Détermination du nombre optimal de composantes dans l'Analyse en Composantes Indépendantes

Amine Kassouf [1] Delphine Jouan-Rimbaud Bouveresse [2,3] Douglas N. Rutledge[2]

[1] Département de Chimie et de Biochimie, Faculté des Sciences Ⅱ, Université Libanaise, 90656 Jdeideth El Matn, Fanar, Liban.
[2] UMR Ingénierie Procédés Aliments, AgroParisTech, INRA, Université Paris-Saclay, F-91300 Massy, France.
[3] UMR 914 Physiologie de la Nutrition et du Comportement Alimentaire, INRA, AgroParisTech, Université Paris-Saclay, F-75005 Paris

Plan

- Introduction
- Aperçu théorique
  - ICA_by_blocks
  - Random_ICA
  - Critère de Durbin-Watson: DW_Residuals
  - Indice de Kaiser-Meyer-Olkin: KMO_ICA_Residuals
  - ICA_corr_y
- Applications
- Conclusion

Introduction

Rutledge et al. *Trends Anal Chem. 2013*

- **Analyse en composantes indépendantes** (ICA): technique de séparation en aveugle de sources.
- Elle consiste à estimer ***F*** signaux sources, statistiquement indépendants, à partir de ***n*** signaux observés, considérés comme étant des combinaisons linéaires de ces signaux sources.

Modèle général

X = A.S

X: matrice des signaux observés

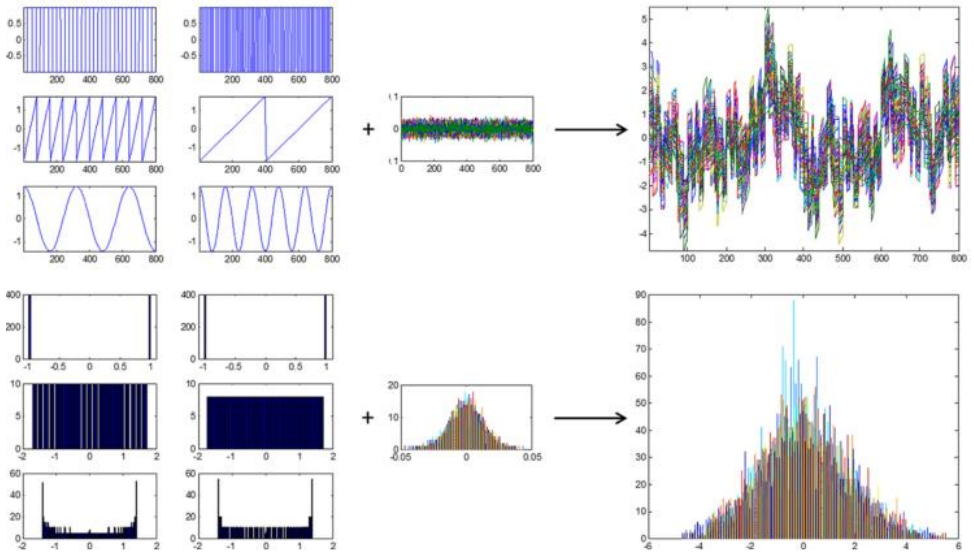
A: matrice des proportions

S: matrice des signaux "sources" (**ICs**)

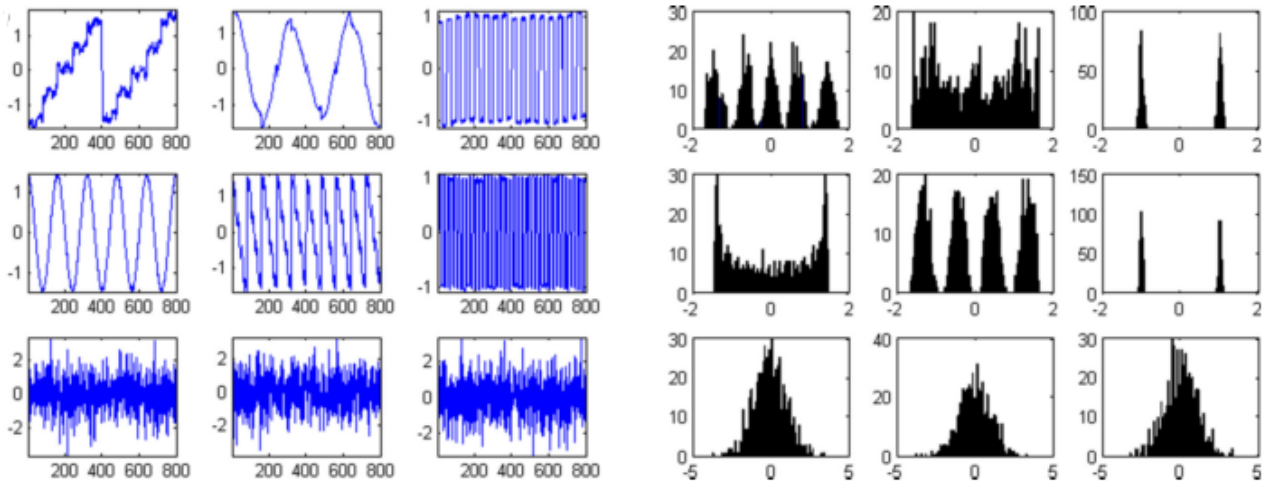
ICA

(JADE algorithm)

ICA cherche à estimer les signaux sources (ICs) par une transformation linéaire visant à maximiser l'indépendance entre les signaux extraits, en maximisant leur caractère **non-gaussien** (Théorème Central Limit).



ICA



Introduction

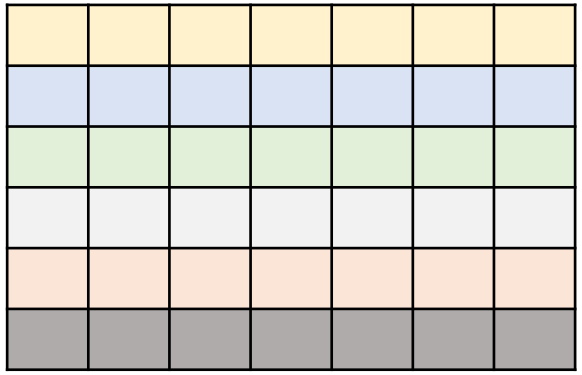
Pourquoi déterminer le nombre optimal de composantes indépendantes (ICs)

- Les ICs ne sont pas classées par ordre d'importance.
- Calculer trop peu d'ICs ⟶ des signaux non-purs.
- Extraire trop d'ICs ⟶ sur-décomposition des signaux / introduction de bruit.

ICA_by_blocks **Random_ICA** **DW_Residuals** **KMO_ICA_Residuals** **ICA_corr_y**

Aperçu théorique: ICA_by_blocks

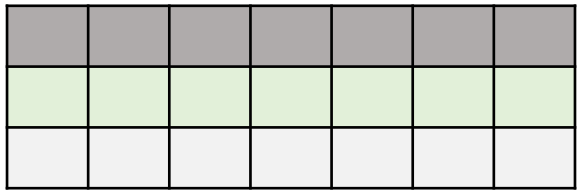
Jouan-Rimbaud Bouveresse et al. *Chem. Intell. Lab. Syst. 2012*



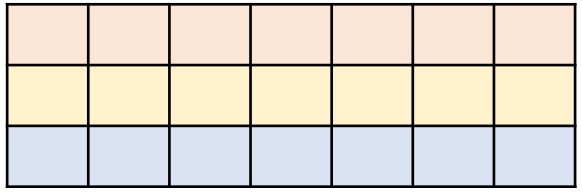
X (n,p)

1

Répartition des *n* individus de **X** en *B*=2 blocs représentatifs de **X**



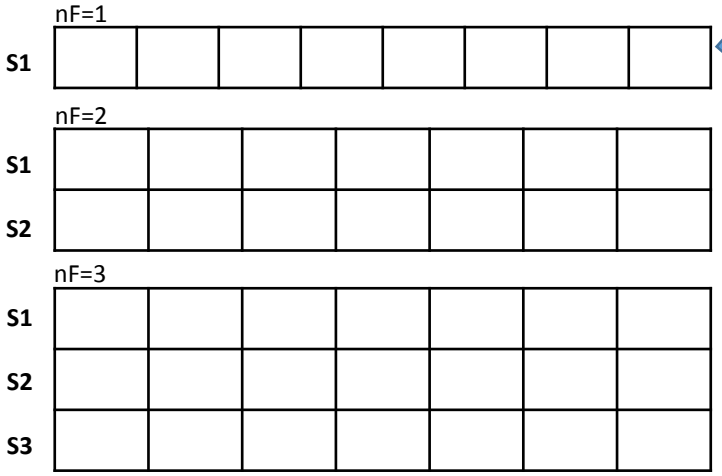
B1 (n/2,p)



B2 (n/2,p)

2

Pour chaque bloc, F_{max} modèles **ICA** sont calculés avec *nF*=1;2;3…F_{max} ICs



3

Les corrélations entre les ICs extraites des deux blocs sont calculées

	S1 (B1)	S2 (B1)	S1 (B2)	S2 (B2)
S1 (B1)			~0	~1
S2 (B1)			~1	~0
S1 (B2)				
S2 (B2)				

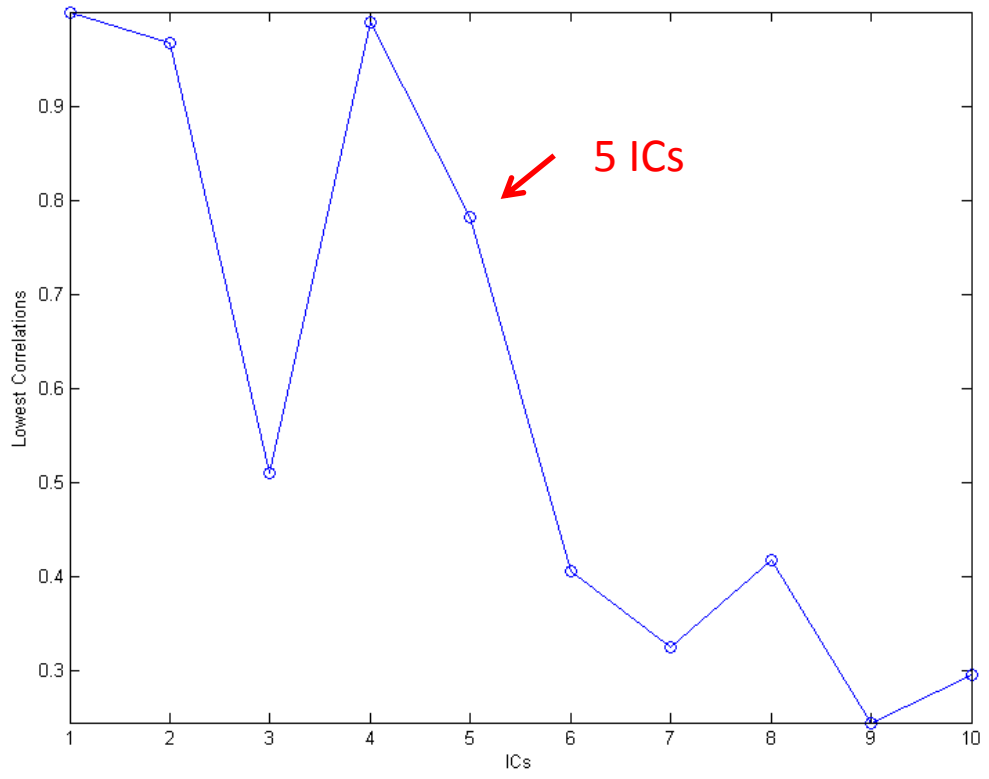
4

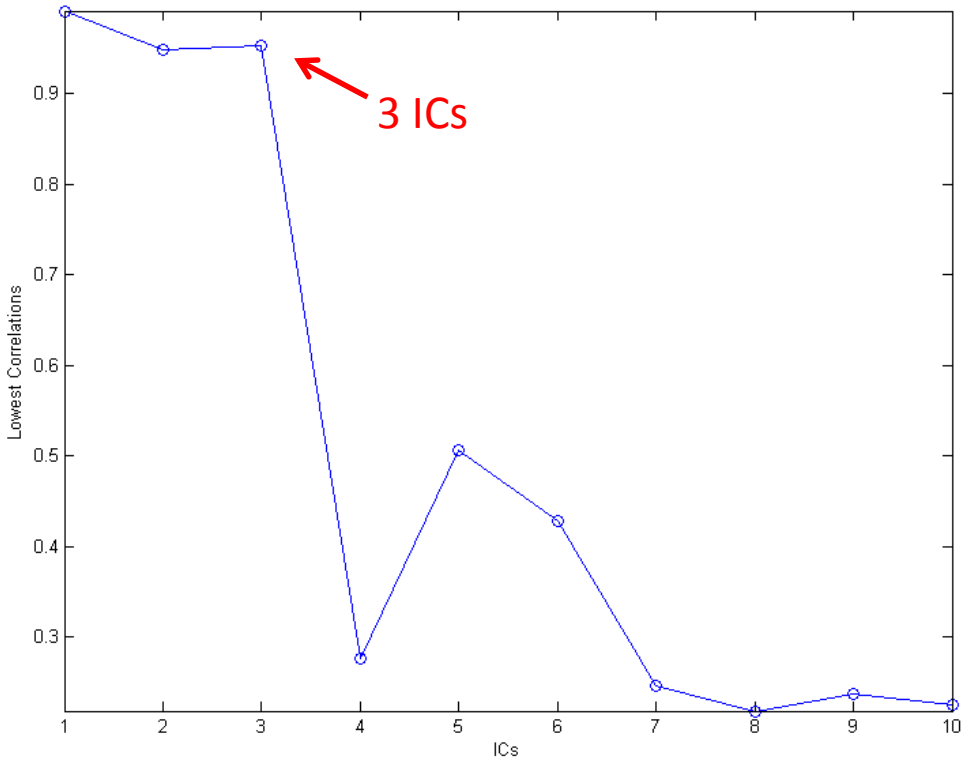
Le modèle donnant le plus grand nombre d'ICs corrélées indique le **nombre optimal d'ICs**

Problème avec la méthode

Lignin data (58 × 260)

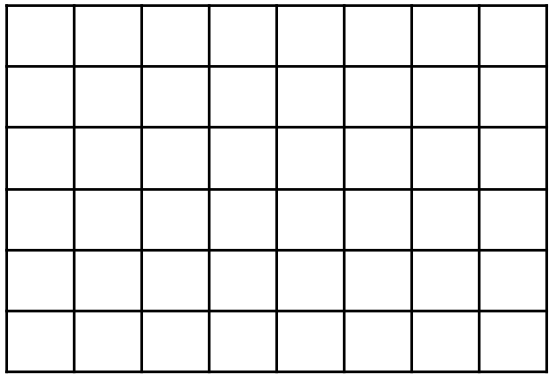
ICA_by_Blocks, B = 2





Aperçu théorique: Random_ICA

Kassouf et al. *Talanta*. 2018



X (n,p)

1

Répartition **aléatoire** des *n* individus de **X** en *B*=2 blocs.

2

Pour chaque bloc, F_{max} modèles **ICA** sont calculés avec nF=1;2;3…F_{max} ICs

3

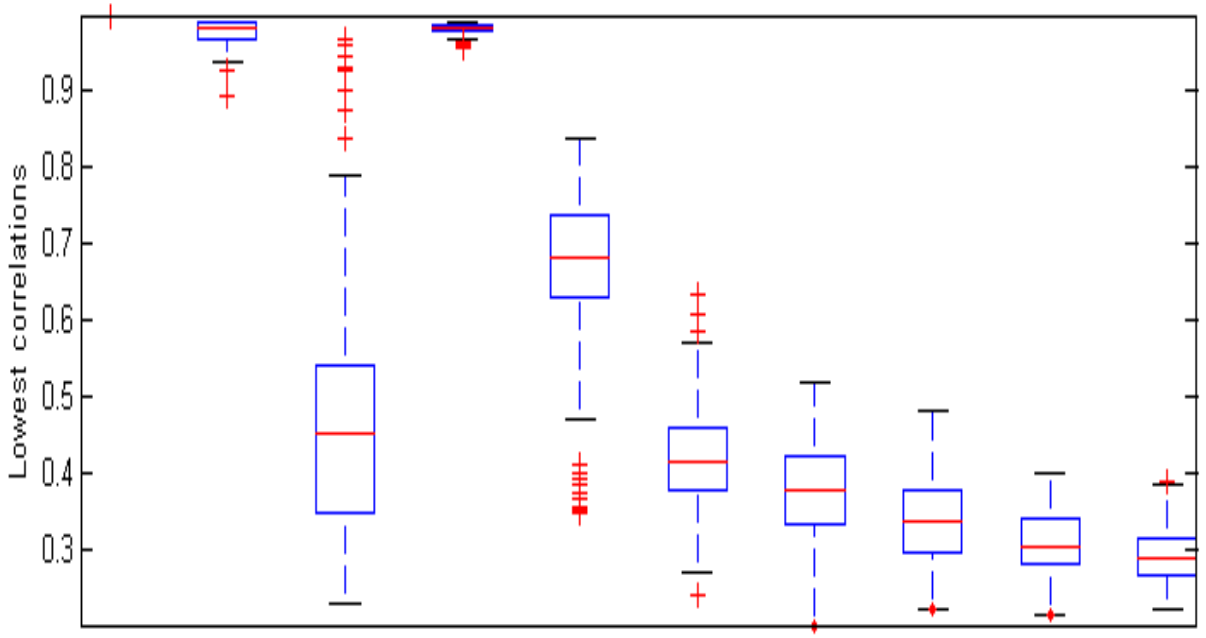
Répétition de la procédure ***k*** fois

Vision plus **large** et **moins biaisée** pour le choix du nombre optimal d'ICs

ICA_by_blocks
(mais avec répartitions **aléatoires** entre blocs)

Random_ICA avec 100 répétitions

Lignin data (58 × 260)



Nombre de signaux extraits

Aperçu théorique: Critère de Durbin-Watson

Jouan-Rimbaud Bouveresse et al. *Chem. Intell. Lab. Syst. 2012*

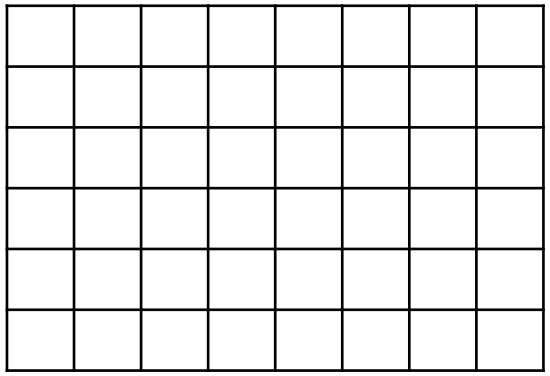
Critère de Durbin-Watson (DW)

$$DW = \frac{\sum_{i=2}^{n} \left(s(i) - s(i-1)\right)^2}{\sum_{i=1}^{n} s(i)^2}$$

s est un signal, s(i) est le $i^{ème}$ point du signal

- DW tend vers 0 s'il n'y a pas de bruit dans le signal
- DW tend vers 2 s'il n'y a que du bruit dans le signal

Aperçu théorique: Critère de Durbin-Watson



X (n,p)

1

Fmax modèles **ICA** sont calculés avec *nF*=1;2;3…*Fmax* ICs

2

Pour chaque modèle ICA:

Matrice résiduelle

R = (X - (AS))

3

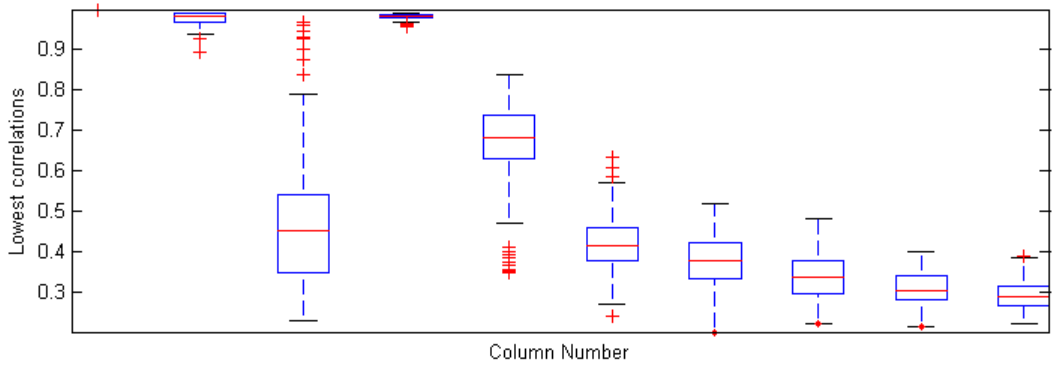
Calcul du critère DW pour chaque signal résiduel (ligne de **R**)

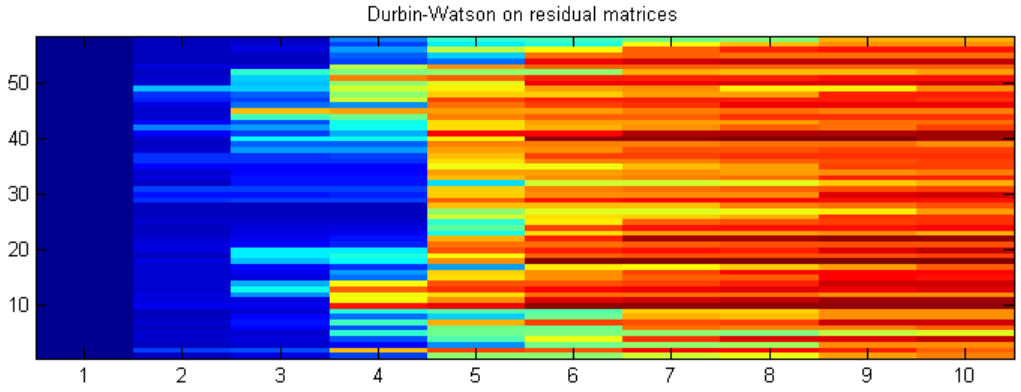
➔

On suit l'évolution des critères DW en fonction du nombre d'ICs dans le modèle

Durbin-Watson

Lignin data (58 × 260)





Nombre de signaux extraits

Aperçu théorique: KMO_ICA_Residuals

Kassouf et al. *Talanta*. 2018

Indice de Kaiser-Meyer-Olkin :

$$\mathrm{KMO} = \frac{\sum_i \sum_{j \neq i} r_{ij}^2}{\sum_i \sum_{j \neq i} r_{ij}^2 + \sum_i \sum_{j \neq i} a_{ij}^2}$$

$$a_{ij} = -\frac{v_{ij}}{\sqrt{v_{ii} + v_{jj}}}$$

$0 \leq \mathrm{KMO} \leq 1$

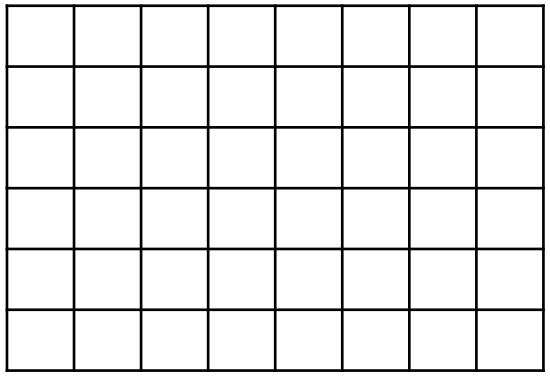
r_{ij} : corrélation entre les variables *i* et *j*
a_{ij} : corrélation partielle entre les variables *i* et *j*

- KMO ≈ 0: Il n'y a pas de corrélation entre les variables
- KMO ≈ 1: Les variables sont corrélées

L'indice KMO est utilisé pour déterminer s'il est utile d'appliquer une Analyse en Composantes Principales à un jeu de données.

Aperçu théorique: KMO_ICA_Residuals

Kassouf et al. *Talanta*. 2018



X (n,p)

1

F_{max} modèles **ICA** sont calculés avec nF=1;2;3…F_{max} ICs

2

Pour chaque modèle ICA:

Matrice résiduelle

$$\mathbf{R = (X - (AS))^T}$$

3

Calcul de l'indice KMO

➔

Corrélations partielles entre les signaux résiduels

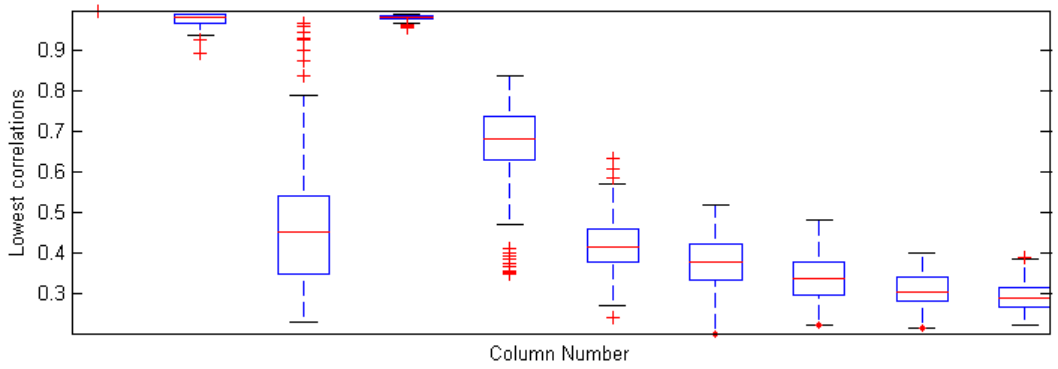
4

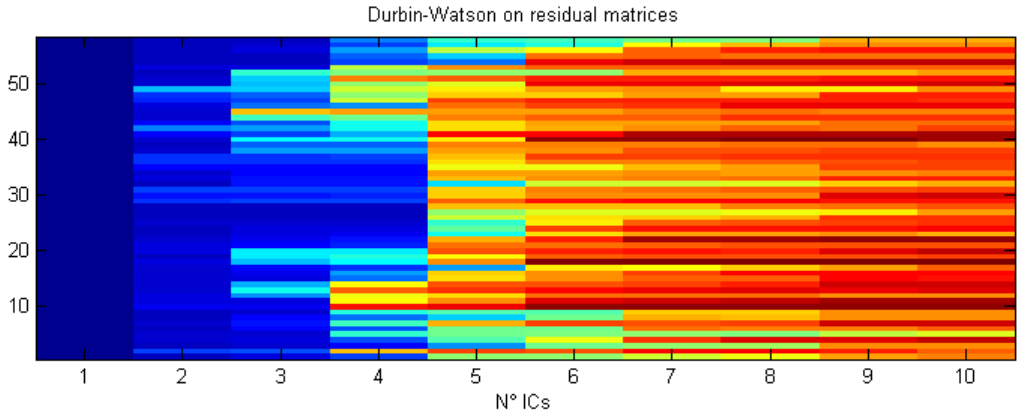
- Indice KMO **proche de 0.5:** pas de corrélations entre les colonnes = tous les signaux sources sont déjà extraits.
- Indice KMO **proche de 1**: il reste encore des signaux sources dans la matrice résiduelle.

Nombre optimal d'ICs

KMO_ICA_Residuals

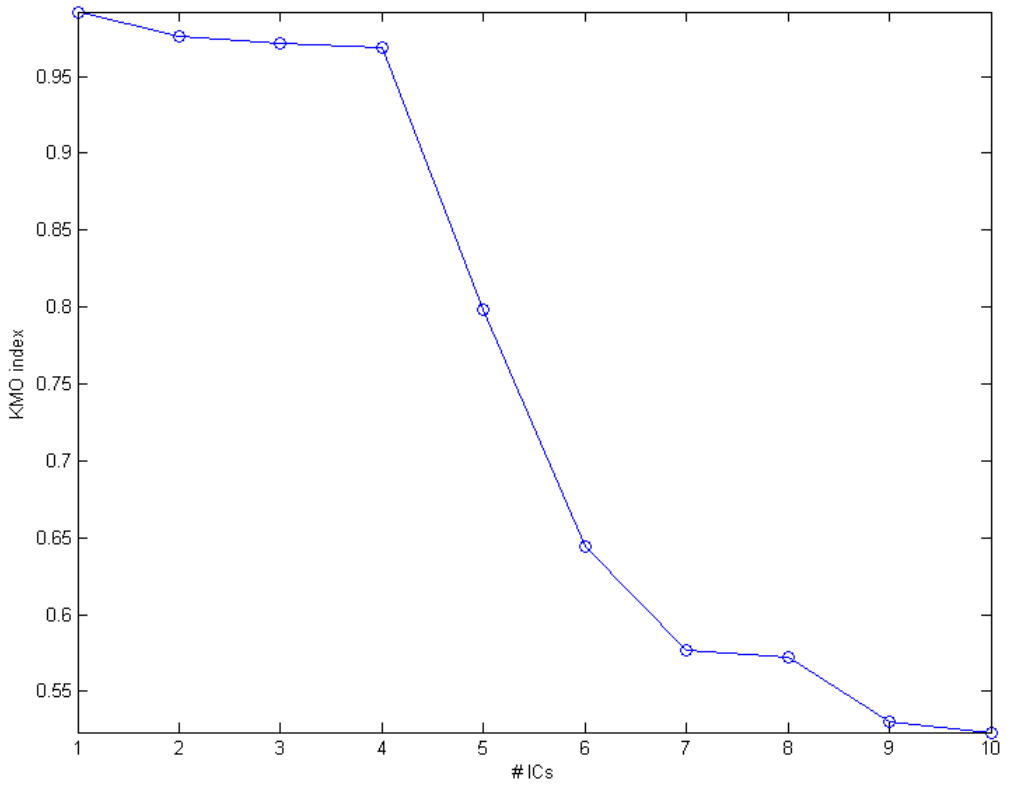
Lignin data (58 × 260)





Nombre de signaux extraits

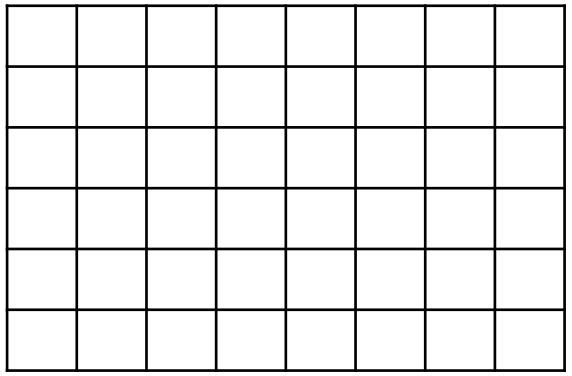
KMO



Nombre de signaux extraits

Aperçu théorique: ICA_corr_y

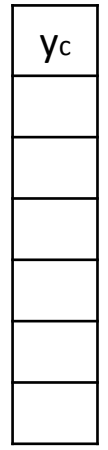
Kassouf et al. *Talanta*. 2018



X (n,p)

1

Fmax modèles **ICA** sont calculés avec *nF*=1;2;3…*Fmax* ICs



Vecteur y_c
(concentrations « vraies »)

2

Vecteur y_s
(signaux « vrais »)

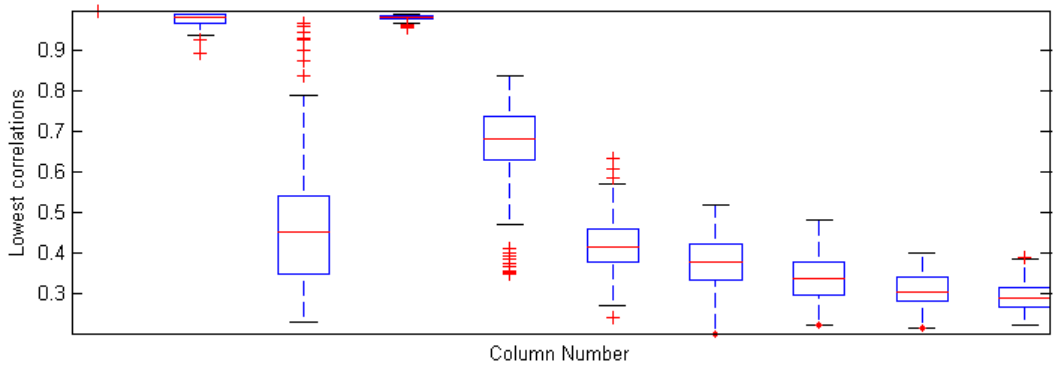
Corrélations R entre **A** et y_c ou entre **S** et y_s , pour nF=1, 2, 3…*Fmax*

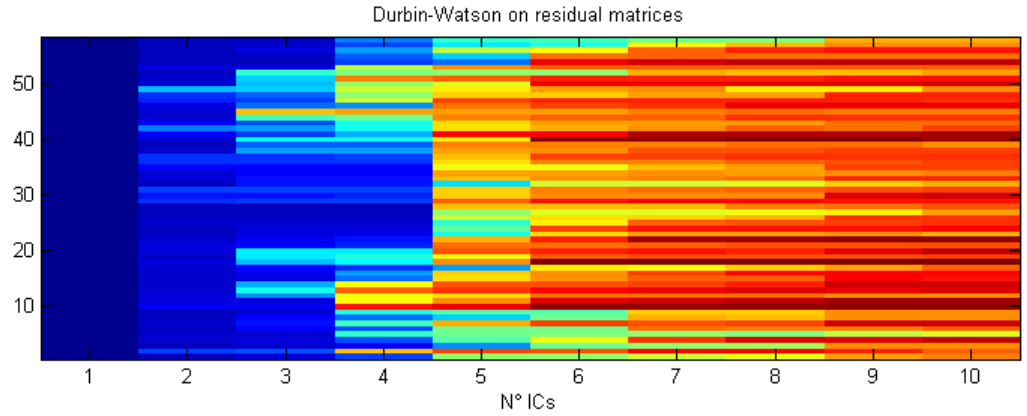
3

Nombre optimal d'ICs est obtenu pour le modèle assurant les valeurs **maximales de corrélation**.

ICA_corr_y

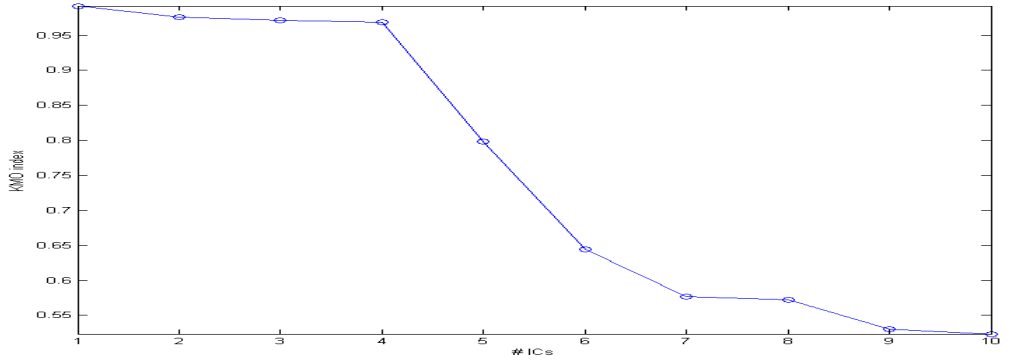
Lignin data (58 × 260)



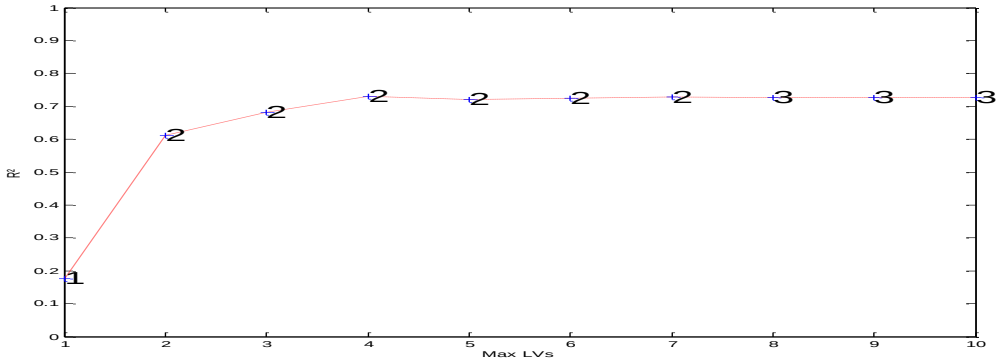


Nombre de signaux extraits

KMO



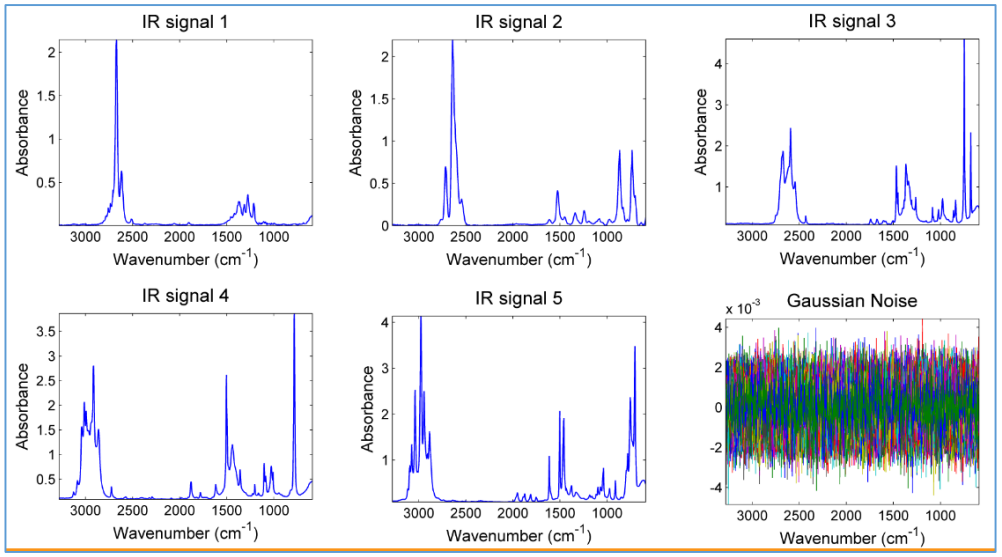
R



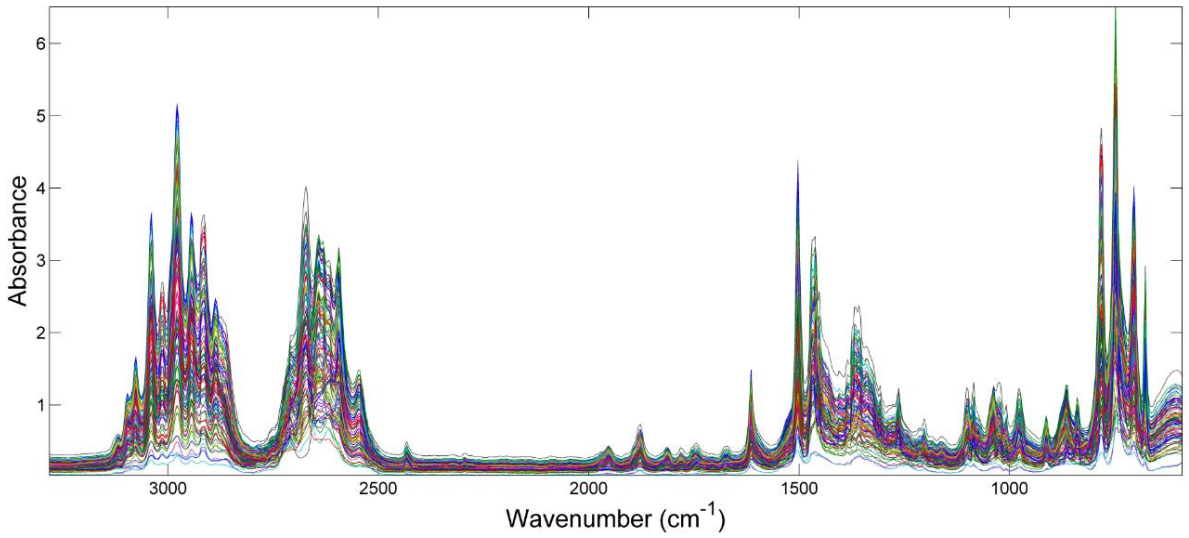
Nombre de signaux extraits

Application: données simulées

Kassouf et al. *Talanta. 2018*

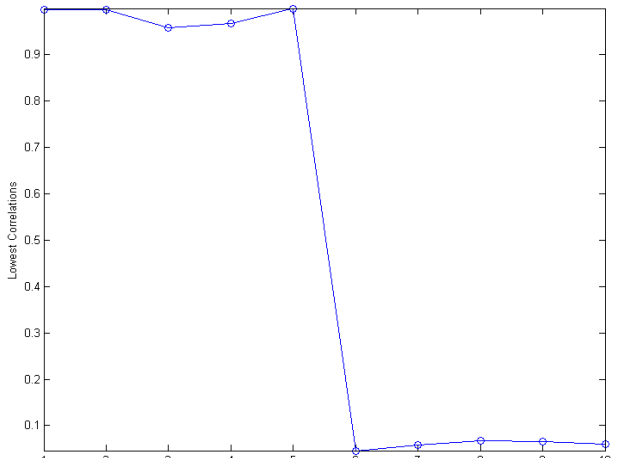


Données IR simulées
X (100,800)

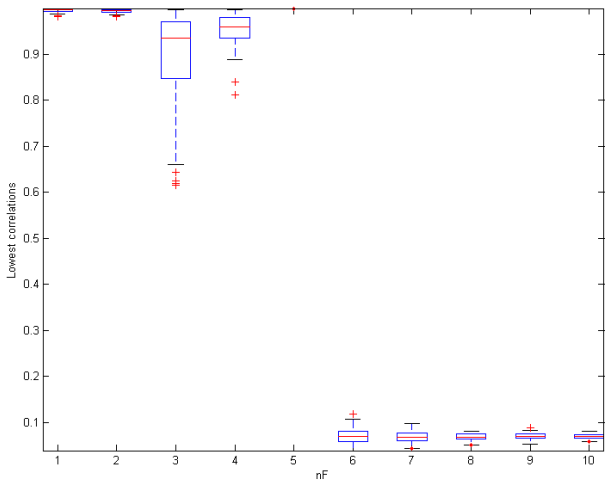


Application: données simulées

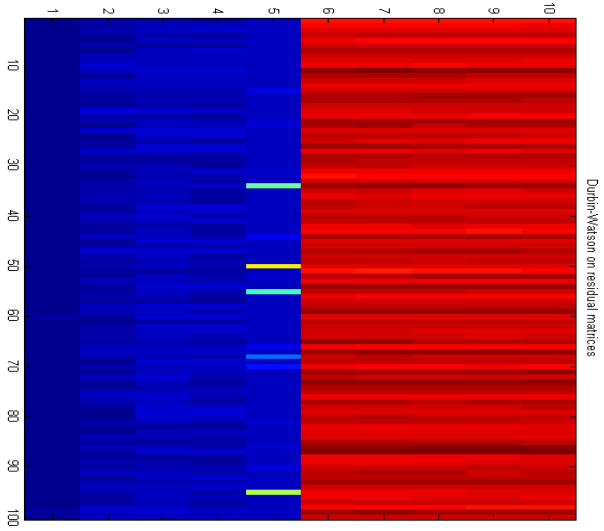
Kassouf et al. *Talanta. 2018*



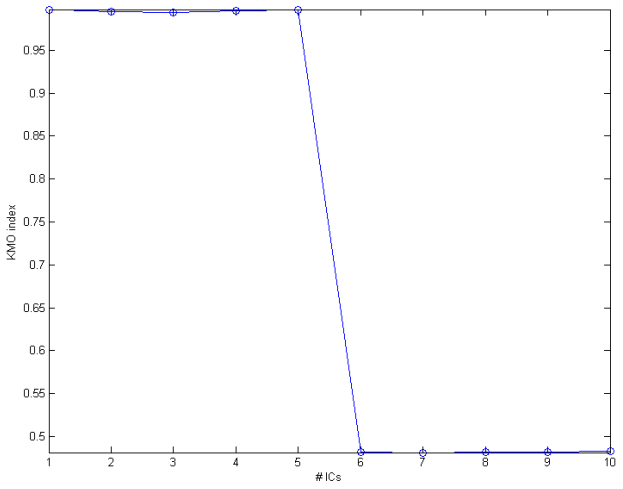
ICA_by_Blocks
B = 2



Random_ICA
50 répétitions



DW



KMO

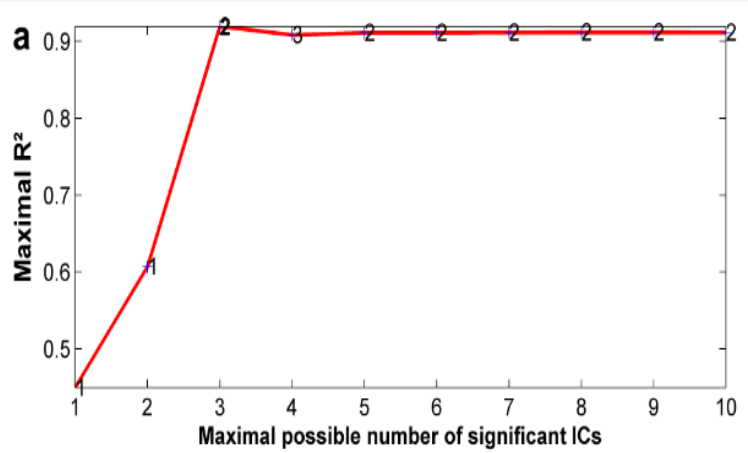
Application: données simulées

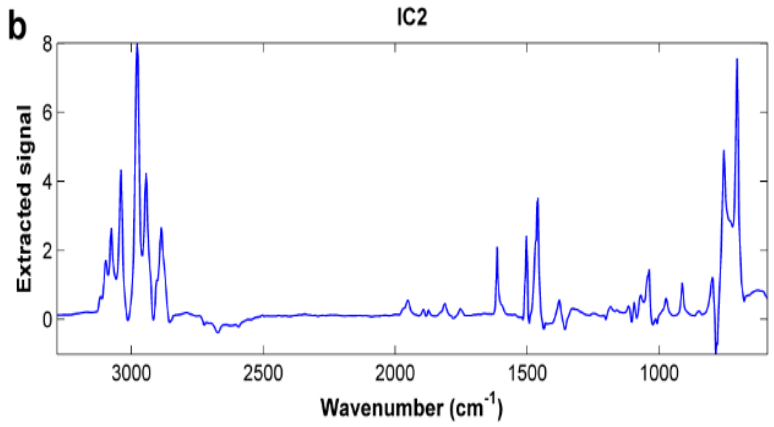
Kassouf et al. *Talanta. 2018*

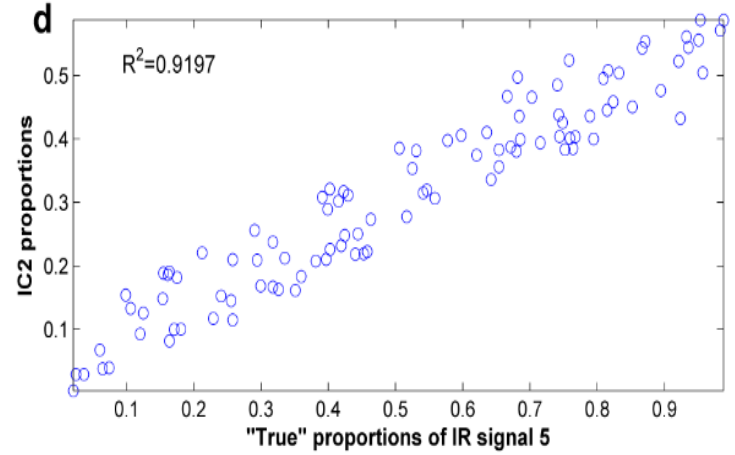
Donnée IR simulée
X (100,800)

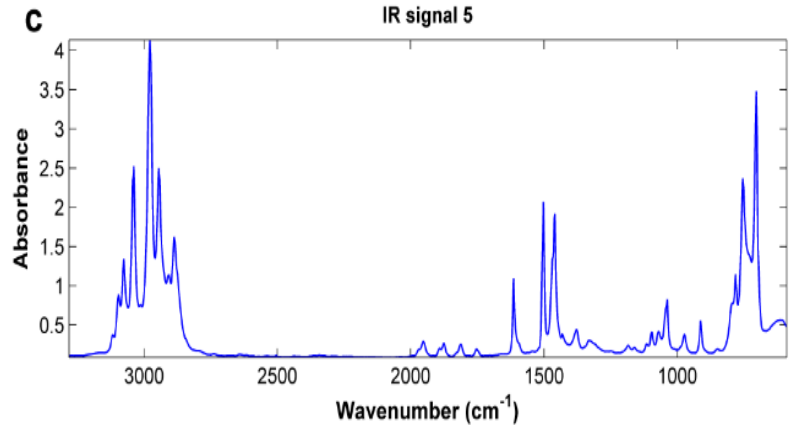
y

Vecteur y_c
(**proportions de IR signal 5**)







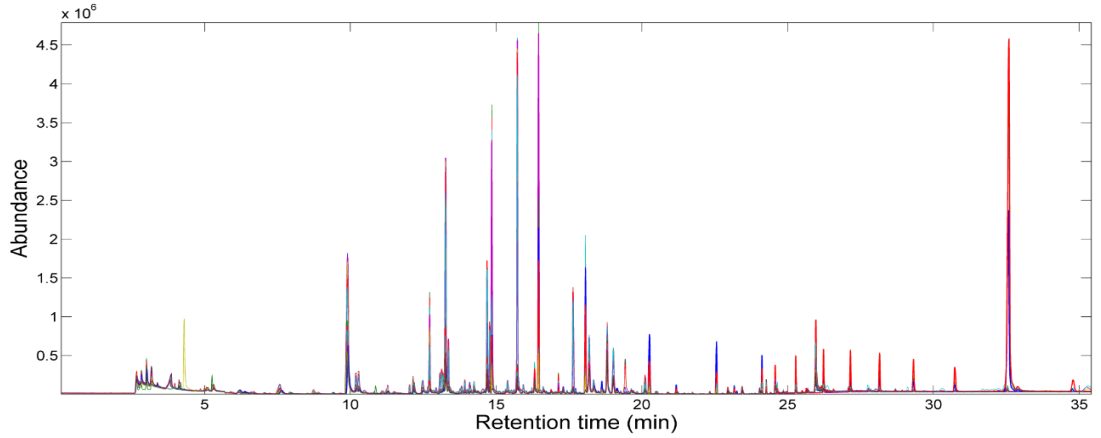


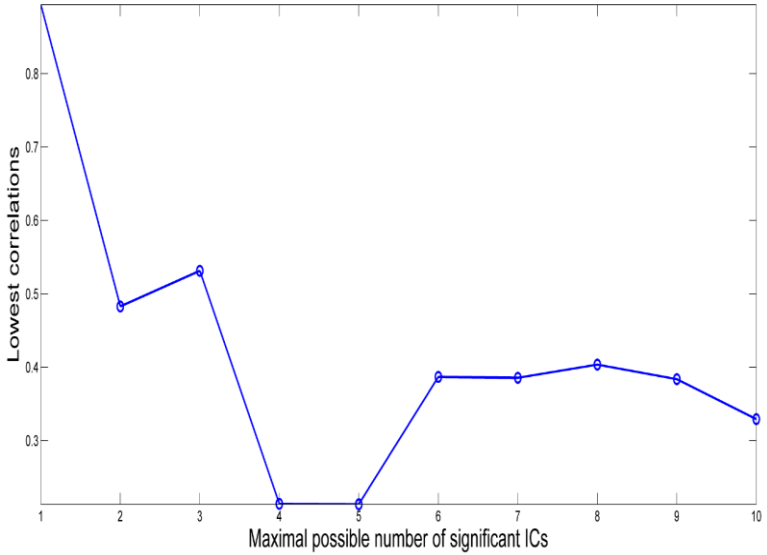
ICA_corr_y

Application: données expérimentales

Kassouf et al. *Talanta. 2018*

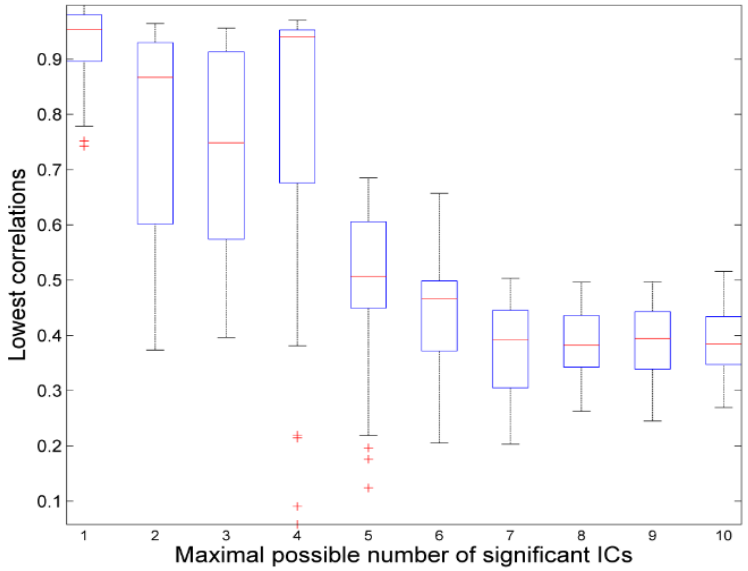
Donnée HS-SPME/GC-MS sur huile d'olive
X (21,5453)





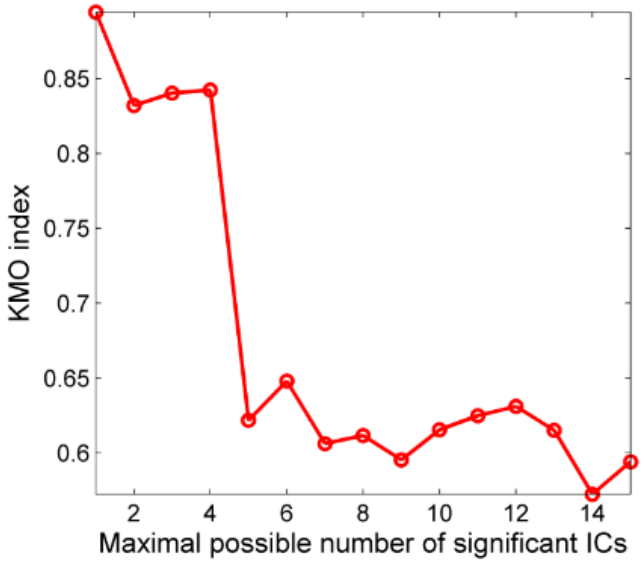
ICA_by_blocks

(B=2; Fmax=10)



Random_ICA

(B=2; Fmax=10); k=50

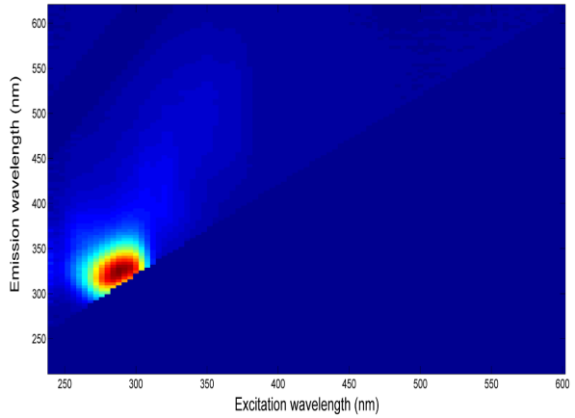


KMO_ICA_Residuals

Fmax=15;

Application: données expérimentales

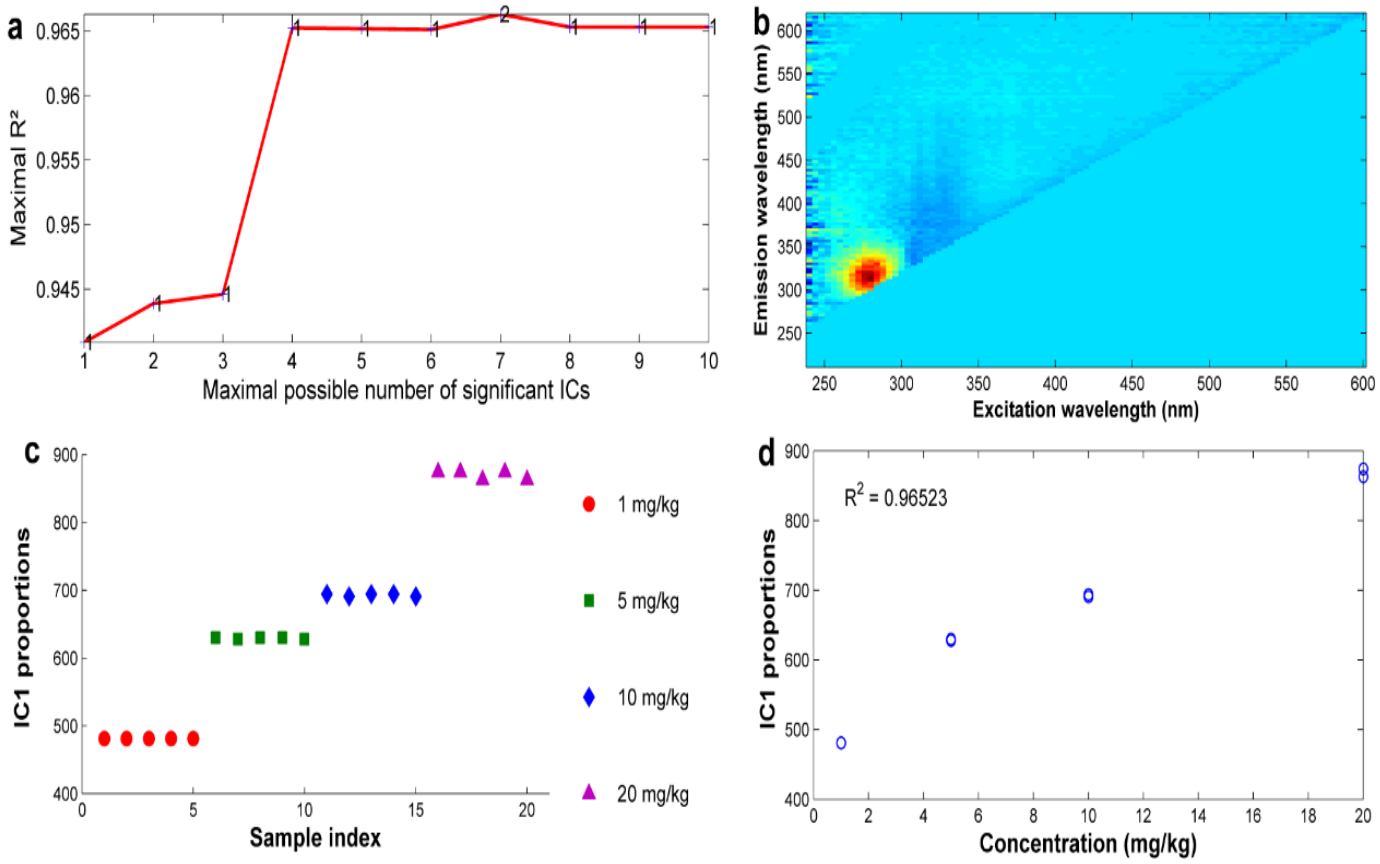
Kassouf et al. *Talanta. 2018*



Donnée fluorescence frontale 3D / Huiles d'olive dopées par des concentrations croissantes d'Irganox 1010
X (20,11375)

Vecteur ***yc***
Concentrations mg/kg Irg1010

yc
1
5
10
20



ICA_corr_y

Conclusion

- La détermination du nombre de composantes indépendantes est un facteur essentiel

- Il existe plusieurs méthodes complémentaires:

    - Le critère de Durbin-Watson permet de voir le nombre de signaux source dans chaque signal; MAIS il n'est applicable qu'aux données structurées
    - La méthode Random_ICA est plus performante qu'ICA_by_blocks
    - La méthode KMO_ICA_Residuals donne des résultats similaires à Random_ICA
    - La méthode ICA_corr_y est utile quand une variable connue existe

- Toutes ces méthodes peuvent être appliquées à l'Analyse en Composantes Principales

Remerciements

Merci pour votre attention

delphine.bouveresse@agroparistech.fr

aminekassouf@hotmail.com

rutledge@agroparistech.fr

Multi_ICA_corr_y

Lignin data (58 × 260)