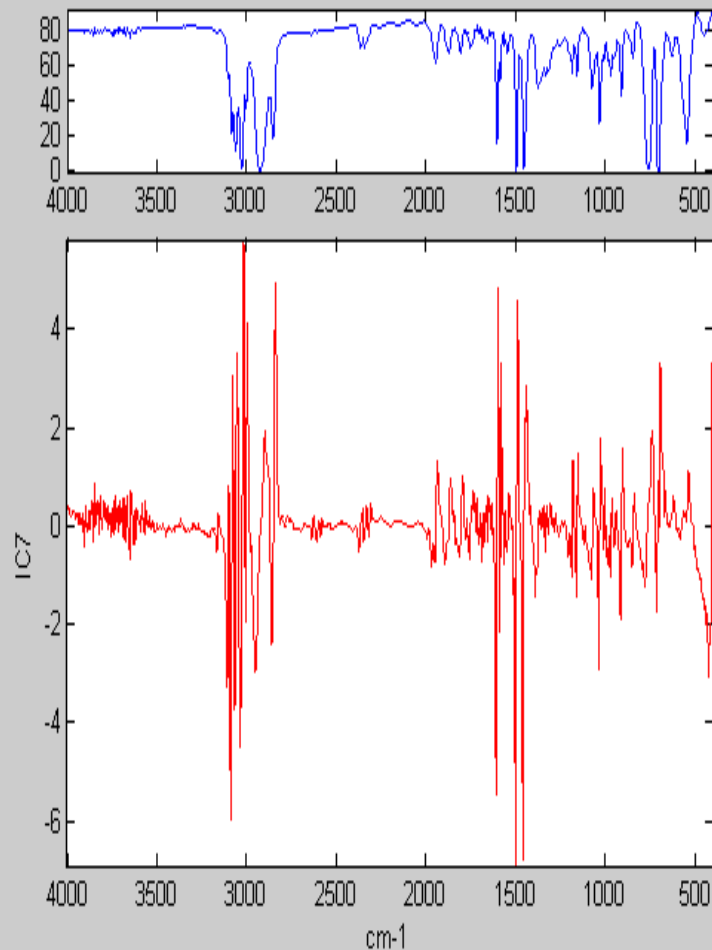
IC2 signal = spectre de vapeur d'eau



IC3 signal = spectre de CO₂



IC4 signal = simple interférogramme



IC7 signal $\approx$ premier dérivé du spectre moyen
Spectromètre N°61 est mal ajusté (décalage de fréquence)

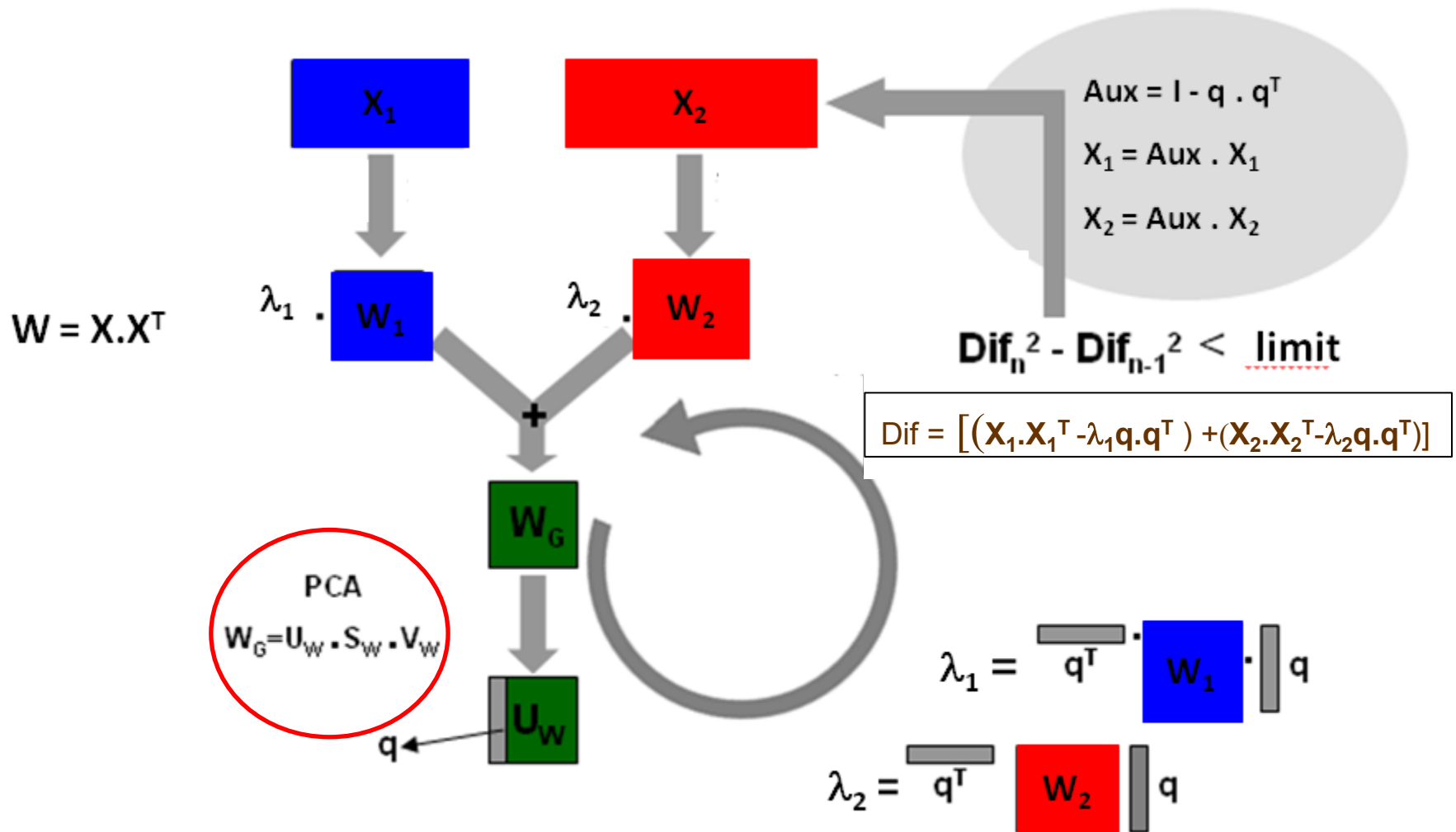
ComDim

- “Analyse en Composantes Communes et Poids Spécifiques”
une méthode exploratoire d’analyse de données multi-bloc
- Analyse simultanée de plusieurs tableaux de données
avec différentes variables décrivant les mêmes objets
- Développée initialement en sensométrie
- Appliquée depuis pour fusionner des données analytiques

E.M. Qannari, I. Wakeling, P. Courcoux, H.J.H. Macfie,
“Defining the underlying sensory dimensions”, *Food Quality and Preference*, 11 (2000) 151-154

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

Algorithme de ComDim



"ComDim" the implementation of CCSWA used here is part of the *SAISIR* toolbox
 SAISIR (2008): Statistics Applied to the Interpretation of Spectra in the InfraRed, D. Bertrand
 (http://www.chimimetrie.fr/saisir_webpage.html)

Une application de ComDim

Données Métabolomiques

Objectif :

Estimer la proportion d'information métabolique commune
extrait des données spectrales de différentes plateformes MS et RMN

Les données

- *Objets*

- 15 membres du laboratoire
 - *échantillons d'urine avec et sans ajouts d'un mélange de métabolites*

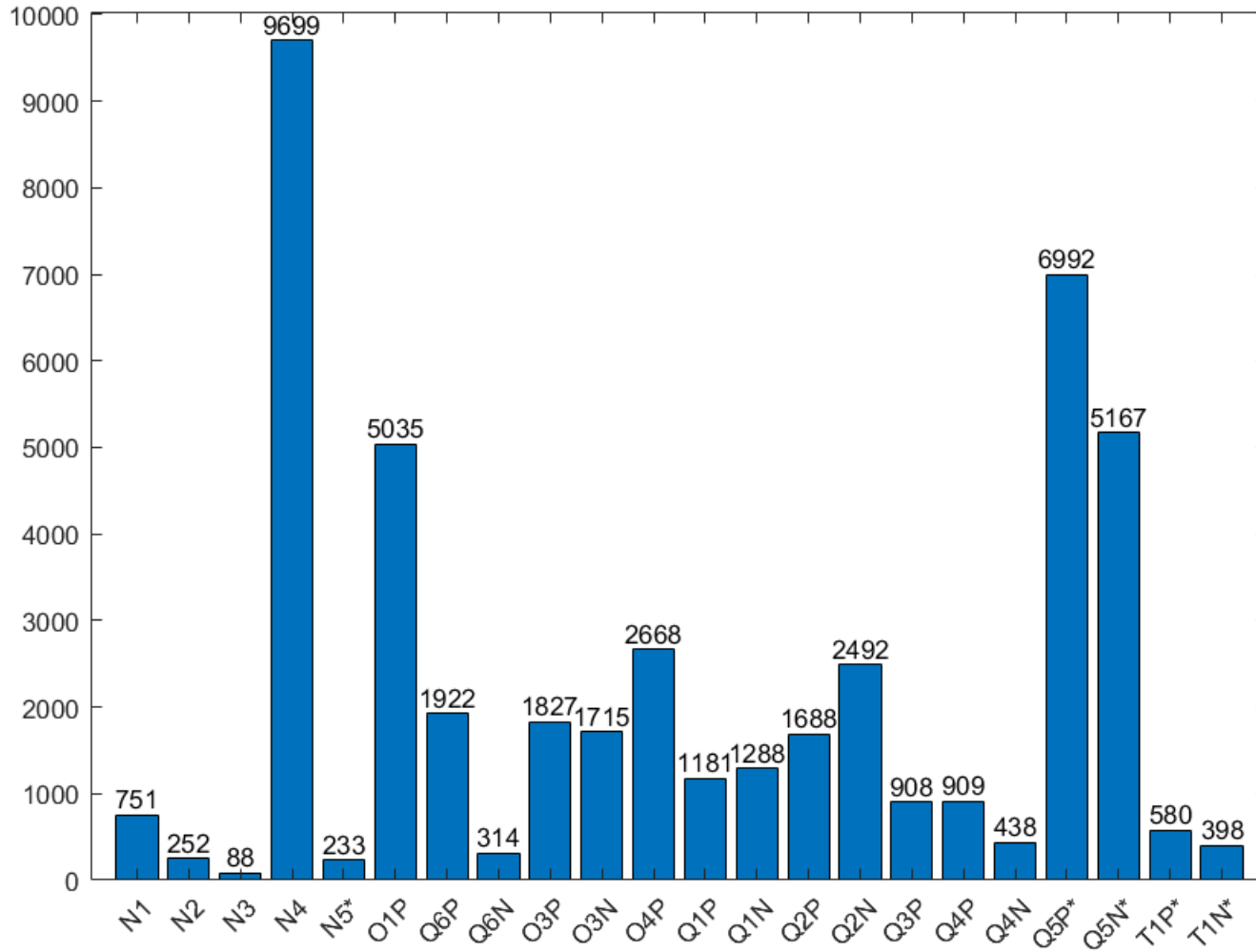
- *Techniques*

- *échantillons* analysés par 5 RMN and 9 MS
- Participants travaillaient comme d'habitude chez eux
- Entre 88 et 9699 variables dans les tableaux

- *Prétraitement*

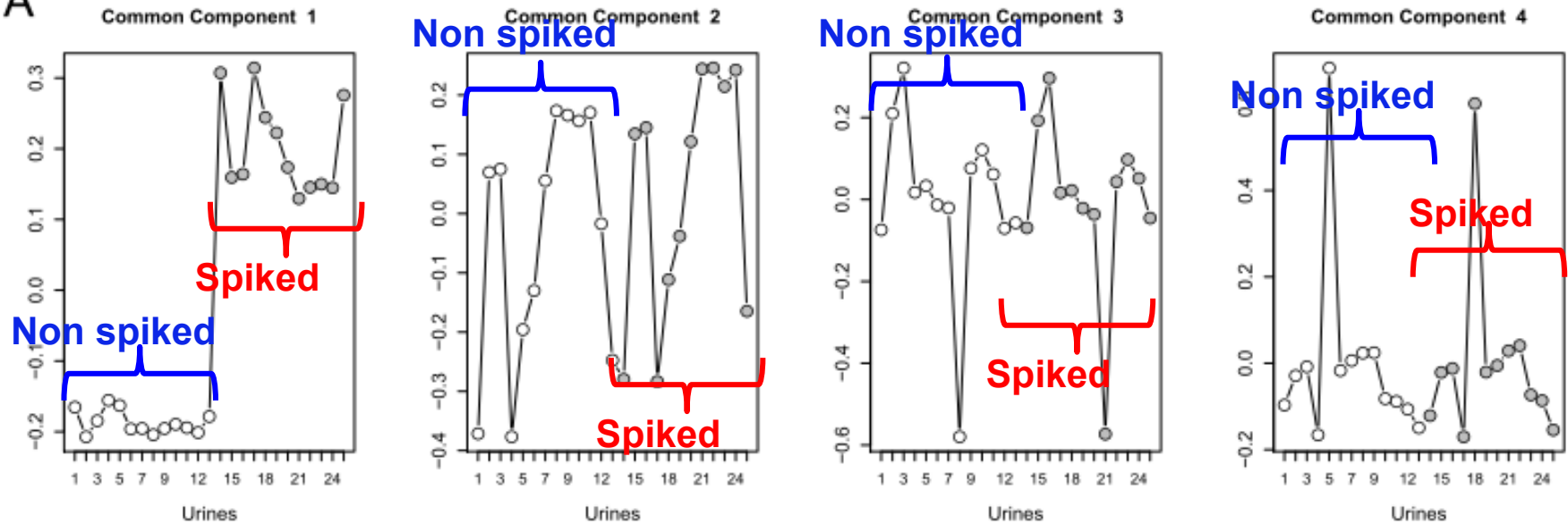
- RMN : Pareto-scaled
- MS : Log10 puis Pareto-scaled

Nombre de variables retenues



Scores et Saliences

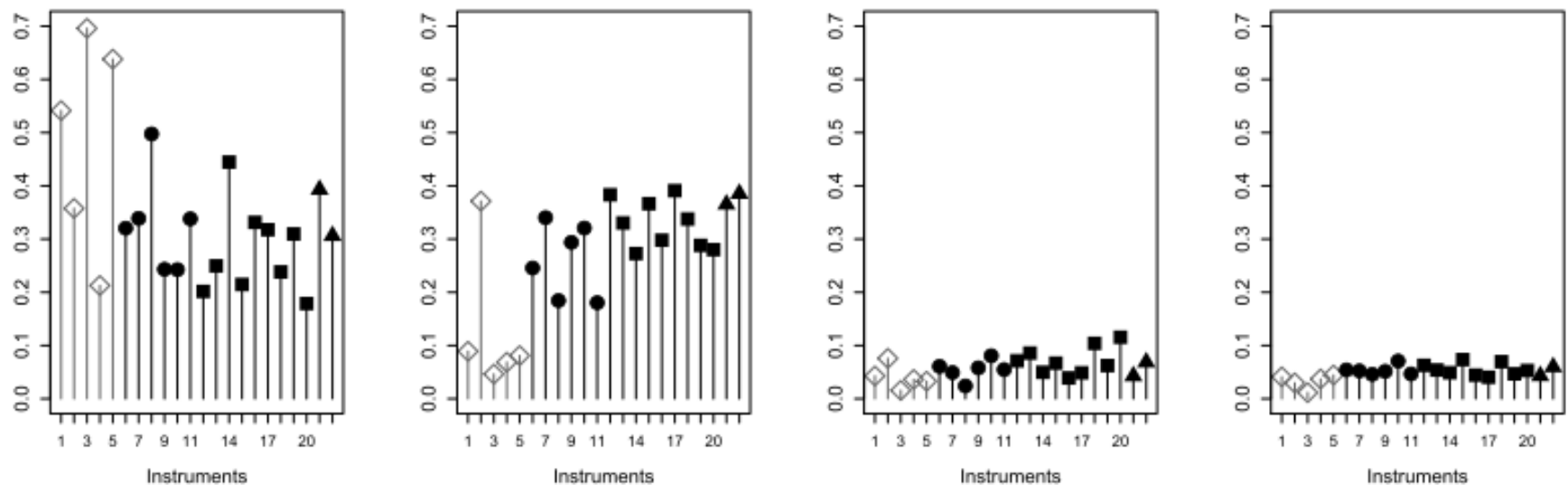
A



B

Symboles ouverts : RMN / Symboles noir : MS

on Component 4



Individu #5

Platform ID	ionization mode	Features ID	Adducts type	Generated formula	Annotated formula	Metabolites ID proposition
Q1	POS	M533T564	[2M+H]1+	C36H37O4	C18H18O2	3-hydroxy-estra-1,3,5(10),6,8-pentaen-17-one
Q1	POS	M535T525	[M+Na2-H]+	C27H44NaO9	C21H35O2	Epipregnanolone glucuronide
Q1	POS	M533T585	[M+Na]1+	C27H42NaO9	C21H34O3	21-hydroxyallopregnanolone-glucuronide
Q1	POS	M528T601	[M+NH4]1+	C27H46NO9	C21H34O3	Dihydroxy-pregnane-beta-one-glucuronide
Q1	POS	M528T554	[M+NH4]1+	C27H46NO9	C21H33O2	Dihydroxy-pregnane-beta-one-glucuronide
Q1	POS	M513T626	[M+H]1+	C27H45O9	C21H36O3	5beta-Pregnane-3alpha,17alpha,20alpha-triol
Q1	POS	M204T623	[M+H -H2O]1+	C8H16NO6	C8H15NO6	N-Acetyl-D-glucosamine (glucuronide like fragment)
Q1	POS	M549T486	[M+H]1+	C21H35O4	C21H34O4	9-deoxy-9-methylene-PGE2 glucuronide
Q1	POS	M571T586				unknown
Q1	POS	M487T469	[M+H -H2O]1+	C18H23O2	C18H24O3	hydroxyestradiol glucuronide
Q1	POS	M467T330	[M+Na]1+			unknown
Q1	POS	M506T349	[M+NH4]1+			unknown
Q1	POS	M531T567	[M+H]1+	C21H33O3	C21H32O3	Hydroxypregnenolone glucuronide
Q1	POS	M204T546	[M +H-H2O]1+	C8H14NO5	C8H15NO6	N acetyl glucosamine residue
Q1	POS	M423T298				unknown
Q1	POS	M231T83	[2M+Na]1+	C8H16NaO6	C4H8O3	a-Hydroxyisobutyric acid
Q1	POS	M285T666	[M+Na]1+	C24H32NaO5	C24H32O5	7alpha-Hydroxy-3,12-dioxochola-1,4-dien-24-oic Acid

Conclusions

Platforms MS (et 1 RMN) distinguent les différences inter-individus

RMN & MS détectent 2 individus inhabituels

Pour MS, pas d'effet de technologie (Orbitrap = TOF = QTOF)

Peu d'influence du nombre de variables extraites (de 88 à 9699)

Untargeted Metabolomics:

**Results of the Metabo-ring Initiative, a Large-scale Multi-instruments Inter-laboratory Study,
Metabolomics (2015) 11 (4), 807-821**

Jean-Charles Martin, Matthieu Maillot, Gérard Mazerolles, Alexandre Verdu, Bernard Lyan, Carole Migne, Catherine Defoort, Cecile Canlet, Christophe Junot, Claude Guillou, Claudine Manach, Christophe Cordella, Daniel Jabob, Delphine Jouan-Rimbaud Bouveresse, Estelle Paris, Estelle Pujos, Fabien Jourdan, Franck Giacomoni, Frédérique Courant, Gaëlle Favé, Gwenaëlle Le Gall, Hubert Chassaigne, Jean-Claude Tabet, Jean-Francois Martin, Jean-Philippe Antignac, Laetitia Shintu, Marianne Defernez, Mark Philo, Marie-Cécile Alexandre-Gouaubau, Marie-Jo Amiot-Carlin, Mathilde Bossis, Mohamed N. Triba, Natali Stojilkovic, Nathalie Banzet, Roland Molinié, Romain Bott, Sophie Goulitquer, Stefano Caldarelli, Douglas N. Rutledge

Merci

New ComDim algorithm

(proposed by M. Hanafi & E.M Qannari)

Start with p matrices $\mathbf{X}_i$ of size $n \times k_i$ ($i = 1$ to p)

Each $\mathbf{X}_i$ column-centered and scaled by dividing by matrix norm to give $\mathbf{Xs}_i$

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

At each iteration, the p weighted $\mathbf{Xs}_i$ matrices are concatenated column-wise ,
resulting in a global $\mathbf{X}_G$ matrix

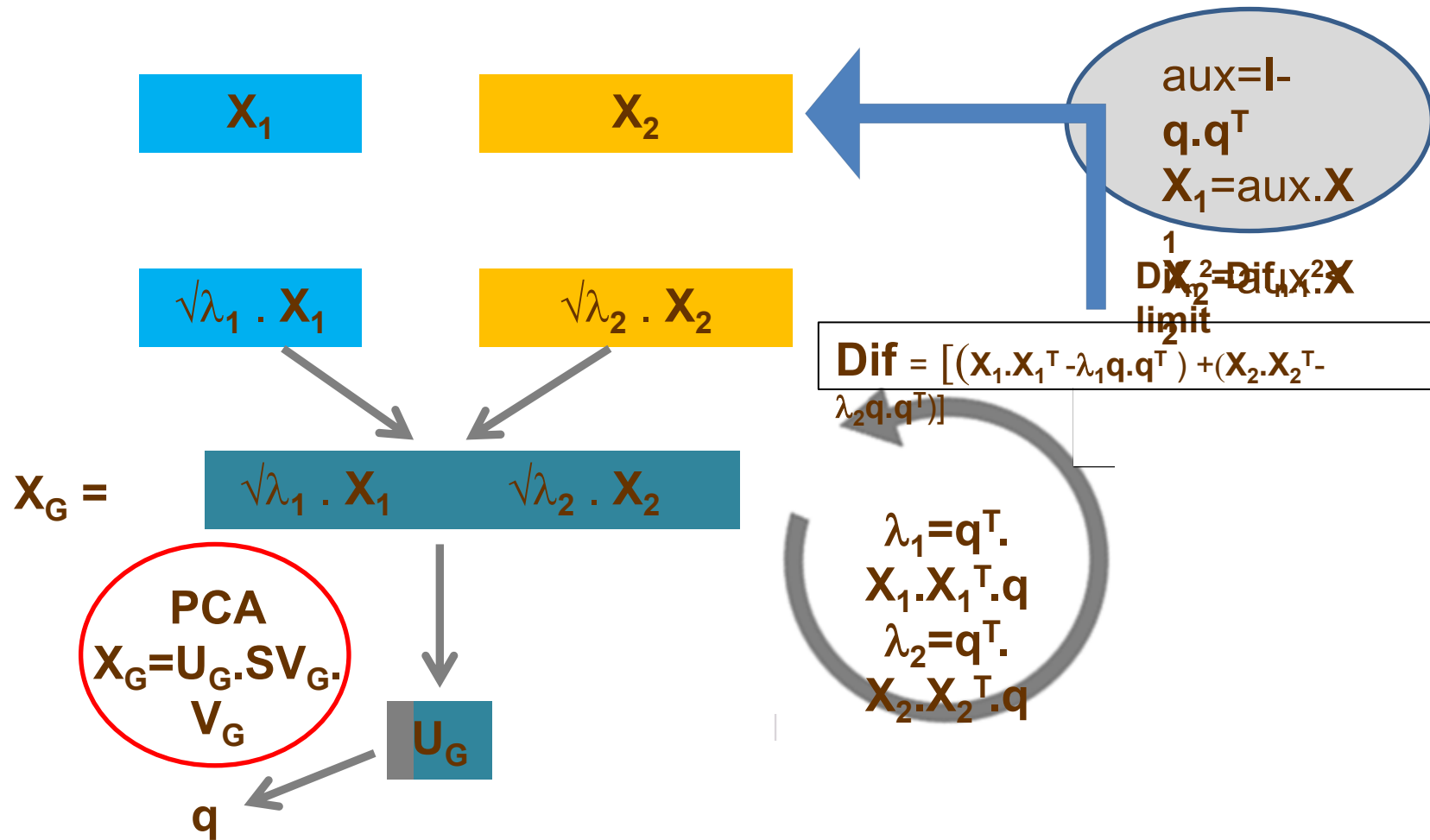
$$\mathbf{X}_G = [\sqrt{\lambda_1}\mathbf{Xs}_1, \sqrt{\lambda_2}\mathbf{Xs}_2, \sqrt{\lambda_3}\mathbf{Xs}_3, \dots, \sqrt{\lambda_p}\mathbf{Xs}_p]$$

M. Hanafi, E.M Qannari,

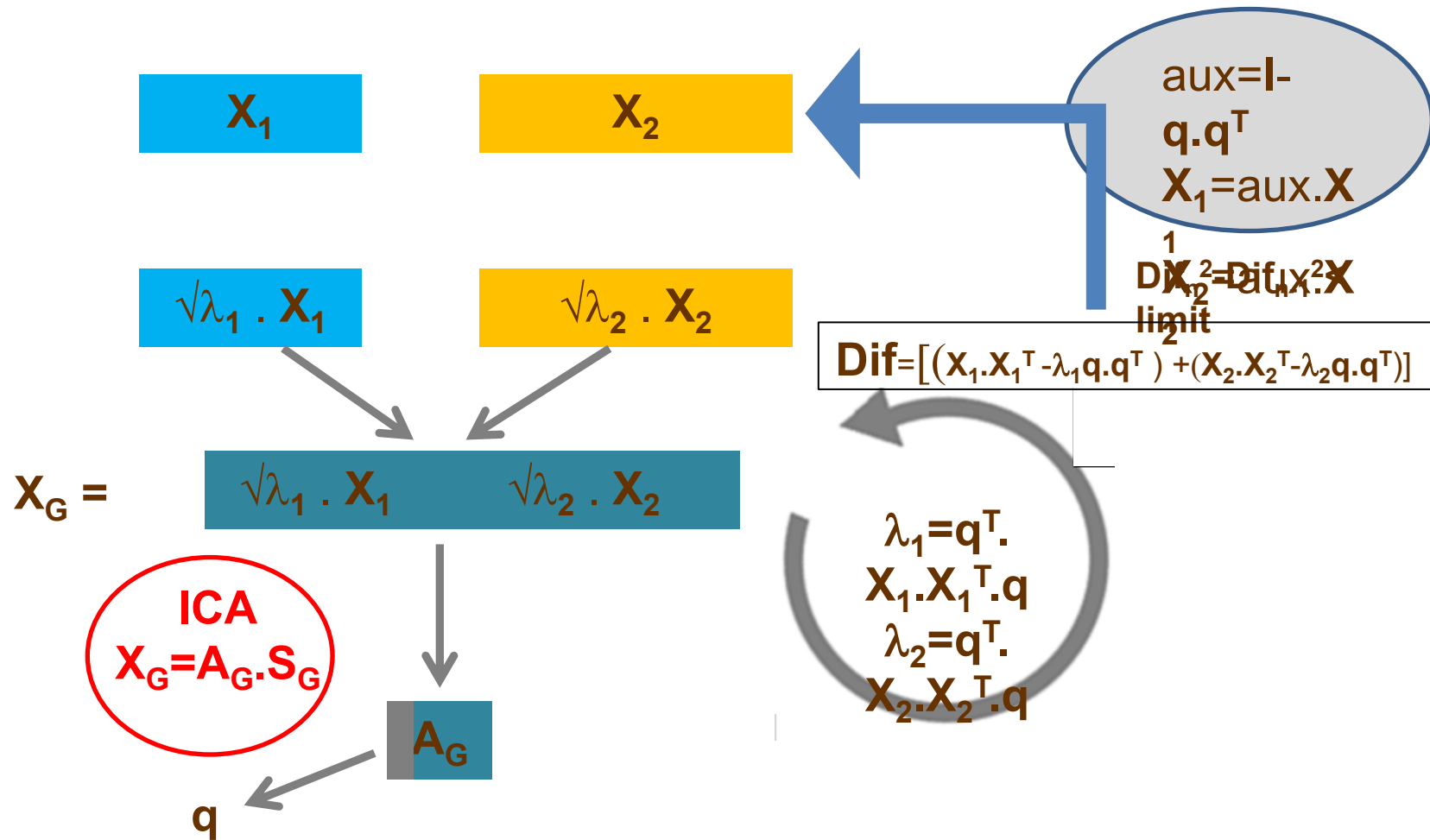
“Nouvelles propriétés de l'analyse en composantes communes et poids spécifiques”

Journal de la société française de statistique, Tome 149 (2008) no. 2, p. 75-97

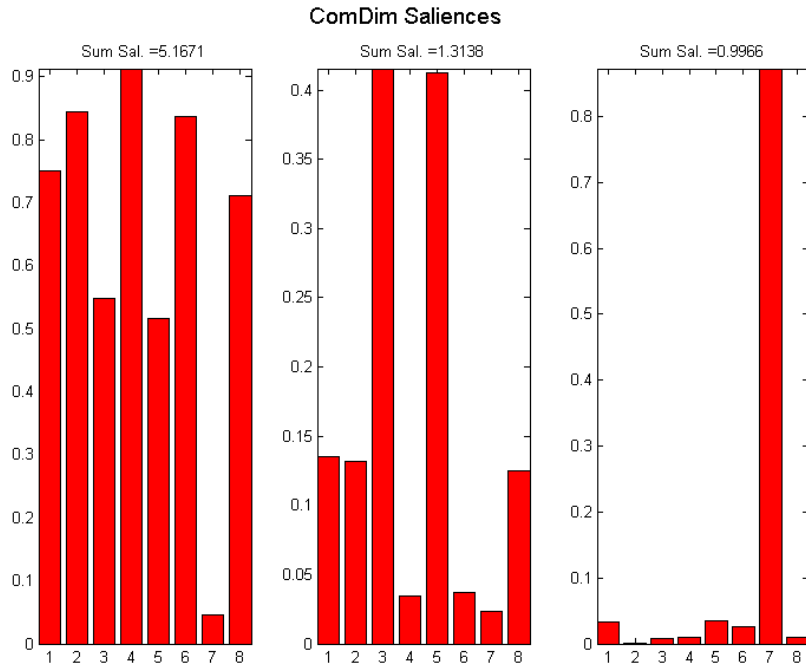
New ComDim algorithm



ICA-based ComDim (MultiBlock ICA !)



PCA-ComDim versus ICA-ComDim

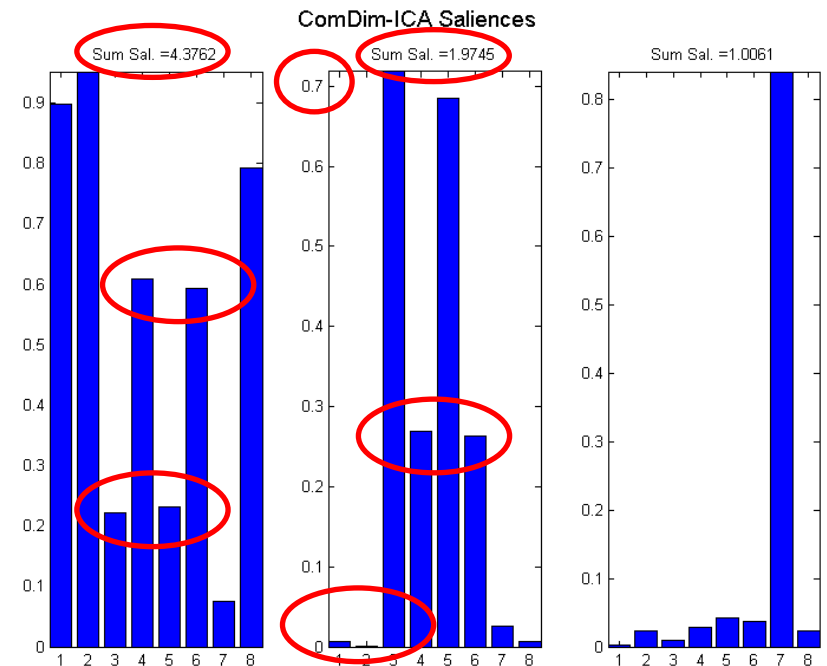


ComDim-ICA Saliences for CC1,
CC2 & CC3

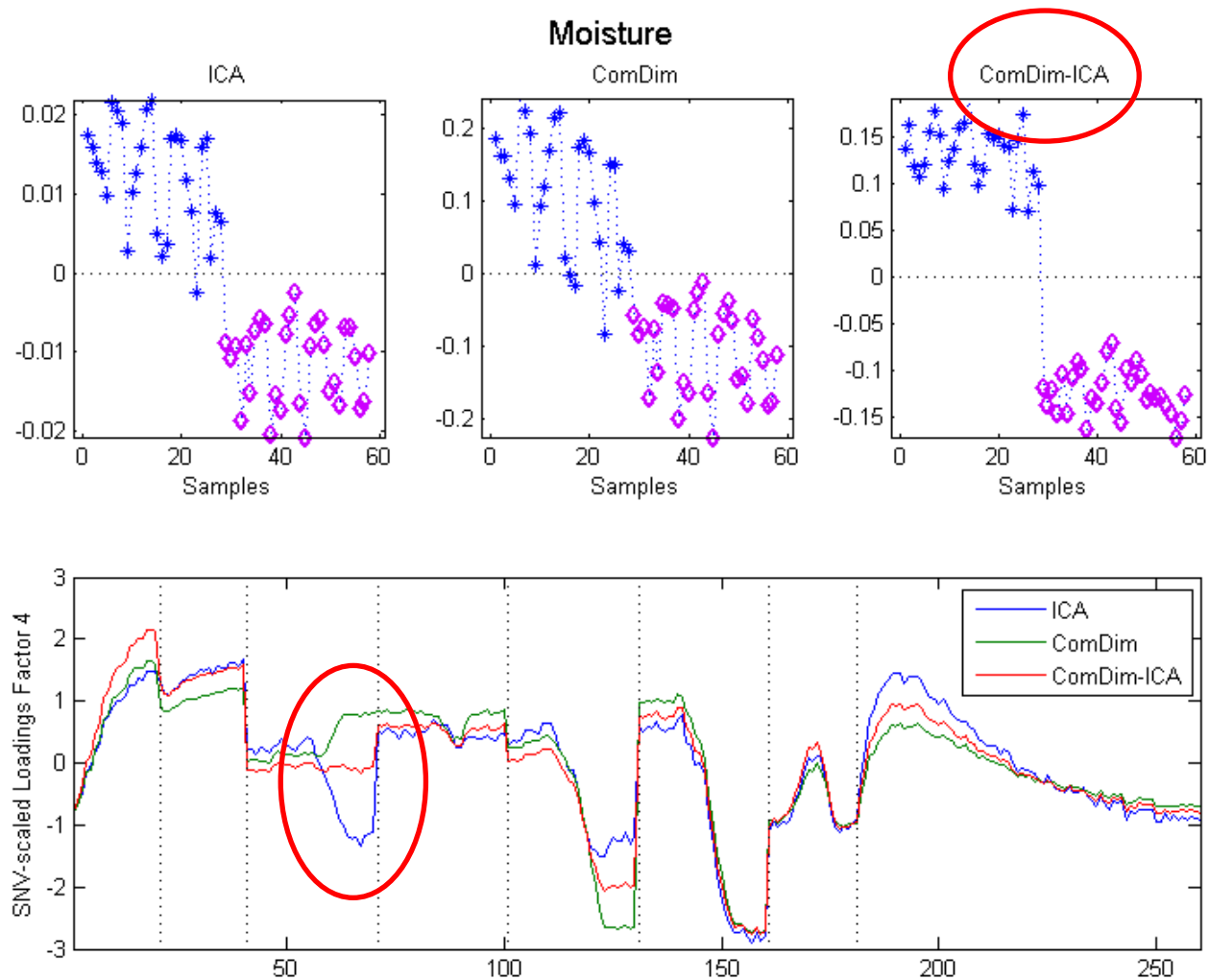
Convergence in 104 mS

ComDim Saliences for CC1,
CC2 & CC3

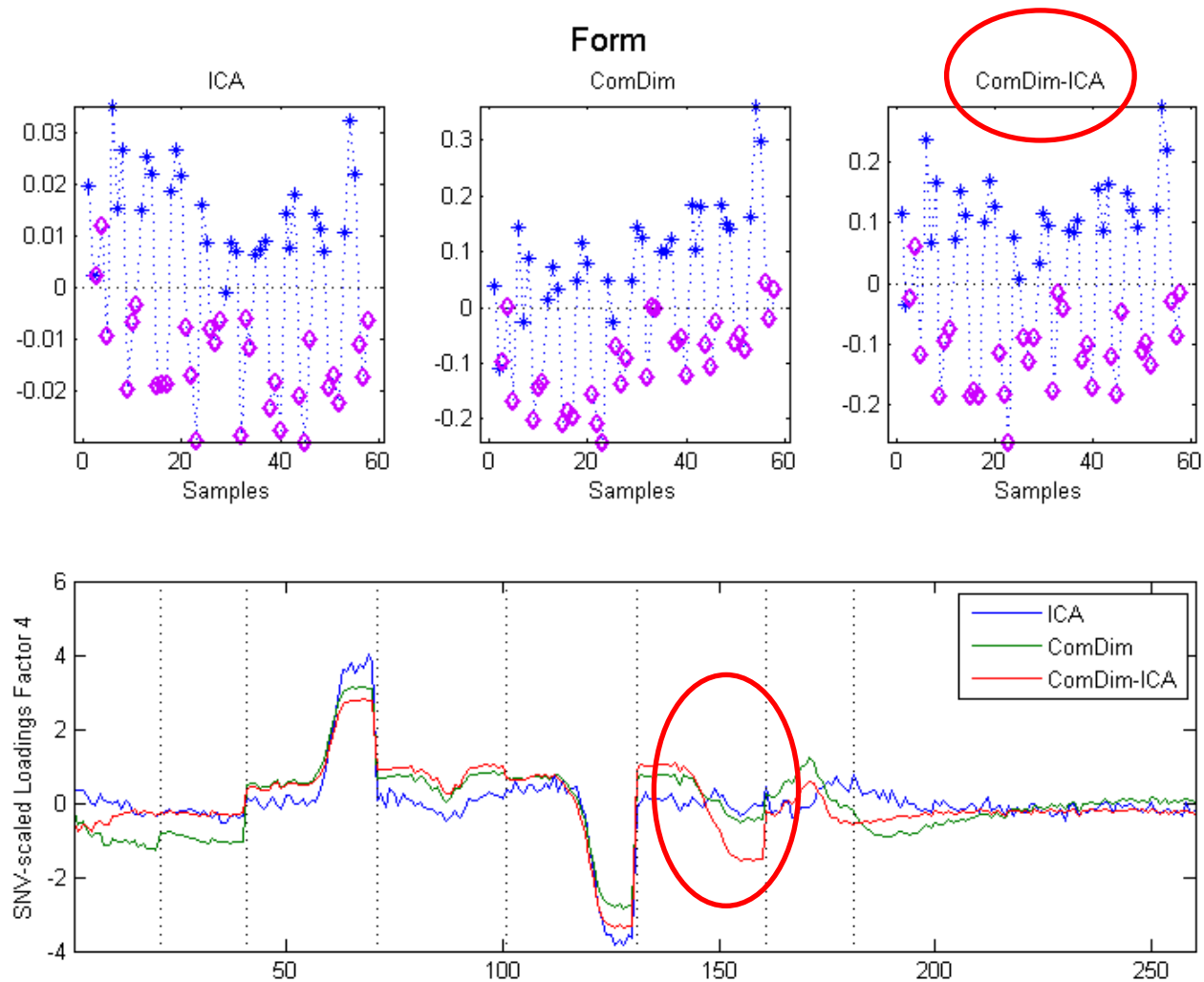
Convergence in 124 mS



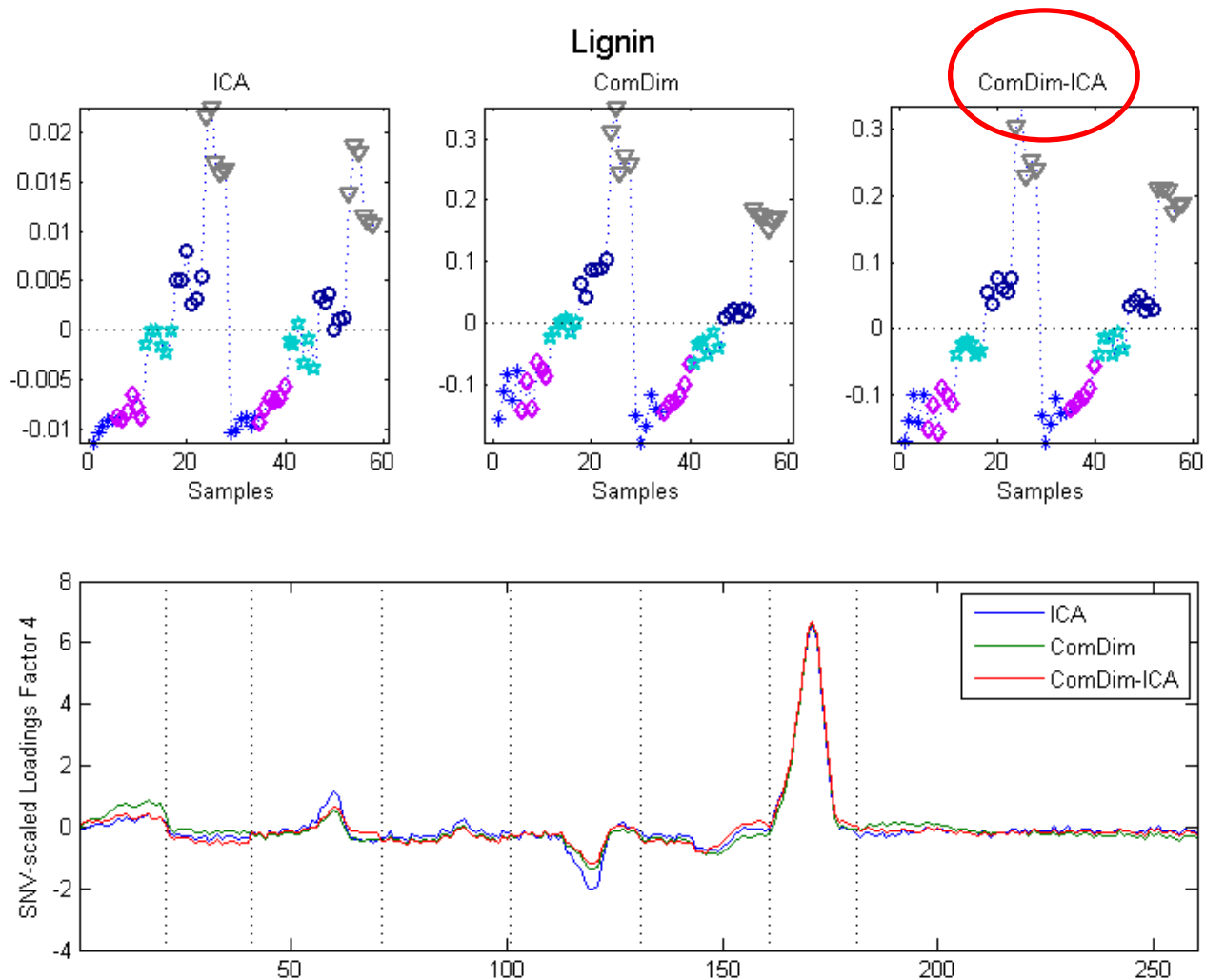
Comparison of ICA and PCA-ComDim with ICA-ComDim



Comparison of ICA and PCA-ComDim with ICA-ComDim



Comparison of ICA and PCA-ComDim with ICA-ComDim



Conclusion

ICA-ComDim

A non-supervised multi-block method

ICA on iteratively re-weighted concatenated data tables

- Better than *PCA-ComDim* !
(*PCA on iteratively re-weighted* concatenated data tables)
- Better than *ICA* !
(*ICA on unweighted* concatenated data tables)