Approche multiblocs en spectroscopie vibrationnelle

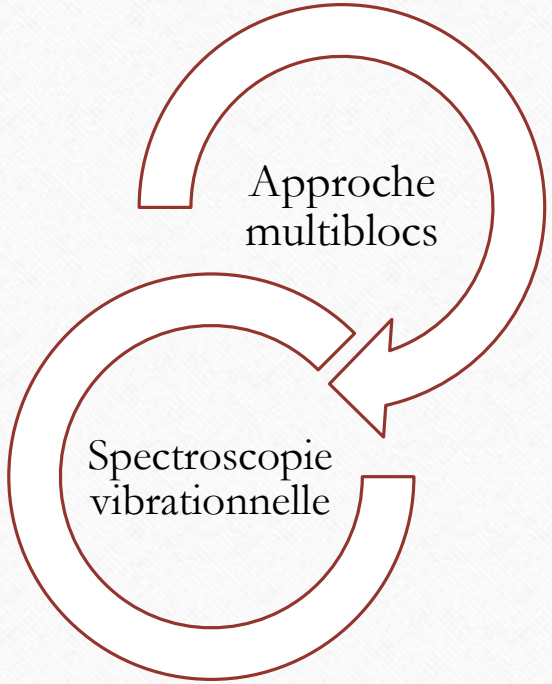
Benoît Jaillais – Mohamed Hanafi

StatSC, ONIRIS-INRAE, Nantes

Sommaire

Avant propos

- Présence des données multiblocs en spectroscopie vibrationnelle.
- Disponibilité d'outils et de méthodes multiblocs.
- Interaction possible.
- Apports réciproques.

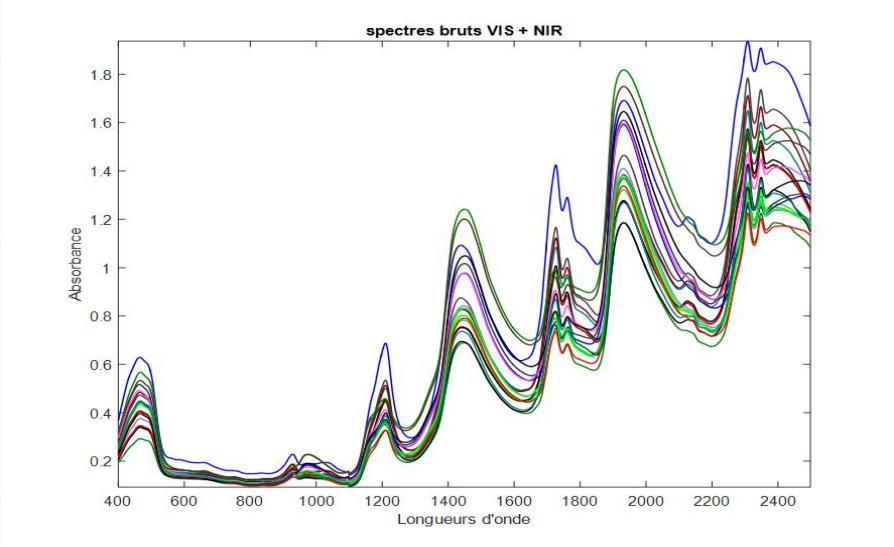


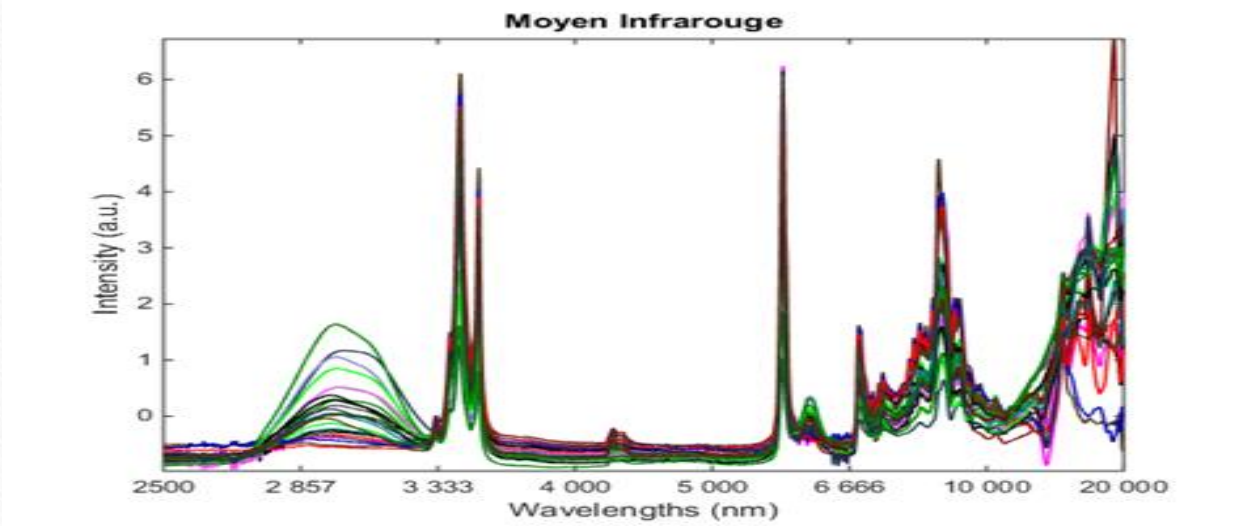
PARTIE 1

Les données multiblocs en spectroscopie vibrationnelle à partir d'un exemple

Contexte

- Caractérisation de la typicité des produits alimentaires (Projet Franco-Irlandais)
- Produits gras - des beurres et des margarines
- Différence entre produits : type de la matière grasse : (animale : lait/végétale : huile) + procédé de fabrication + teneur en matière grasse + leur origine géographique,
- 2 spectromètres : Visible-proche infrarouge (VIS-NIR) et moyen infrarouge (MIR).

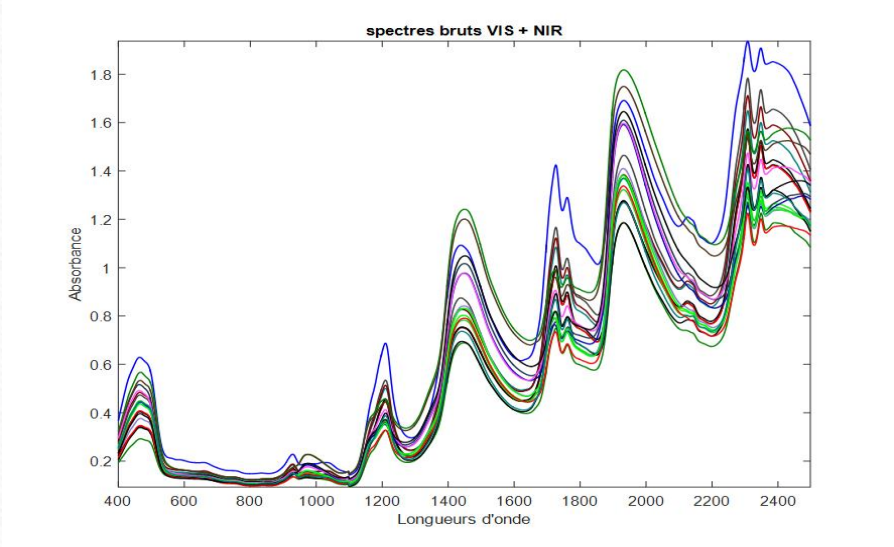


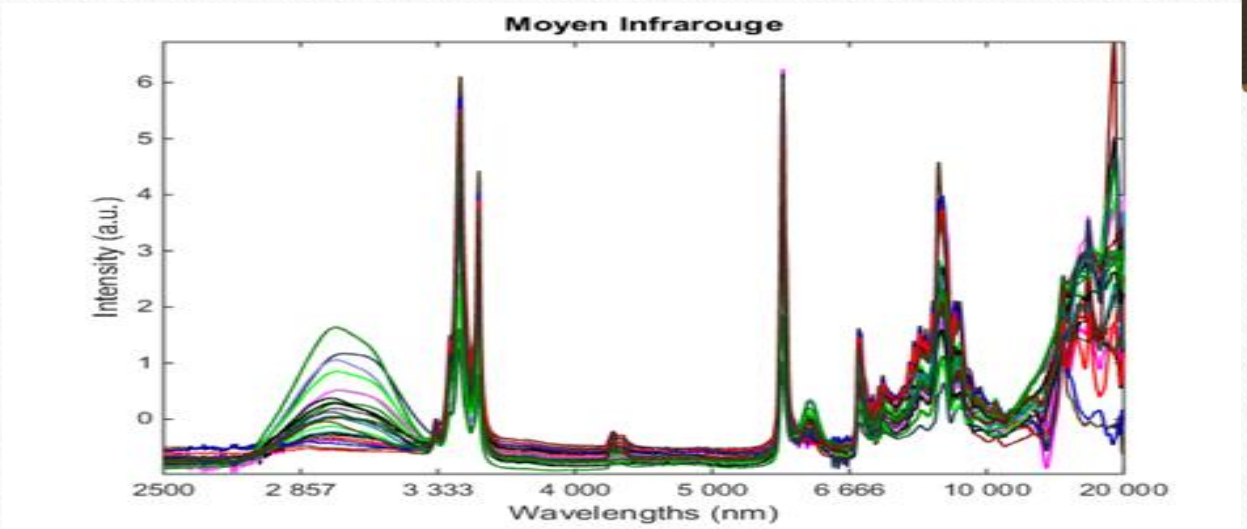


Jaillais B, Morrin V, Downey G (2007) Image processing of outer-product matrices - A new way to classify samples. Examples using visible/NIR/MIR spectral data. Chemometrics and Intelligent. Laboratory Systems, 86, 179-188.

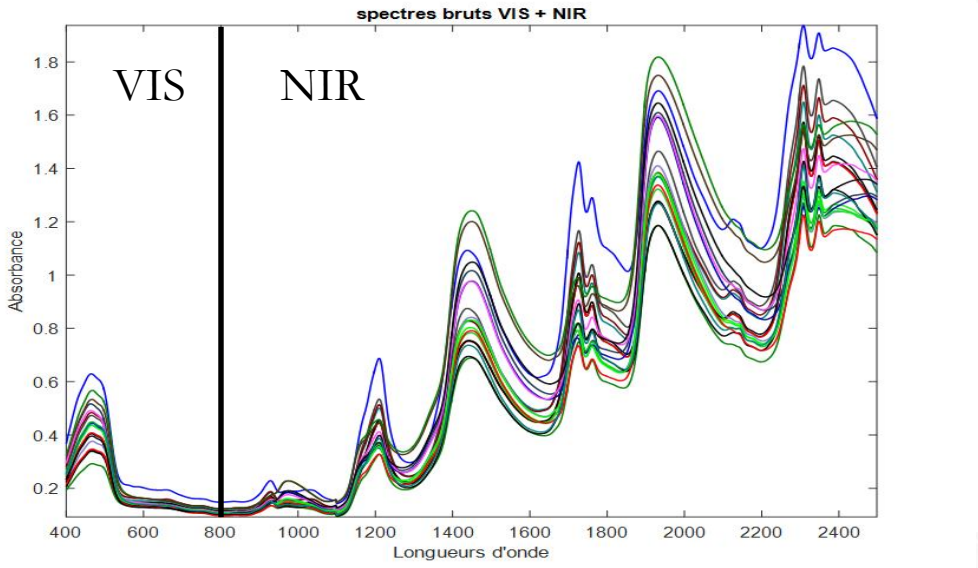
Pourquoi une double acquisition ?

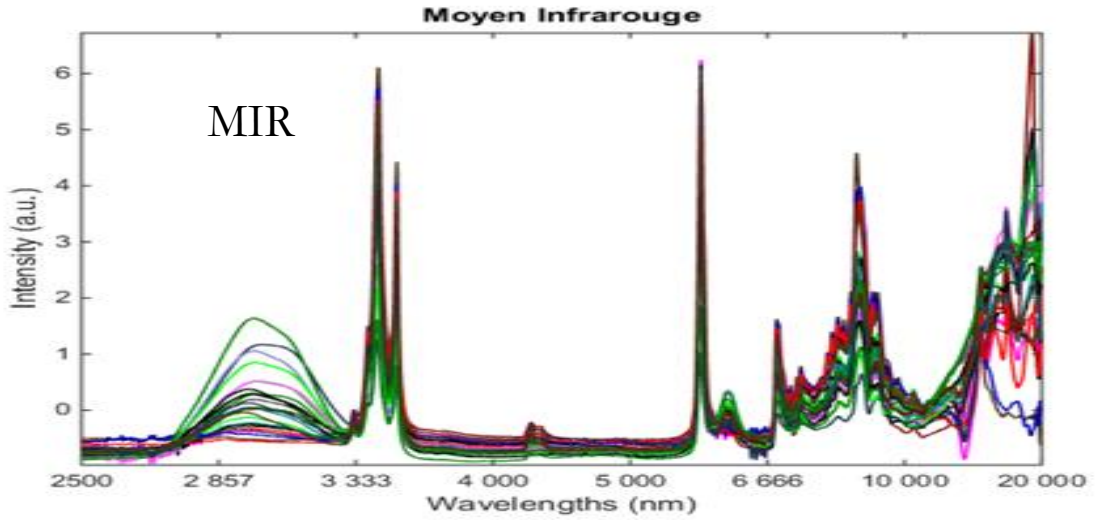
- Comparer l'apport informatif de deux instruments pour caractériser les échantillons.
- Les mécanismes mis en jeu sont différents. Dans le visible : excitation des électrons qui changent d'état, et dans le NIR (variation des fréquences de vibration des liaisons chimiques).
- Fréquences de vibrations fondamentales (MIR)
- Ces spectroscopies diffèrent par les interactions rayonnement-matière mis en jeu, ce qui ne conduit pas au même résultat au niveau de l'interprétation chimique.
- Dans le domaine du MIR, on s'intéresse aux fréquences de vibrations fondamentales, alors que dans le NIR on a accès à des combinaisons de ces fréquences ou à des multiplications par un entier.



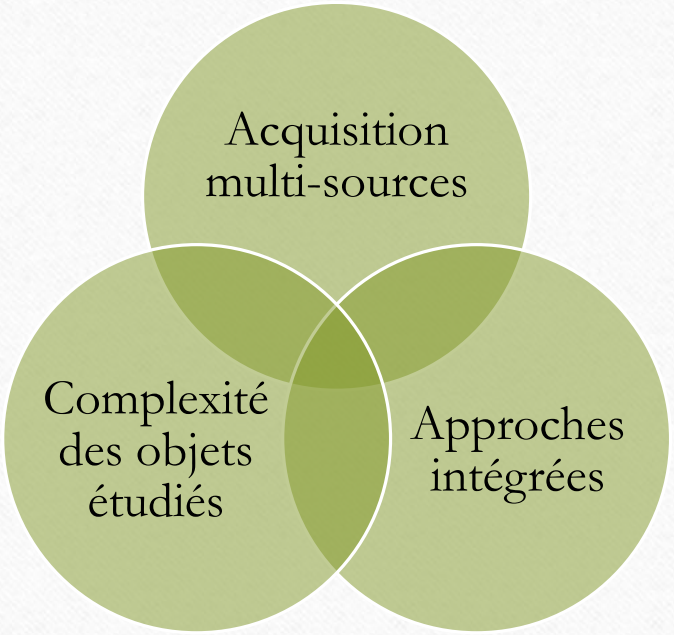


Trois blocs de données






DISCUSSION
TiME

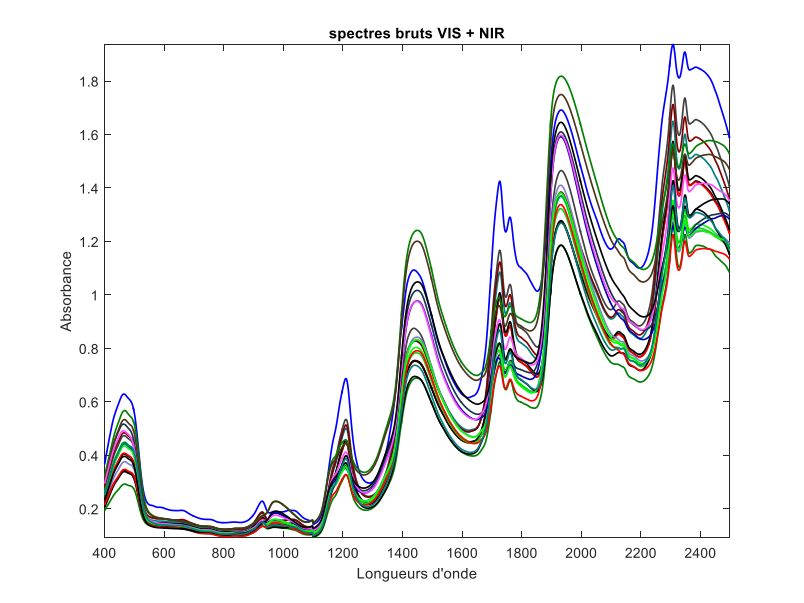

Acquisition multi-sources
Complexité des objets étudiés
Approches intégrées

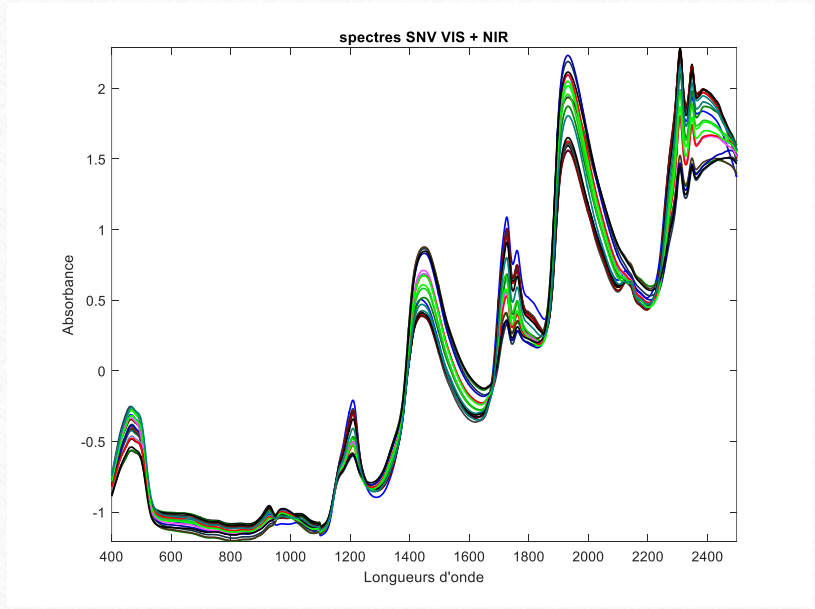
PARTIE 2.
Principes et méthodes

A. Réduction de la dimensionnalité de données spectrales monobloc par ACP

- A1. Un exemple l'ACP

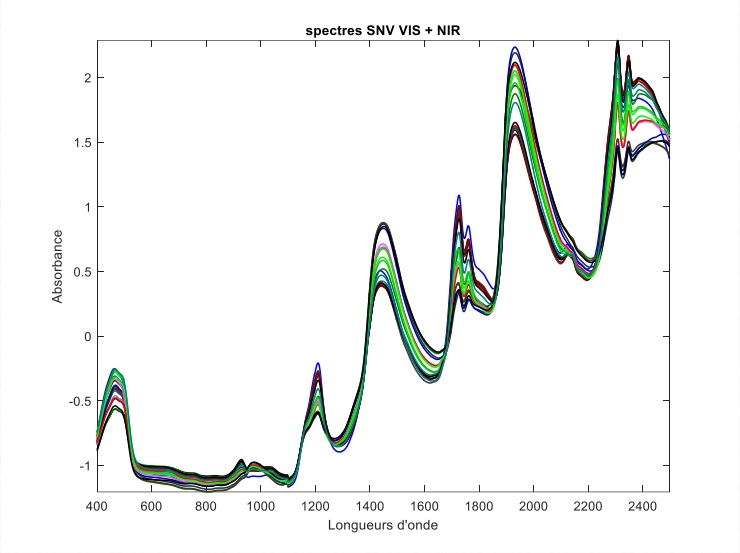
Prétraitement SNV « global »





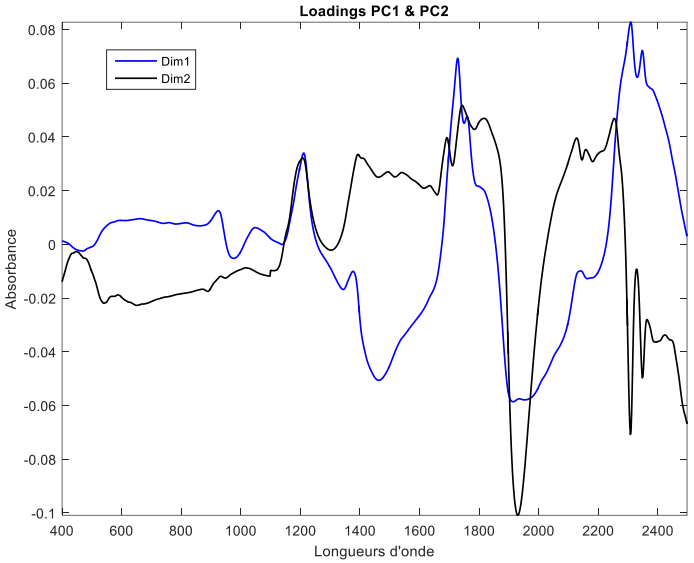
SNV = centrage et réduction de chaque spectre

Réduction de la dimensionnalité



Collection de n spectres

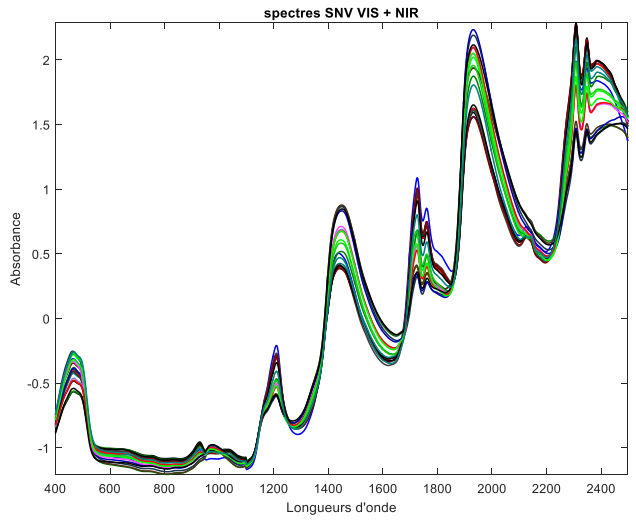
Réduction

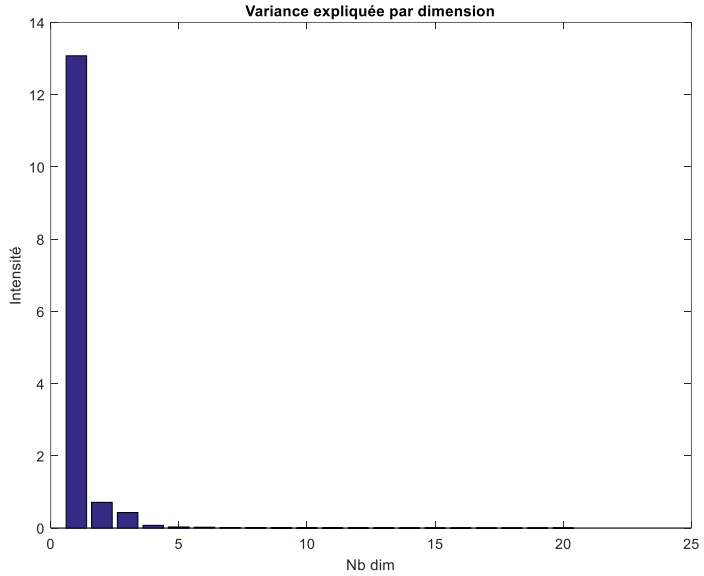


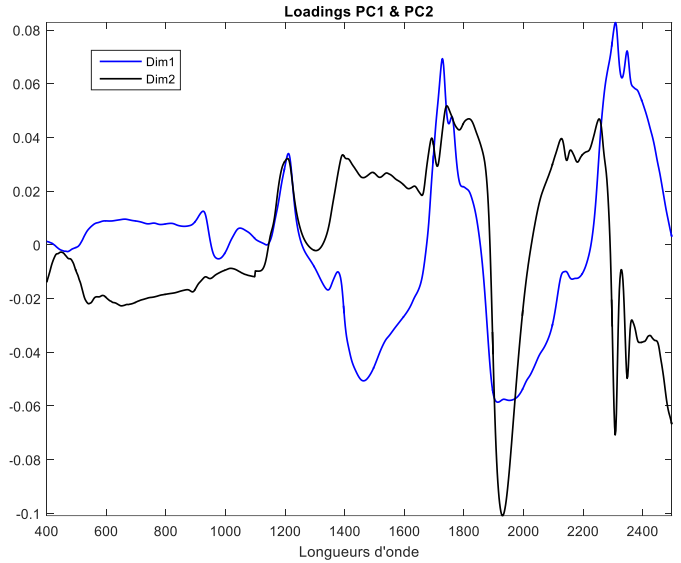
Deux ou plusieurs profils spectraux (<n)
Résumé de la collection de n spectres

Profil spectral = combinaison linéaire des spectres de la collection
Profil spectral = loading

Perte d'information

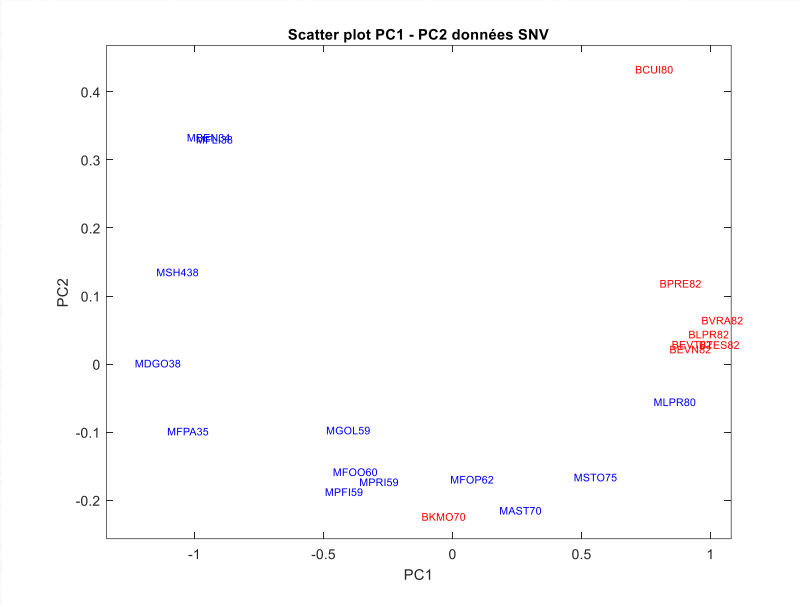


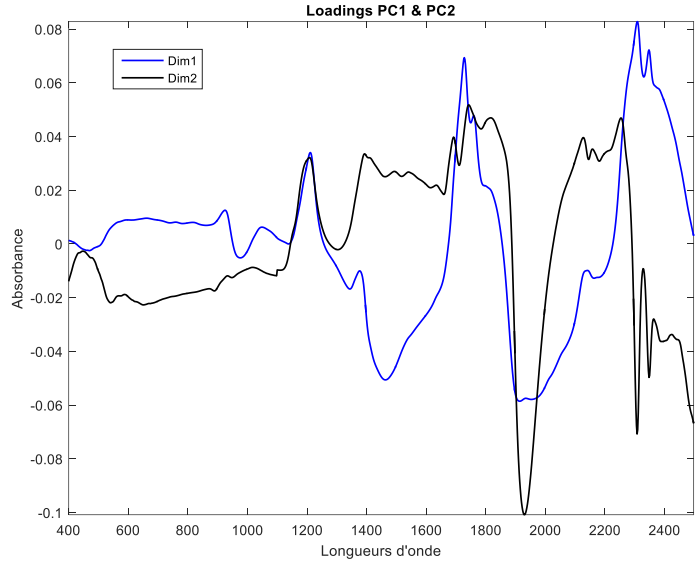




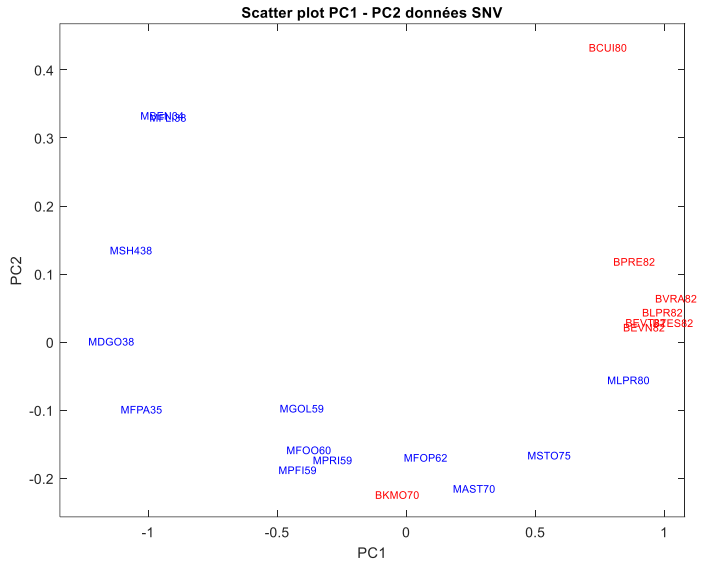
- les spectres sont souvent fortement corrélés.
- Variances expliquées des profils spectraux

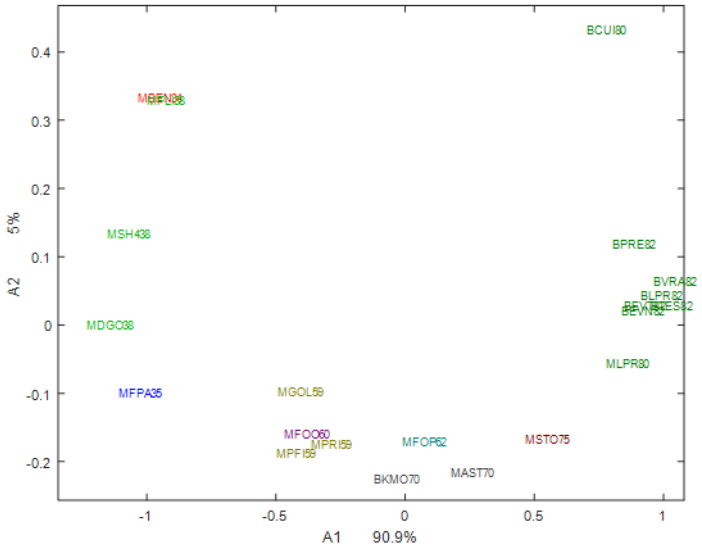
Typologies des spectres

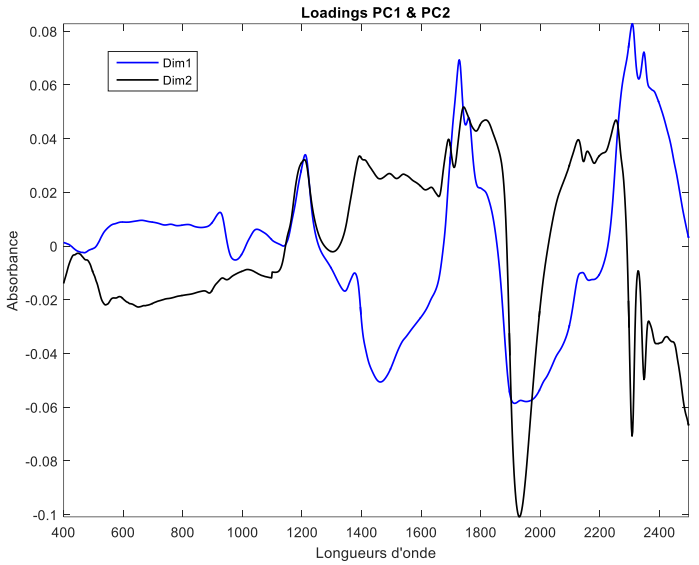




- Corrélations (angles) des spectres de la collection de spectres avec les profils spectraux
- Carte factorielle des composantes principales

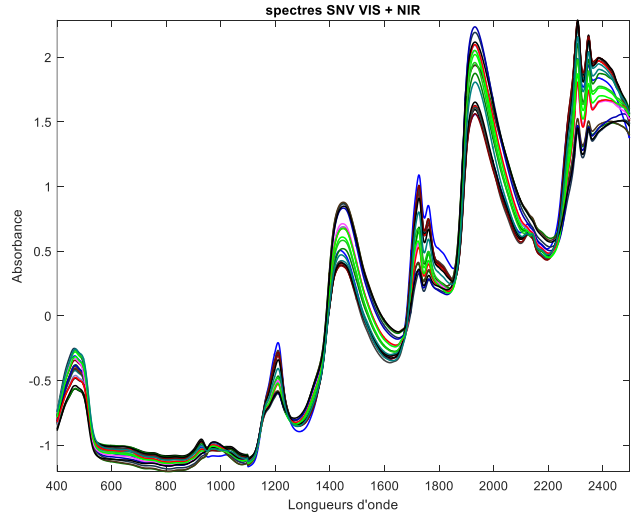


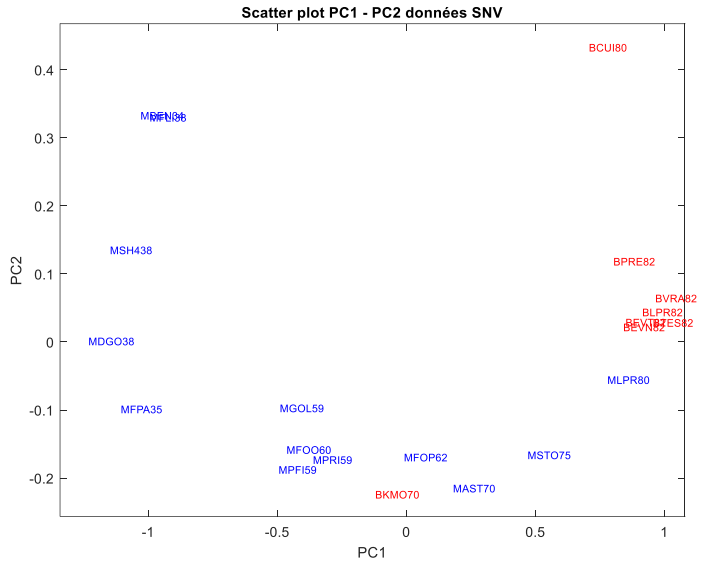


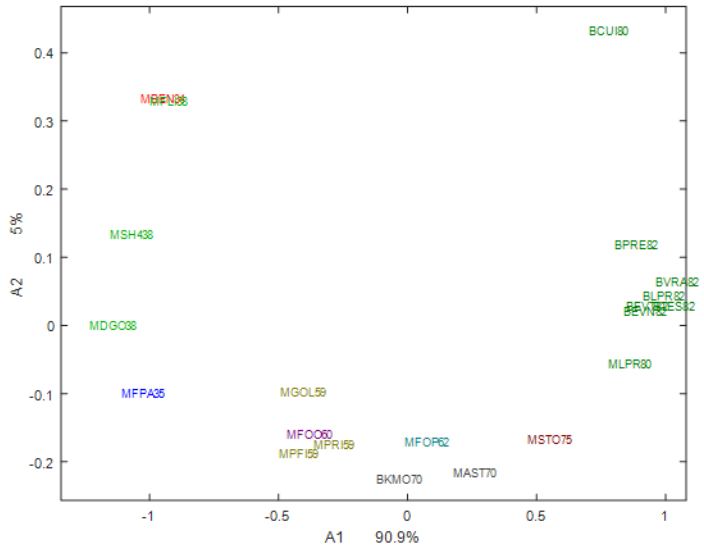


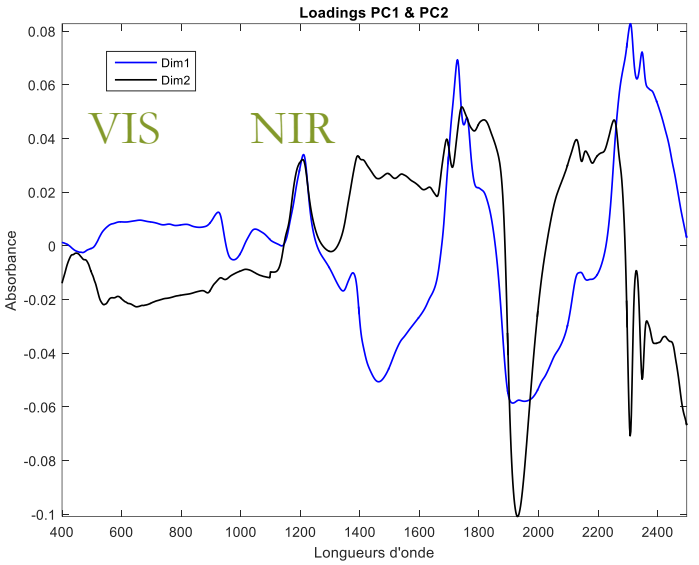
Interprétation

Le profil spectral 1 est relié à la teneur en matière grasse des échantillons. La partie positive est constitué de longueurs d'ondes caractéristiques de la matière grasse, et la partie négative par des longueurs d'ondes caractéristiques de l'eau liée.



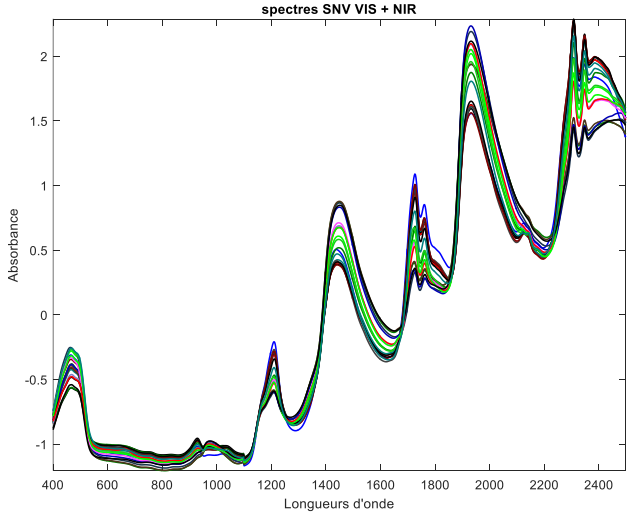






Interprétation

Le profil spectral 2 serait lié à la teneur en acides gras avec des doubles liaisons conjuguées pour la partie positive. Les acides gras correspondant à la partie négative de l'axe 2 sont plutôt monoinsaturés (2308, 2348 nm), et donc insolubles à l'eau (eau libre à 1930 nm).



PARTIE 2.
Principes et méthodes

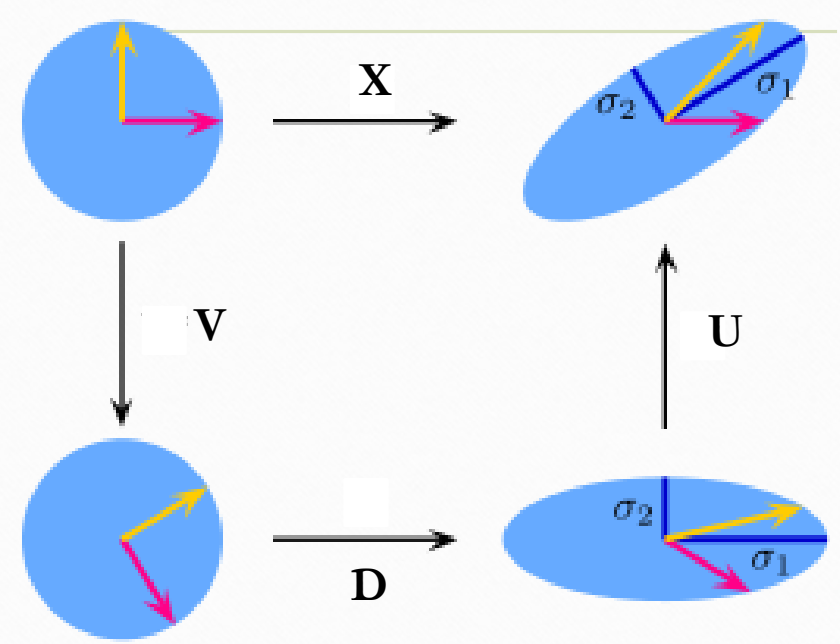
- A2. La parenthèse géométrique (ACP Inside)

Deux formes géométriques courantes

Une sphère

Une ellipsoïde

Un théorème (Singular Value Decomposition)



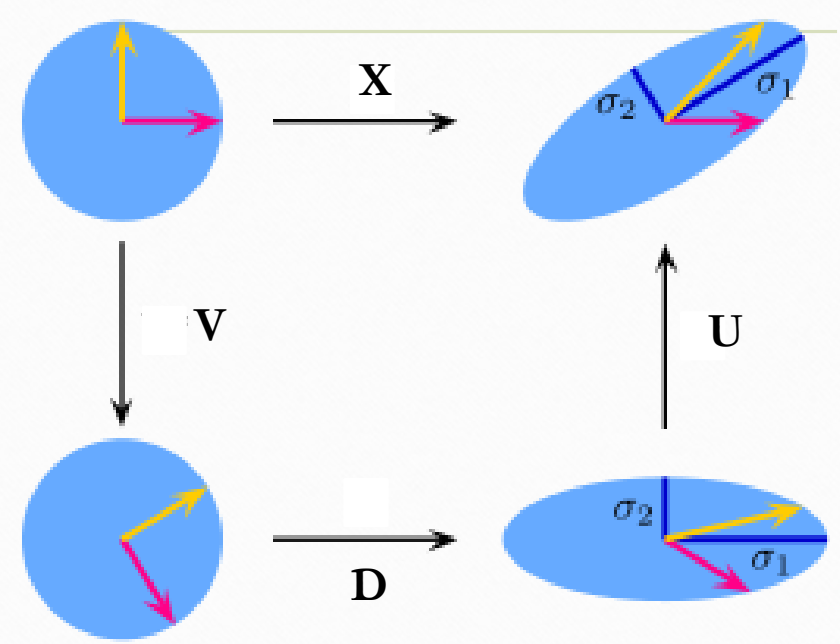
https://fr.wikipedia.org/

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

Un théorème qui établit que les profils spectraux d'un tableau forme une ellipse située dans un espace multidimensionnel (dimension p)

Profil spectral = combinaison linéaire de spectres

Un théorème (Singular Value Decomposition)



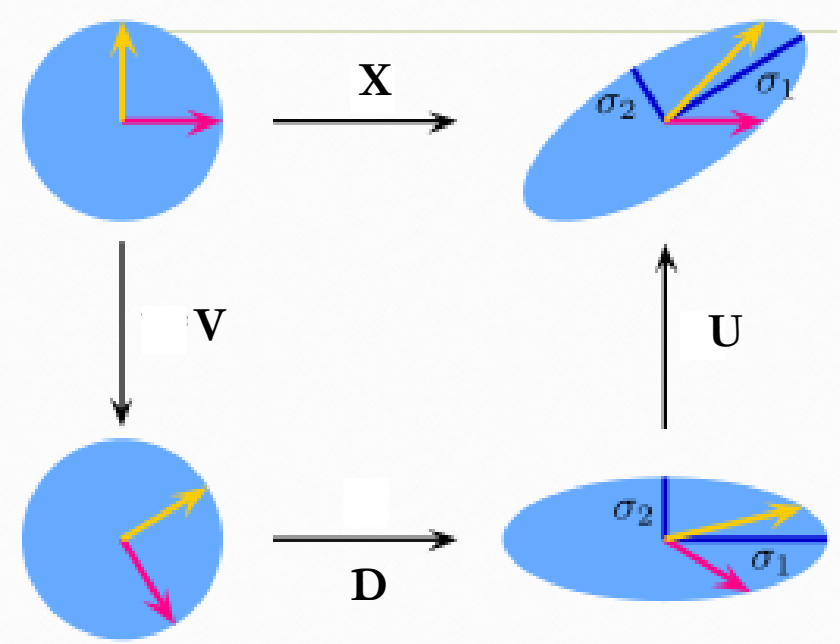
https://fr.wikipedia.org/

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

Un théorème qui établit que les profils spectraux d'un tableau forme une ellipse située dans un espace multidimensionnel (dimension p)

Profil spectral = combinaison linéaire de spectres

Un théorème (Singular Value Decomposition)



https://fr.wikipedia.org/

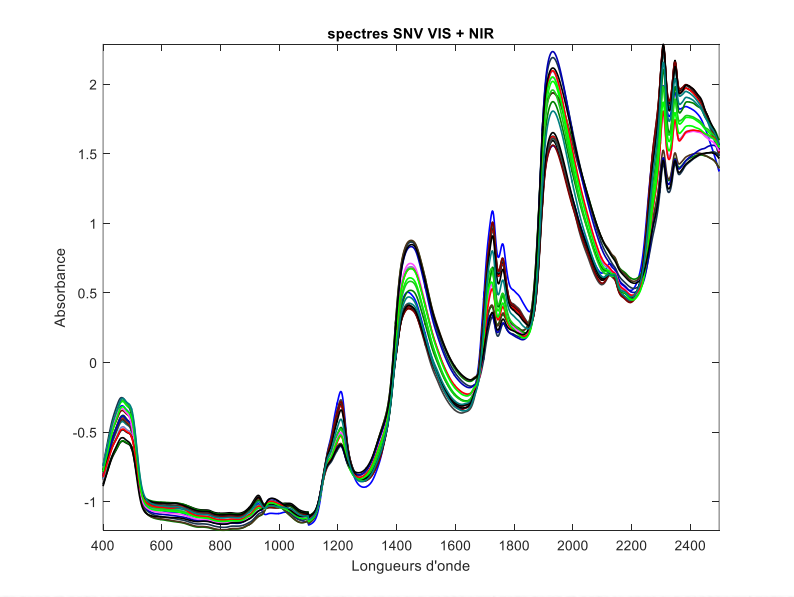
$$\mathbf{X} = \mathbf{V}\,\mathbf{D}\,\mathbf{U}^{T}$$

Un théorème qui établit que les profils spectraux d'un tableau forme une ellipse située dans un espace multidimensionnel (dimension p)

Profil spectral = combinaison linéaire de spectres

Le point de vue euclidien

*une longueur d'onde est un point (vecteur) d'un espace de dimension **n***



un spectre est un point (vecteur) d'un espace de dimension p

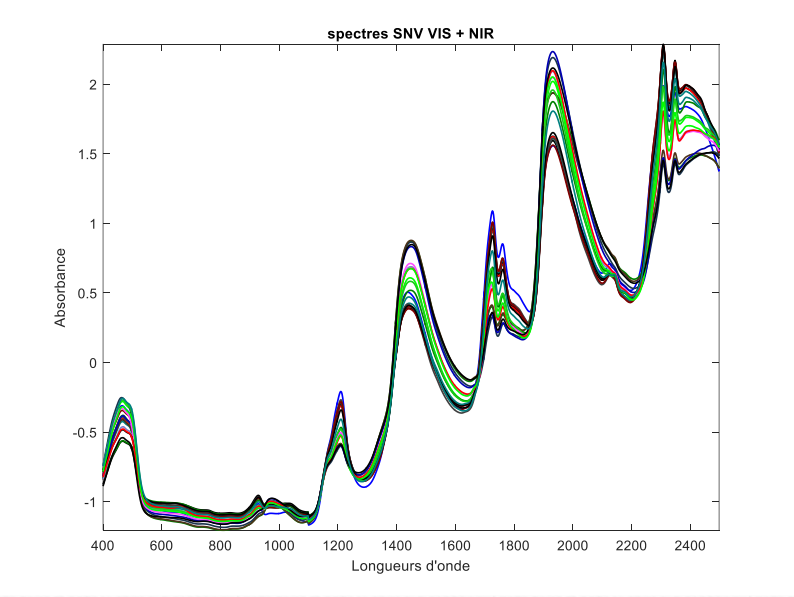
Un collection de spectres déforme un cercle en une ellipse

Les profils spectraux possibles décrivent une ellipse

Les axes de l'ellipse sont les profils spectraux (loadings) de l'ACP

Le point de vue euclidien

*une longueur d'onde est un point (vecteur) d'un espace de dimension **n***



un spectre est un point (vecteur) d'un espace de dimension p

Un collection de spectres déforme un cercle en une ellipse

Les profils spectraux possibles décrivent une ellispe

Les axes de l'ellipse sont les profils spectraux (loadings) de l'ACP

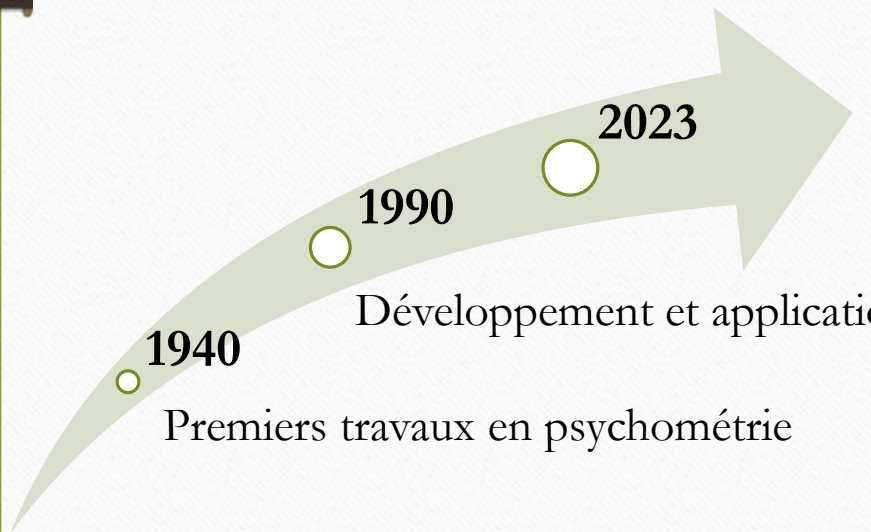

DISCUSSION
TiME

PARTIE 2.
Principes et méthodes

B. Réduction de la dimensionnalité de données multiblocs

- B1. Remarques préalables

La petite histoire…

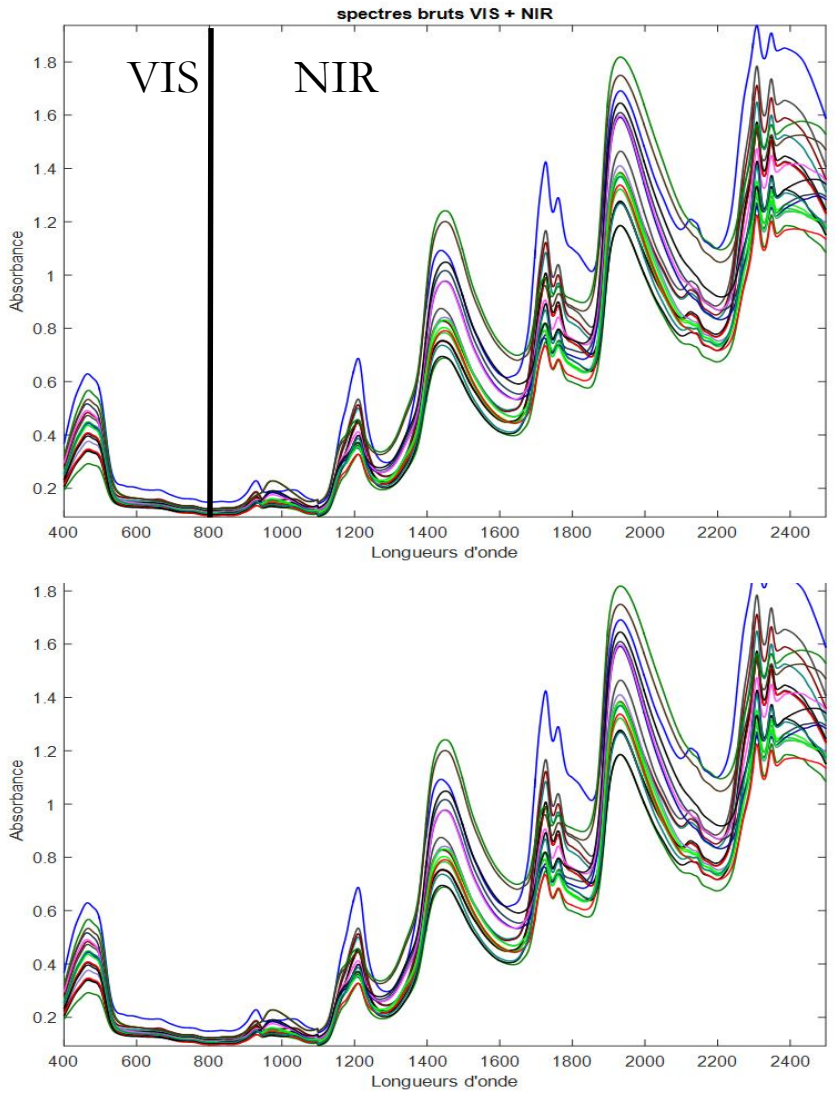


- Abondance et variété des travaux et développements en lien souvent avec des domaines générateurs de données (plateformes analytiques, techniques d'imagerie, essor des données omiques,…)
- Périmètre des méthodes difficile à cerner
- Approche exhaustive des méthodes quasi impossible

Dimensionnalité

- Les données monobloc sont des données multi-blocs où chaque bloc est unidimensionnel (généralisation)
- Les données multiblocs induisent une multidimensionnalité à l'échelle de chaque bloc et à l'échelle de l'ensemble (Une double réduction)

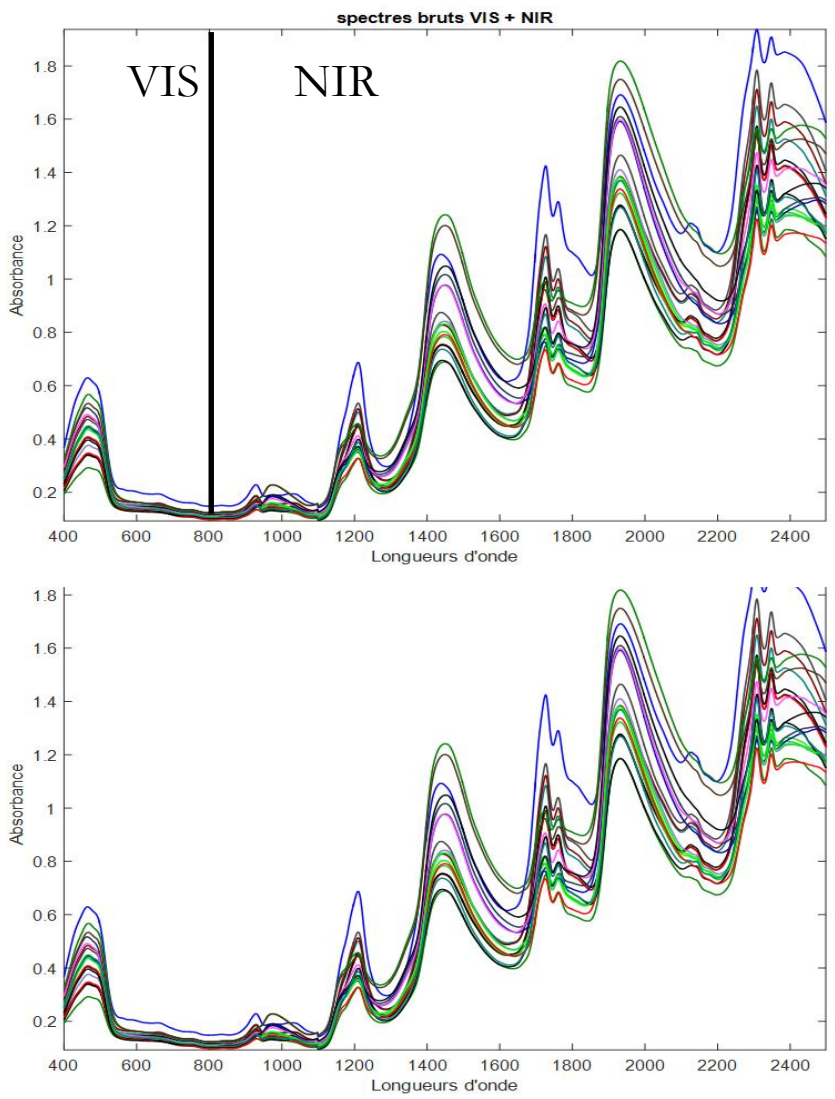
Multiblocs (2 blocs)



Partition

- Données multiblocs sont des monobloc muni d'une partition à priori
- Les méthodes monoblocs sont conçues sans prise en compte de cette partition
- Les méthodes multiblocs sont par essence conçues pour prendre en compte cette partition
- Les méthodes monobloc présentent des limitations pour analyser des données multiblocs.

Multiblocs (2 blocs)

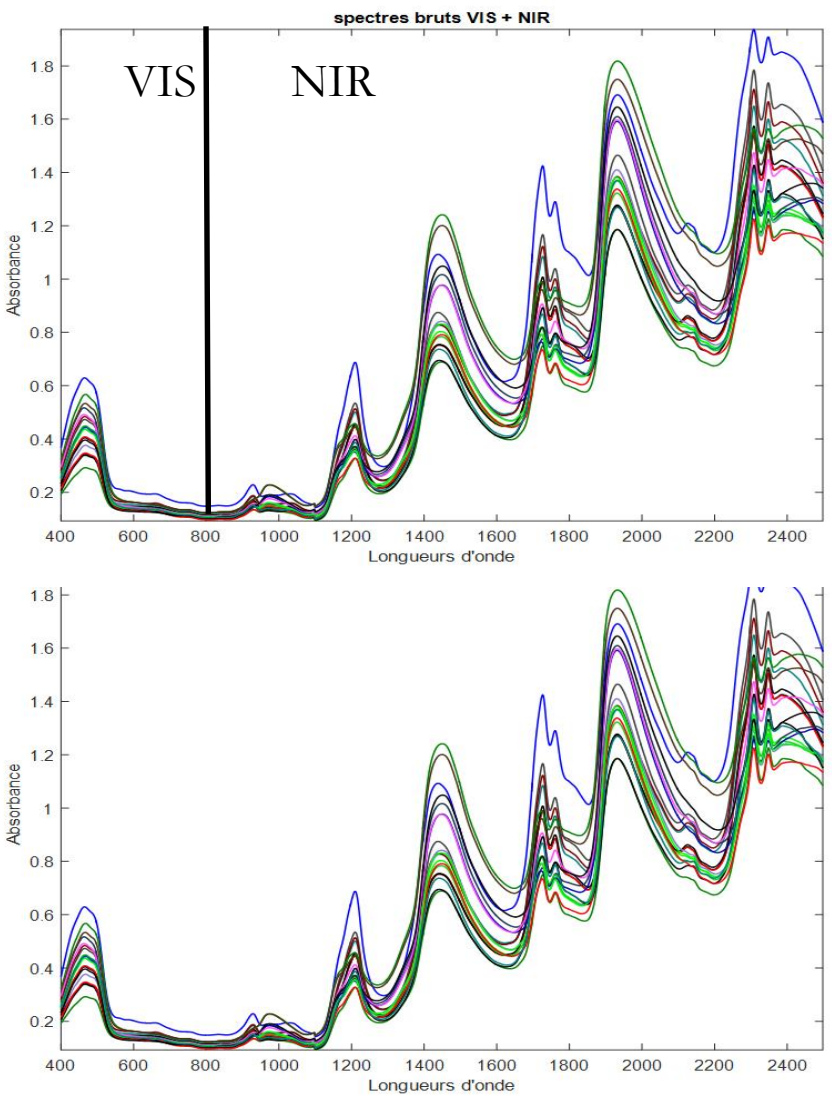


Attendus

- L'attente d'une analyse multiblocs doit permettre une description à l'échelle des blocs

- L'objectif principal est de résumer (compresser l'information) et de savoir comment cette information est reproduite à partir des différents blocs.

Multi-blocs (2 blocs)



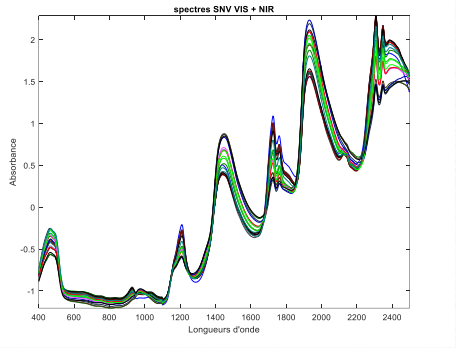
PARTIE 2.
Principes et méthodes

B. Réduction de la dimensionnalité de données multiblocs

- B2. Version multiblocs de l'ACP

Partition de Données versus Partition de modèles

monobloc (1 bloc)

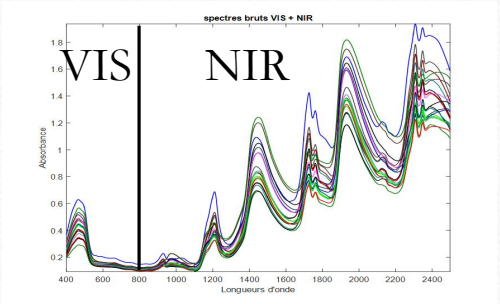


ACP

$$X = U D V^{T}$$

PARTITON

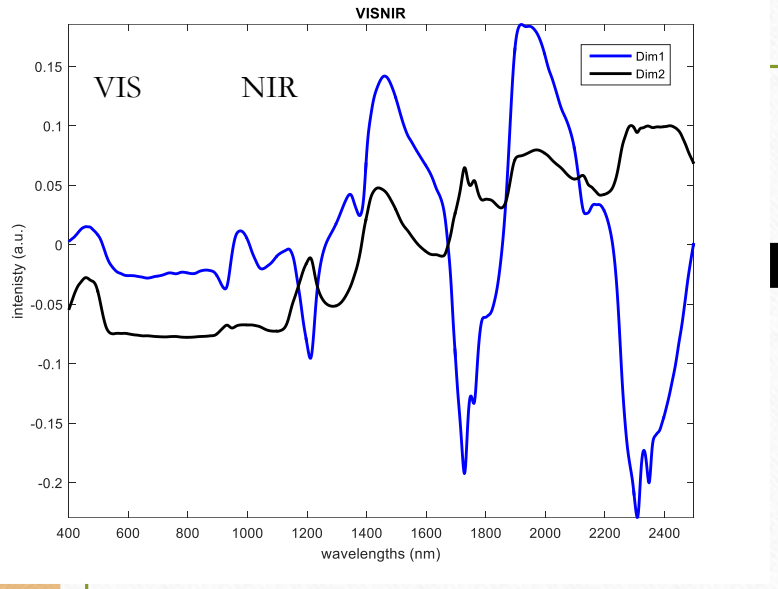
Multiblocs (2 blocs)



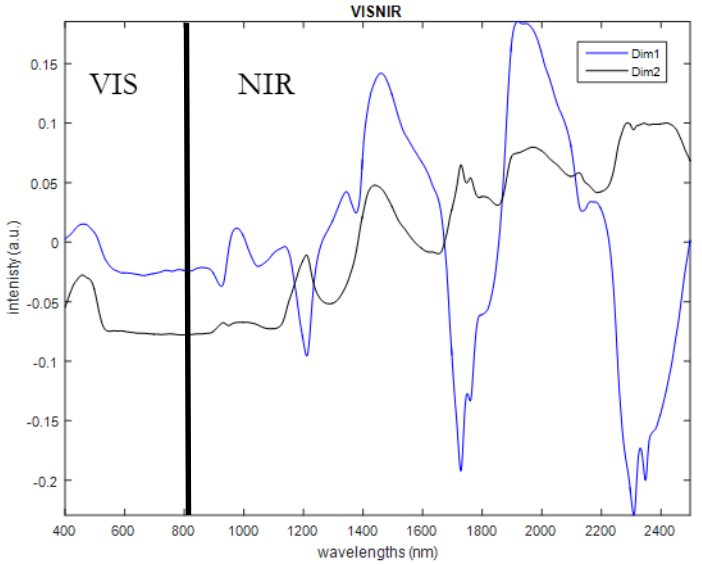
$$VIS = U D_1 V_1^{T}$$

$$NIR = U D_2 V_2^{T}$$

Concrètement : Profils spectraux par bloc

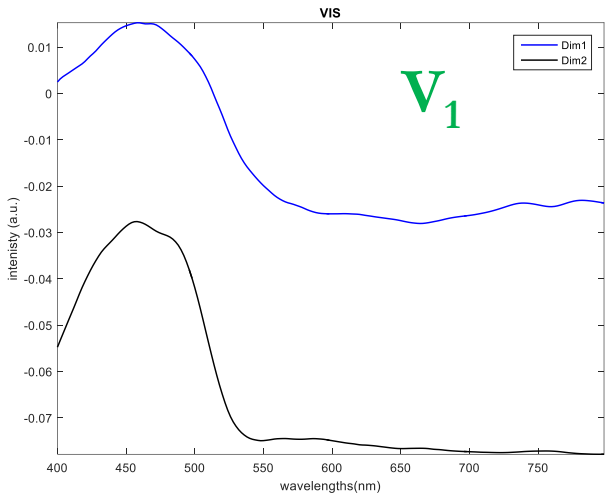


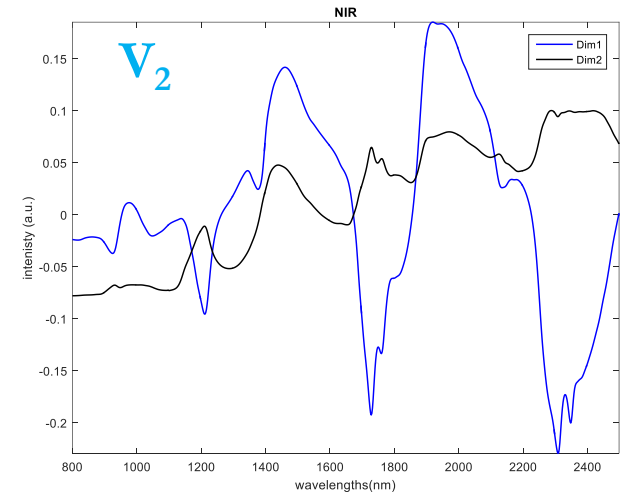
Profils spectraux de l'ACP



Partition

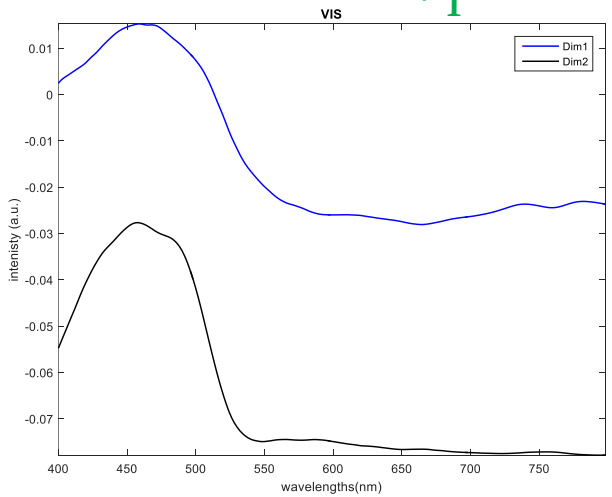
Standardisation

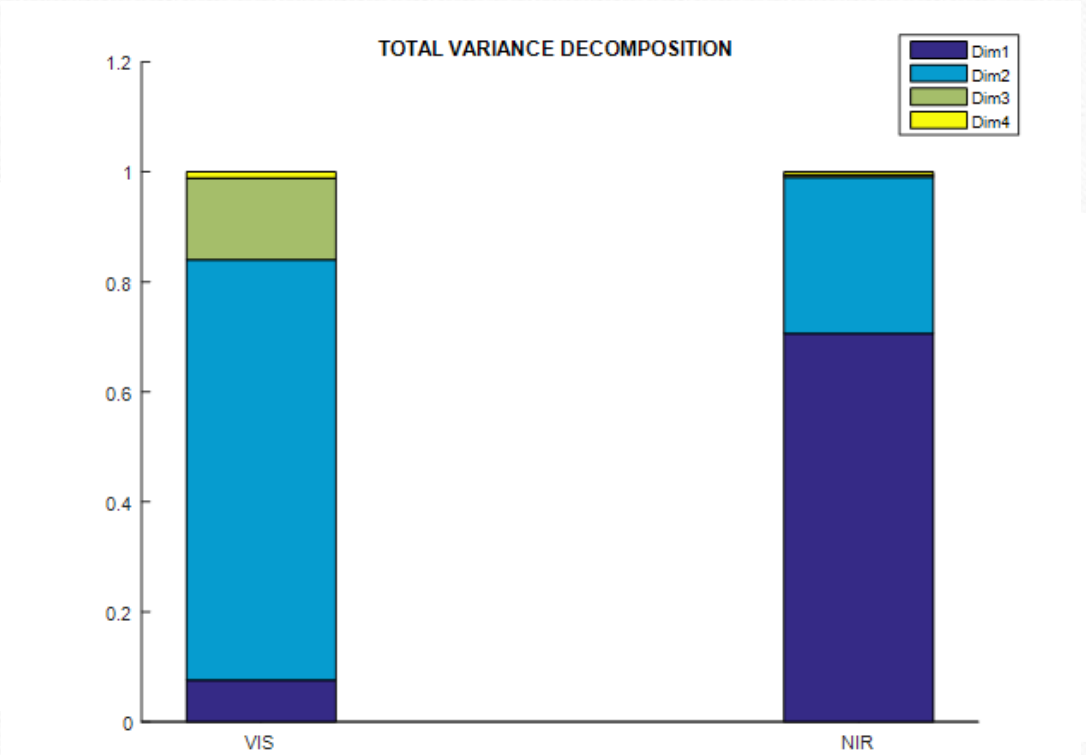




Concrètement : Perte d'information l'échelle de chaque bloc

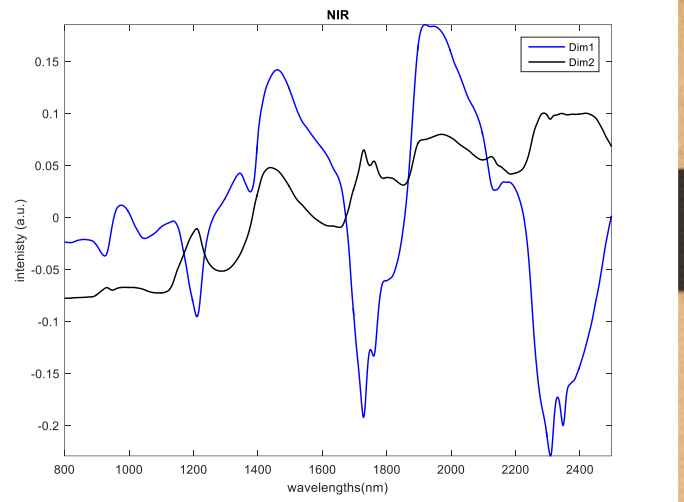
V_1





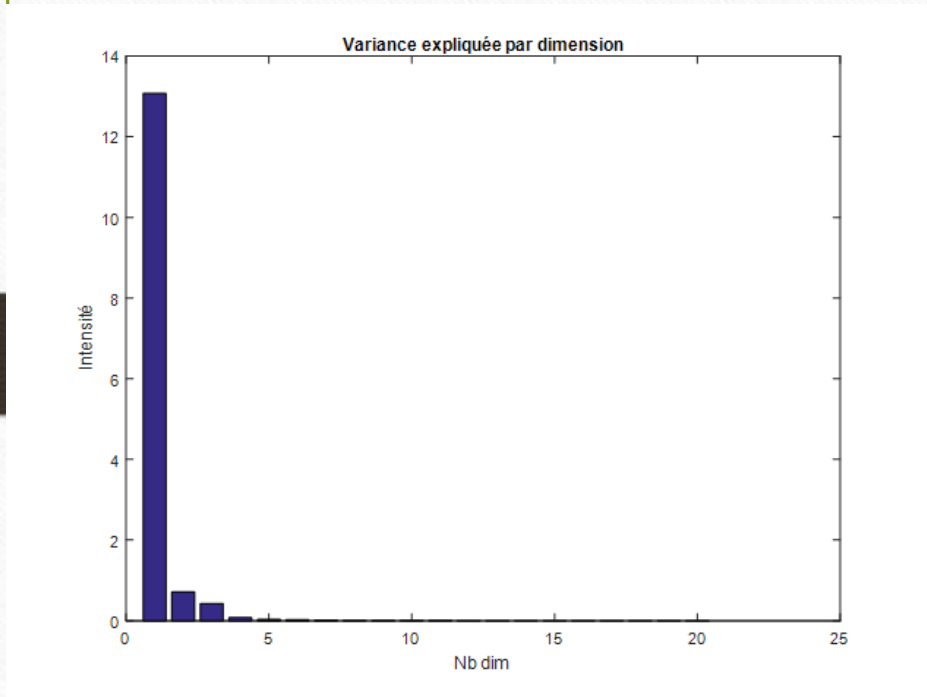
D_1^2 D_2^2

V_2

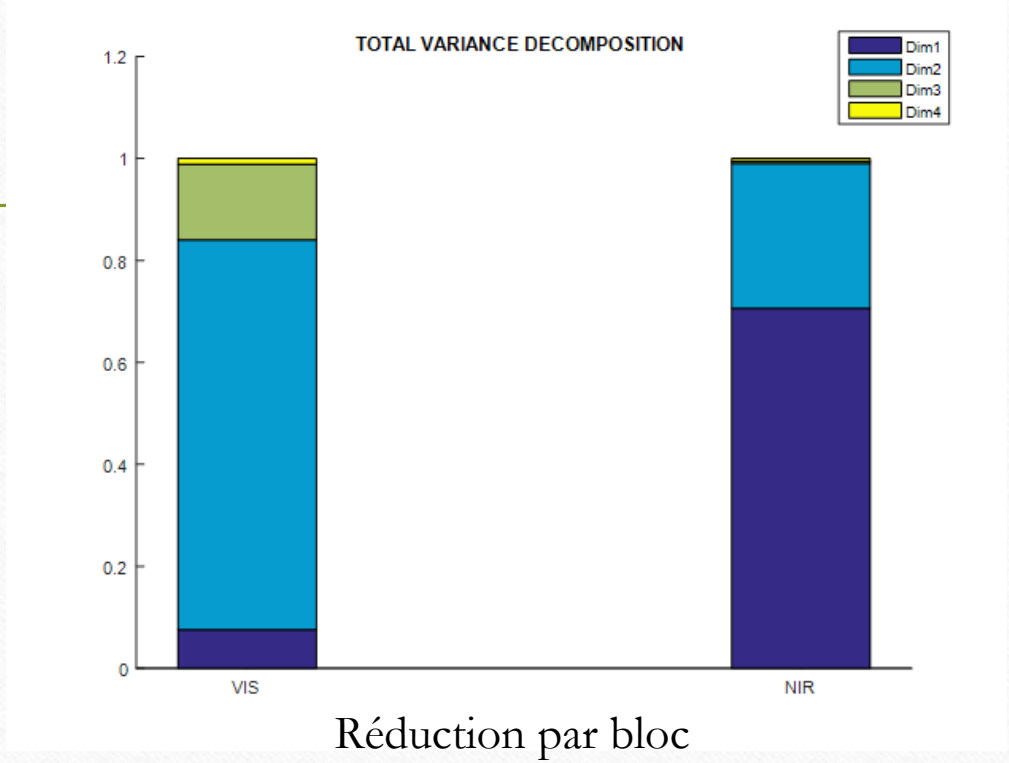


Information commune versus information spécifique
Importance du visible sur la deuxième dimension

Concrètement : Double réduction de dimensionnalité



Réduction de l'ensemble des blocs

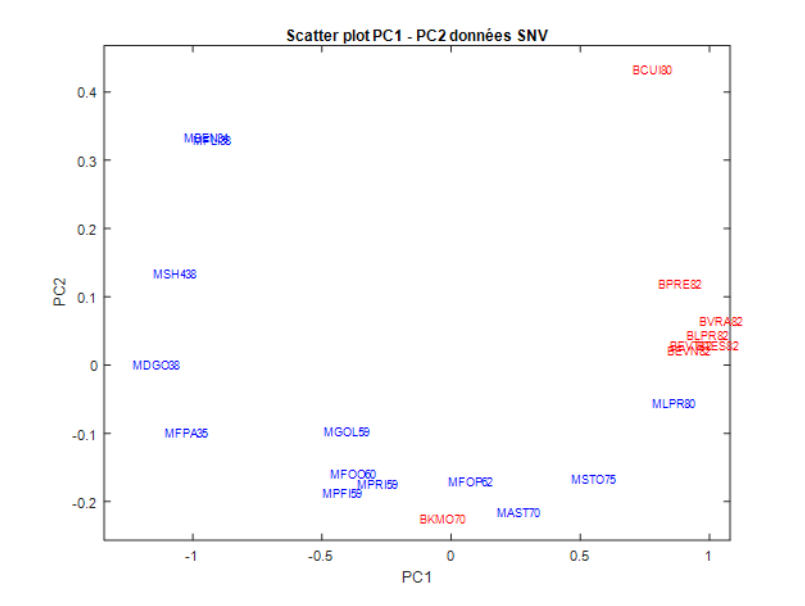


Réduction par bloc

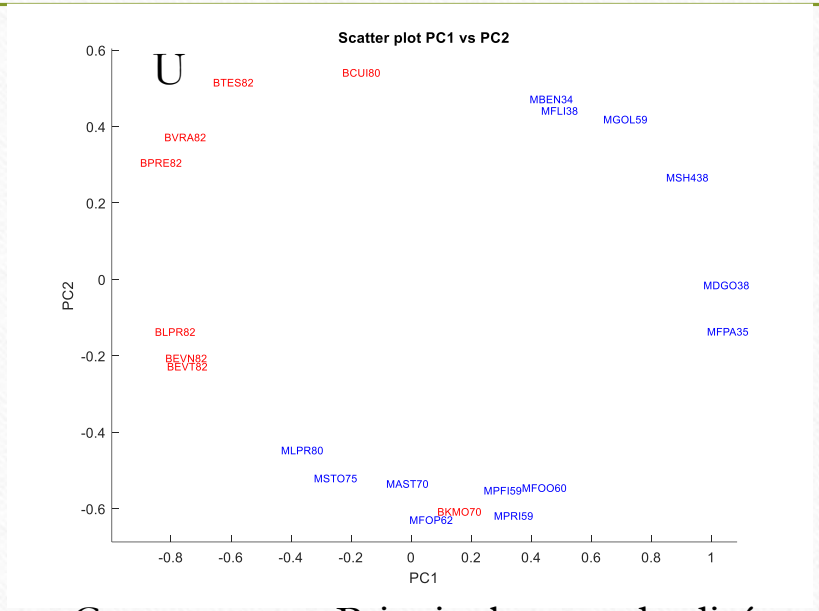
Retour à l'échelle des blocs

Meilleure analyse réalisant une double réduction optimale au sens de la moyenne des variabilités résiduelles de chacun des blocs et de l'ensemble.

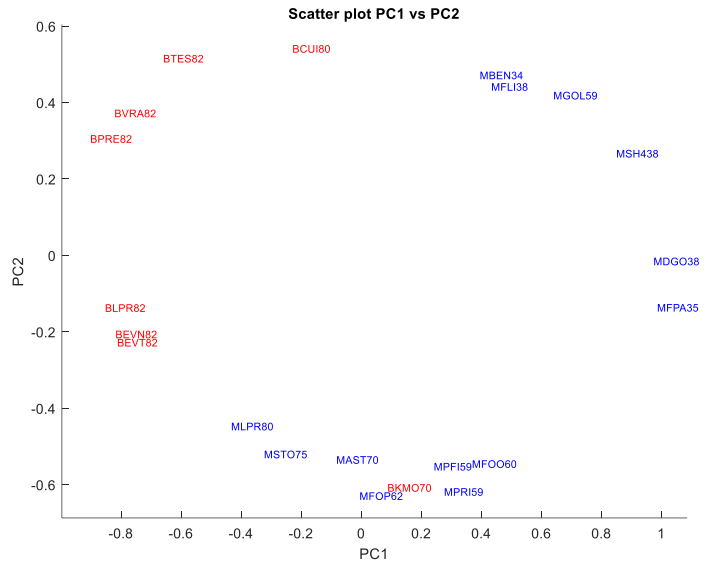
Concrètement : Un référentiel commun entre blocs et l'ensemble

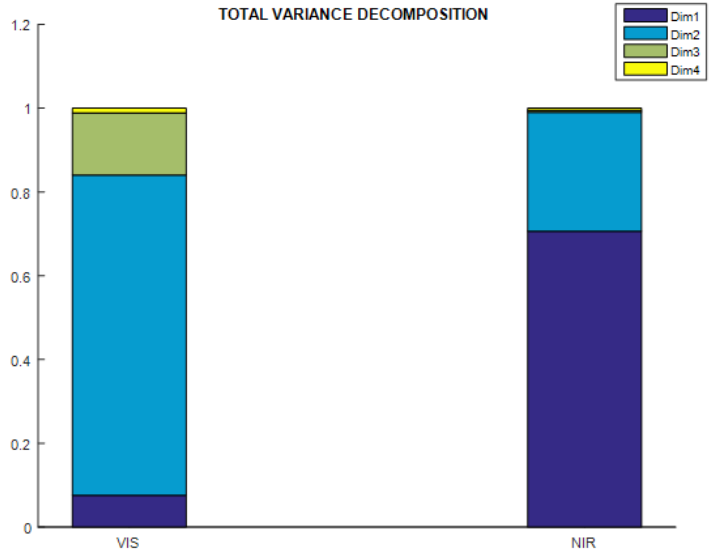


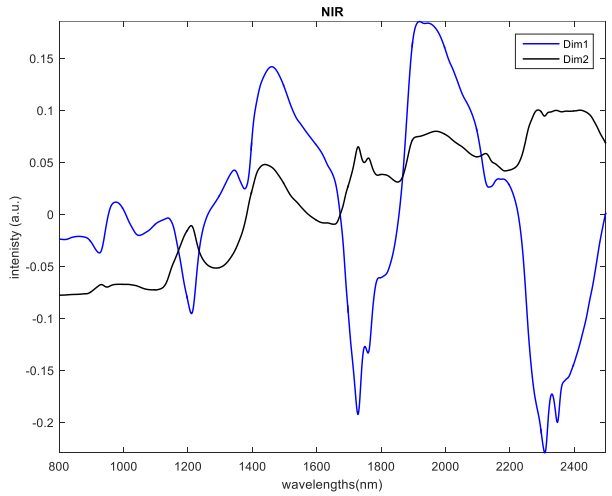
Composantes Principales



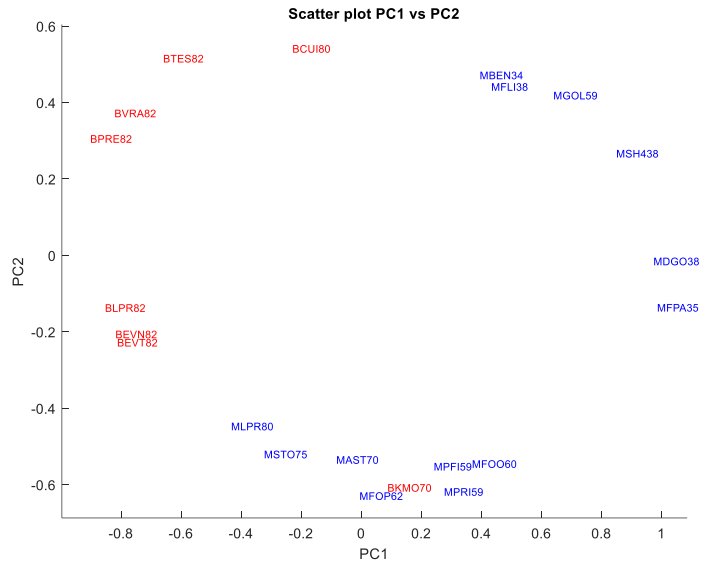
Composantes Principales standardisés
Information de référence
Base de discussion entre blocs

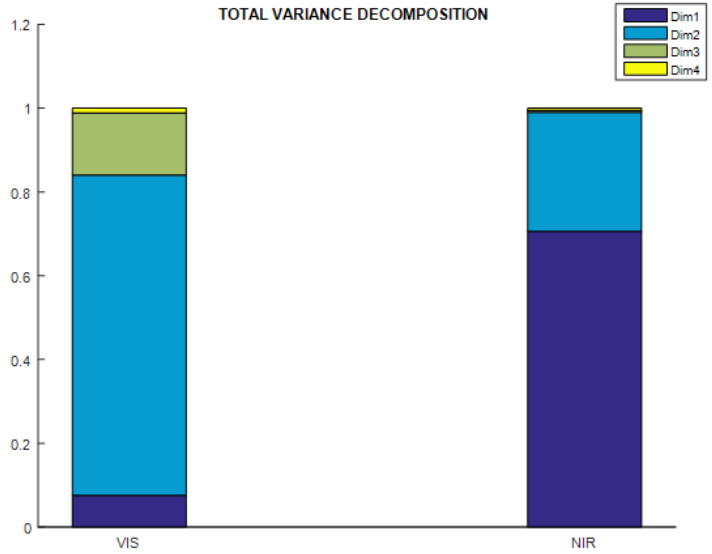


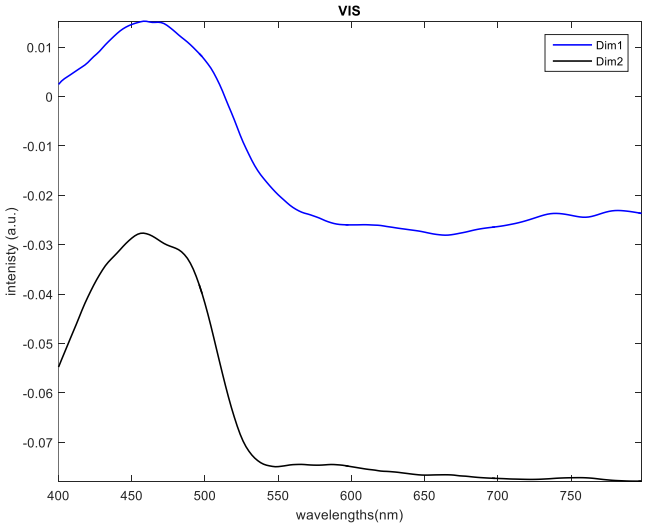




La dimension 1 est majoritairement construite par le bloc NIR et conduit à l'obtention de scores et de loadings similaires à ceux de l'ACP, c-a-d une répartition des échantillons sur l'axe en fonction de la teneur en MG.







Il apparait que la dimension 2 est clairement différente en terme de loadings, puisque le pic à 1930 nm en NIR est absent. Ceci est cohérent avec le fait que cette dimension est principalement synthétisée à partir des variables du bloc Visible. Le loading du bloc visible correspond au spectre du β-carotène, qui est ajouté à la margarines à des doses comprises entre 0.3 et 0.9 mg/100g, et absent naturellement dans les beurres.

Apport par rapport à l'ACP monobloc

1. La version multiblocs de l'ACP a mis en évidence l'importance du bloc du visible

2. Chaque bloc est spécifique (absence de co-variation entre les deux blocs)

3. Base de l'analyse : les composantes principales de l'ensemble des blocs (un référentiel).

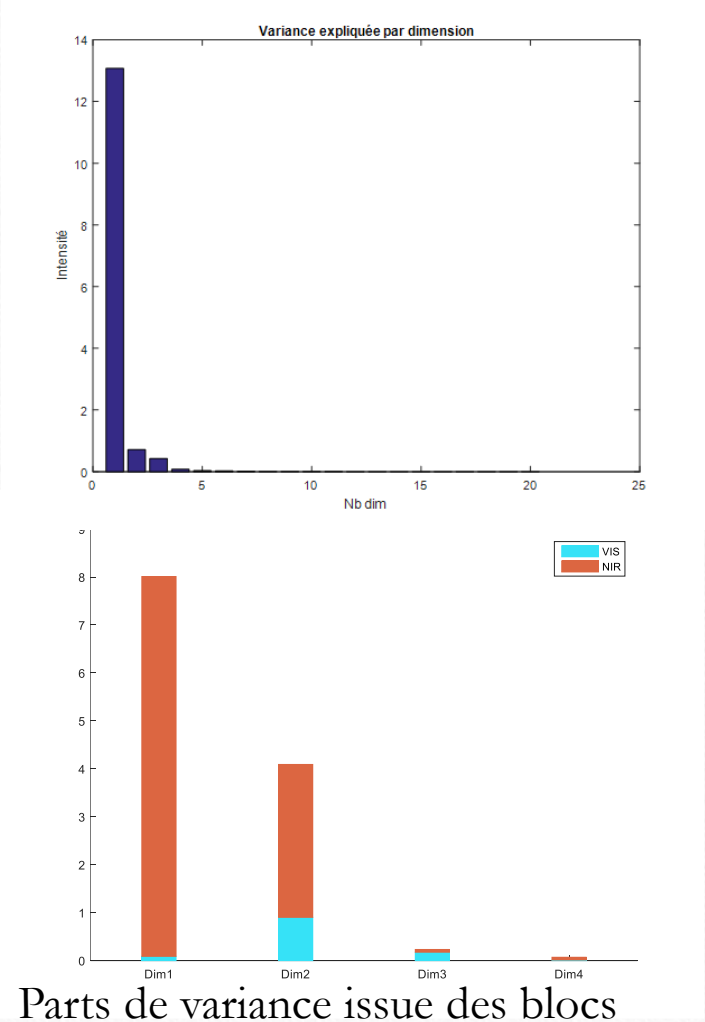
PARTIE 2.
Principes et méthodes

B. Réduction de la dimensionnalité de données multiblocs

- B3. Remise en question du référentiel commun à partir d'un exemple

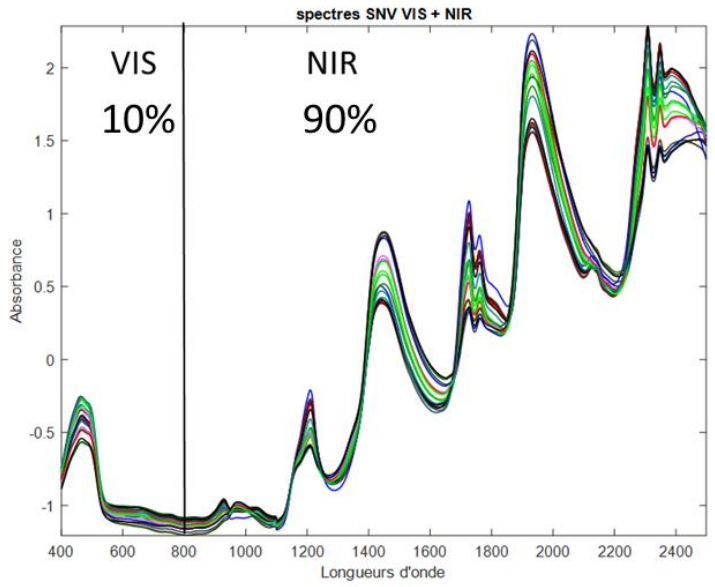
Remise en question du référentiel commun

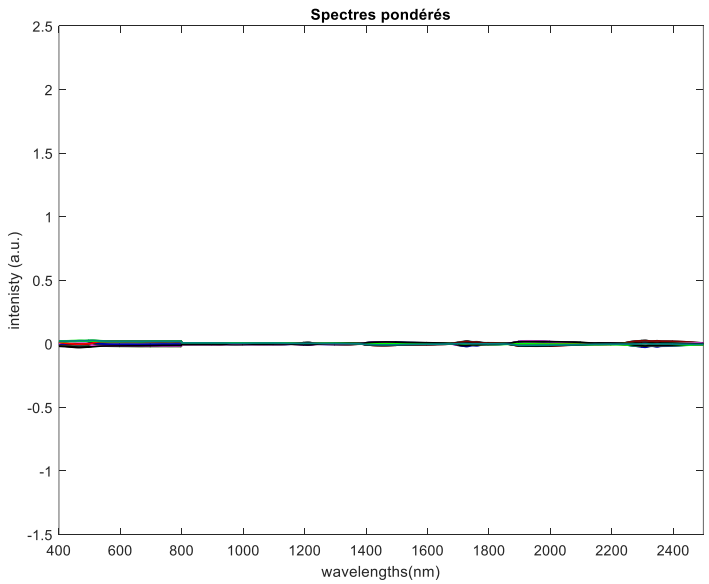
- Cause : Disparité de variabilité des deux plages spectrales
- la Co variation entre blocs (ou l'analyse) est-t-elle affectée par la différence d'échelle de variabilité des données ?



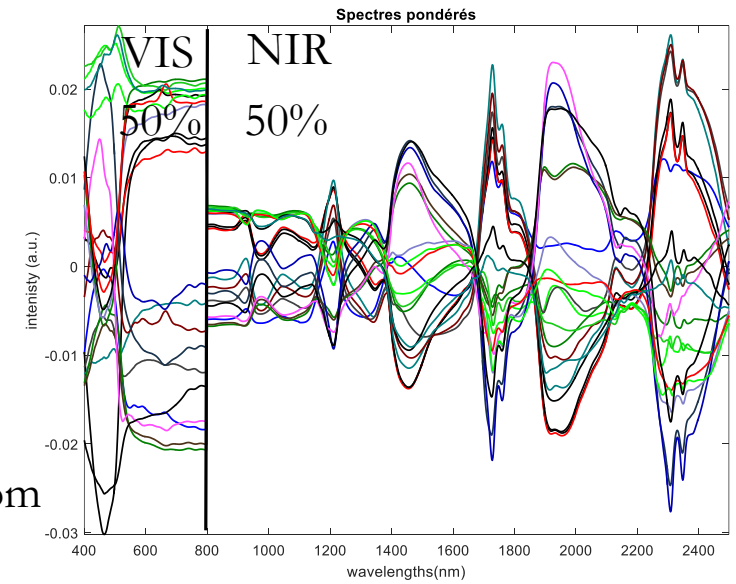
Parts de variance issue des blocs

Variance totale de chaque bloc = 1

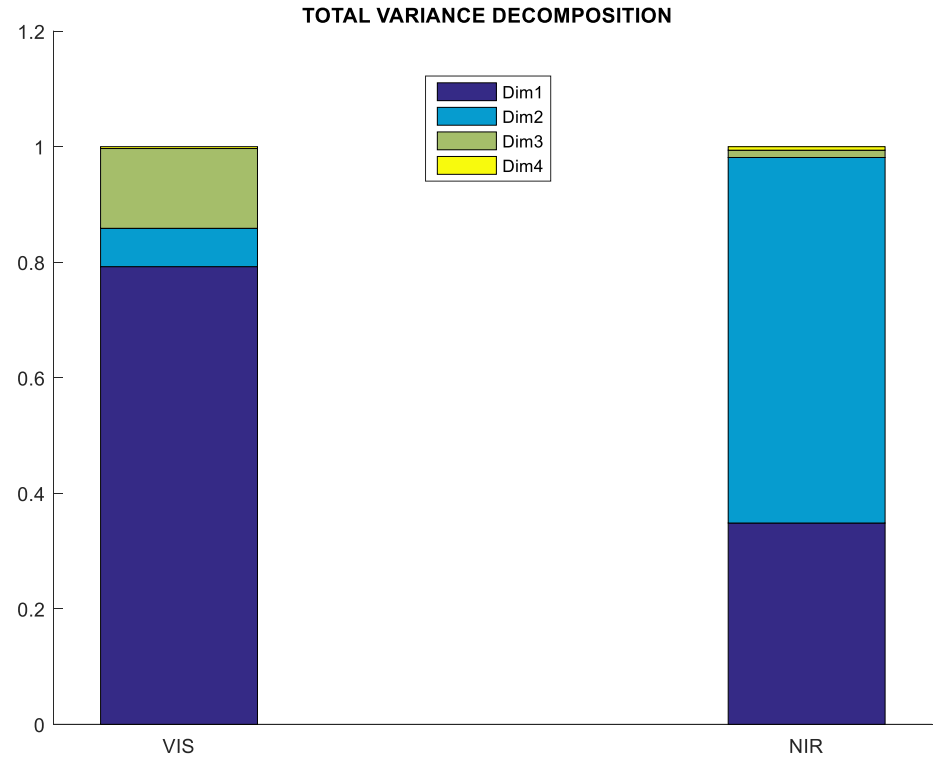


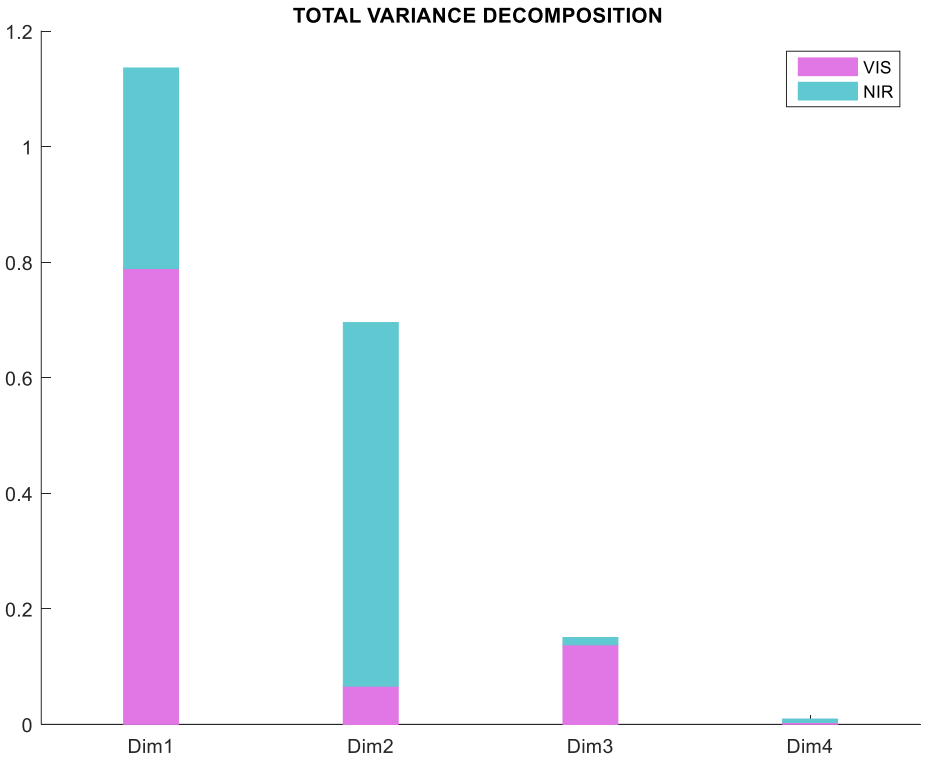


zoom



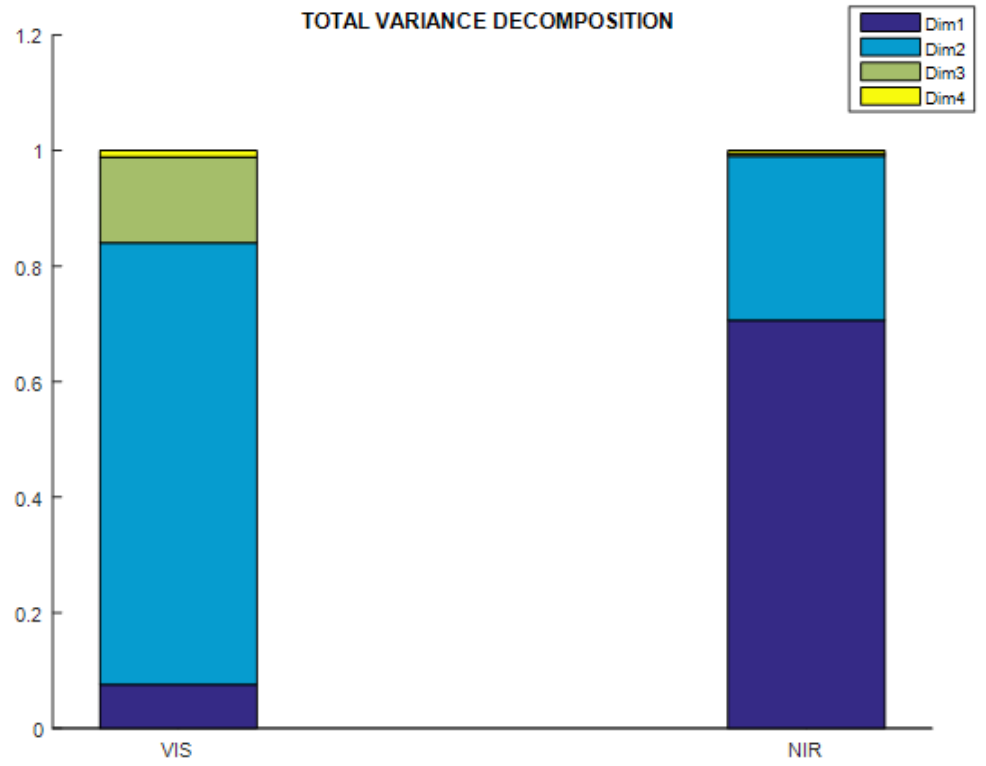
Variance expliquée



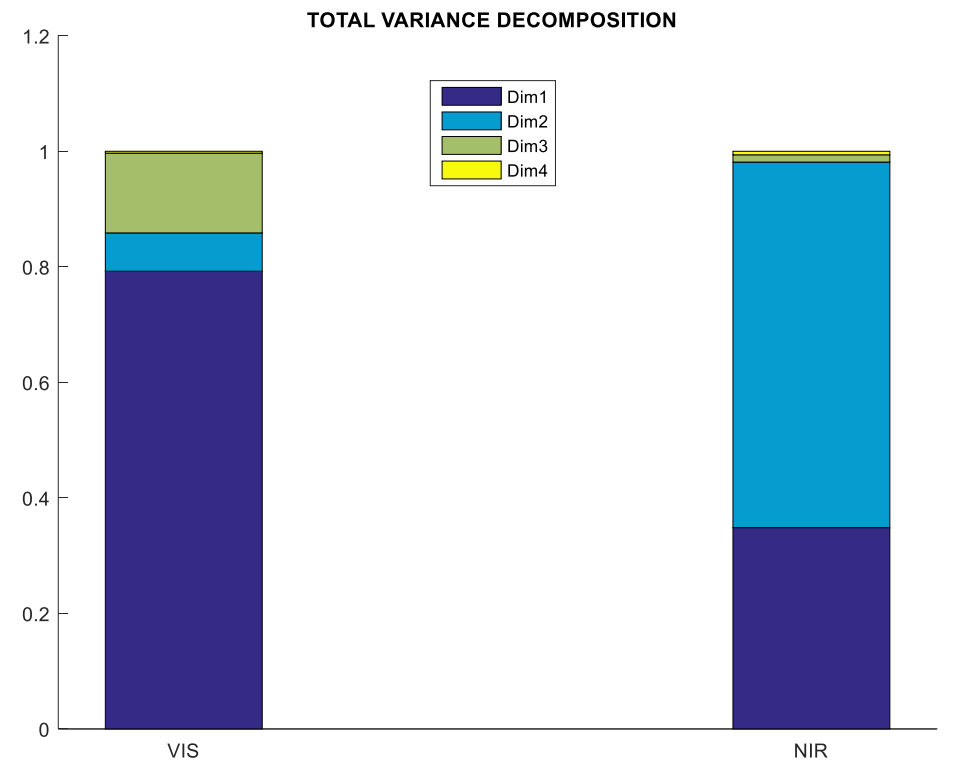


Variance totale de chaque bloc = 1

Variance expliquée

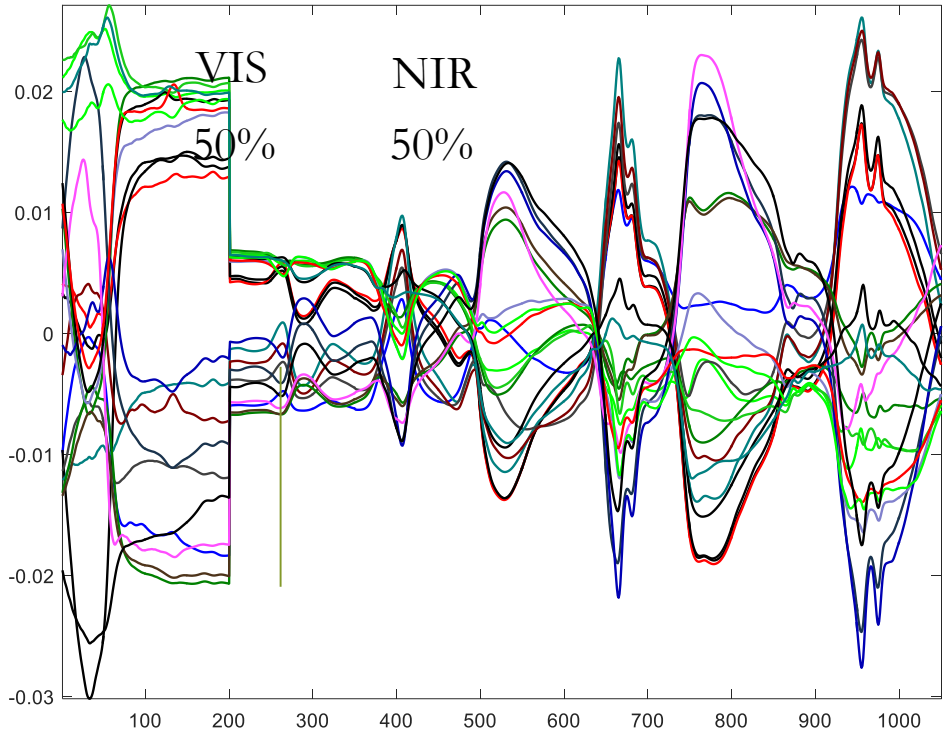


Sans standardisation des blocs

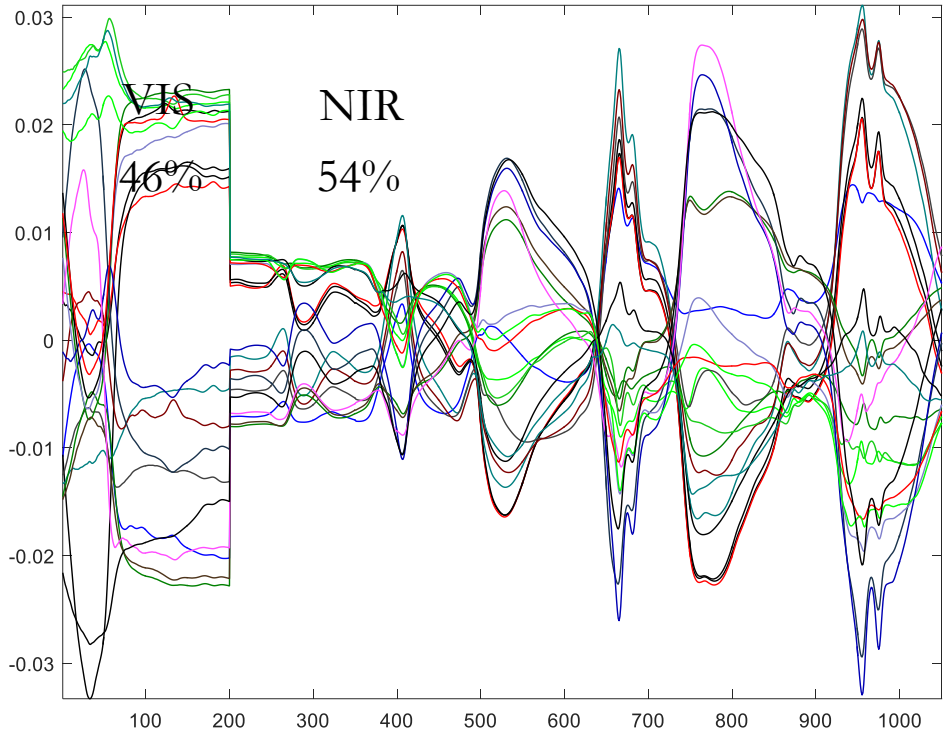


Variance totale de chaque bloc = 1

Alternative

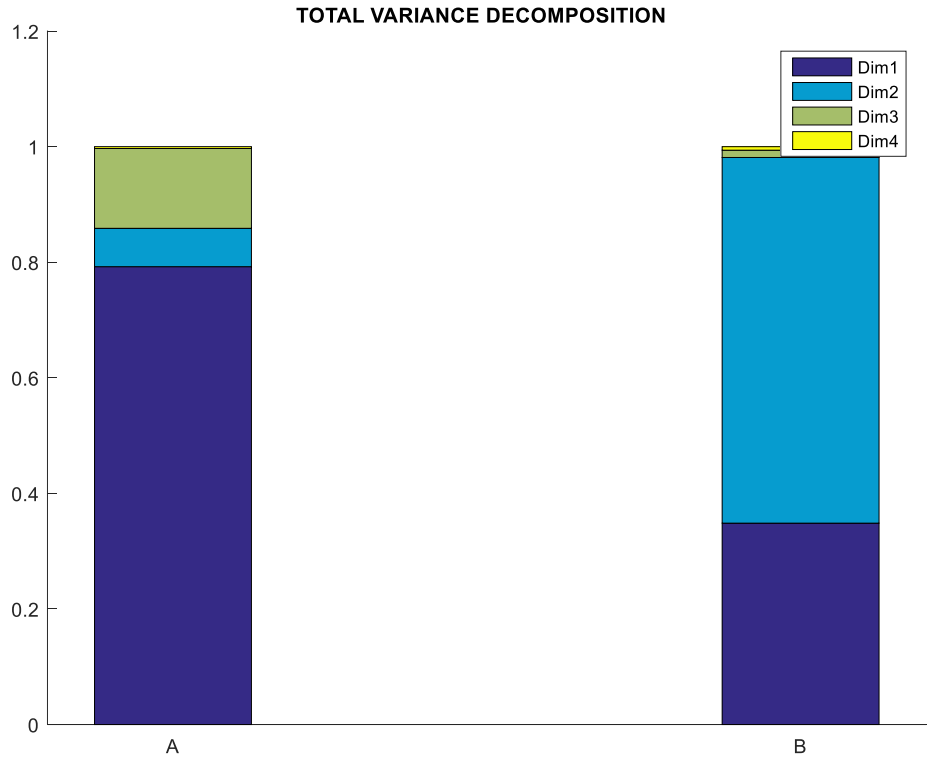


Variance totale de chaque bloc = 1

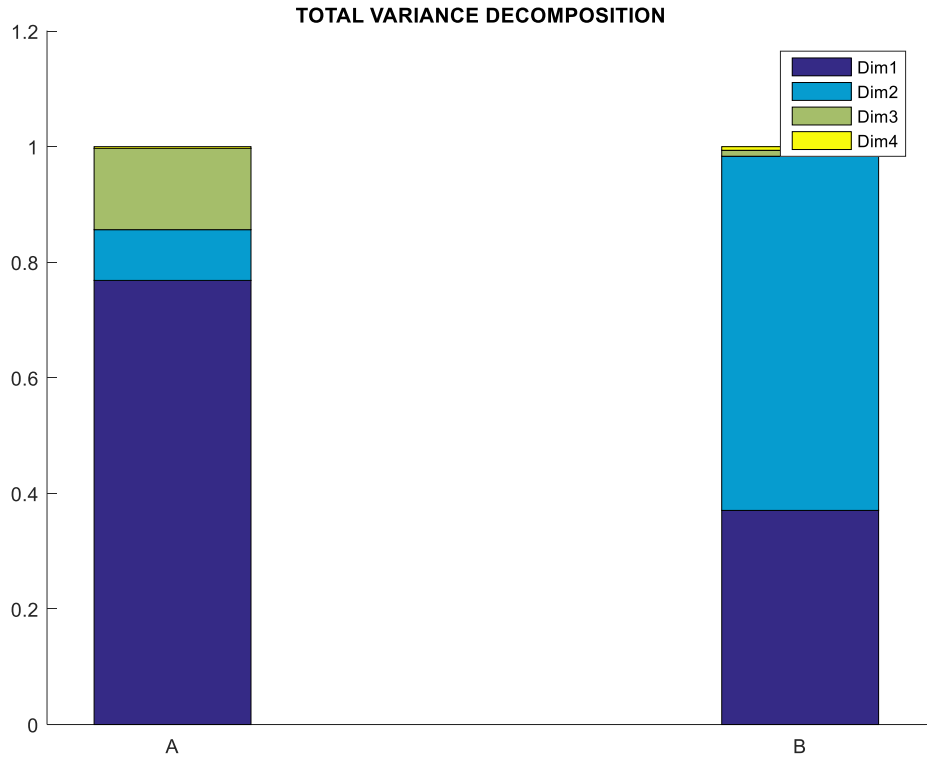


Variance CP1 de chaque bloc = 1

Permutation des dimensions



Variance totale de chaque bloc = 1



Variance CP1 de chaque bloc = 1

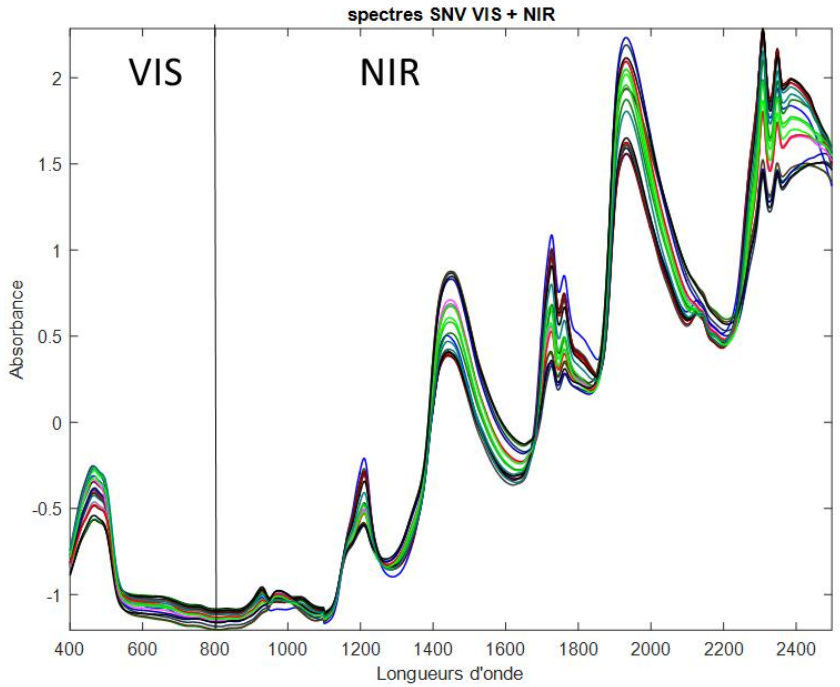
En présence de blocs tous spécifiques, le changement du référentiel par standardisation des blocs n'affecte pas l'analyse

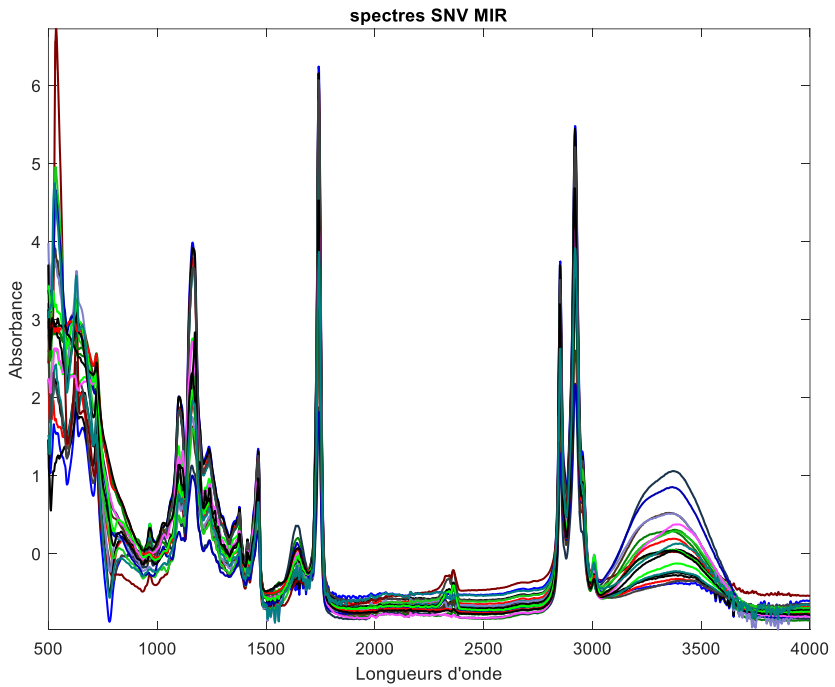
PARTIE 3.
Principes et méthodes

B. Réduction de la dimensionnalité de données multiblocs

- B3. Analyse des trois blocs (VIS, NIR et MIR)

Avant propos



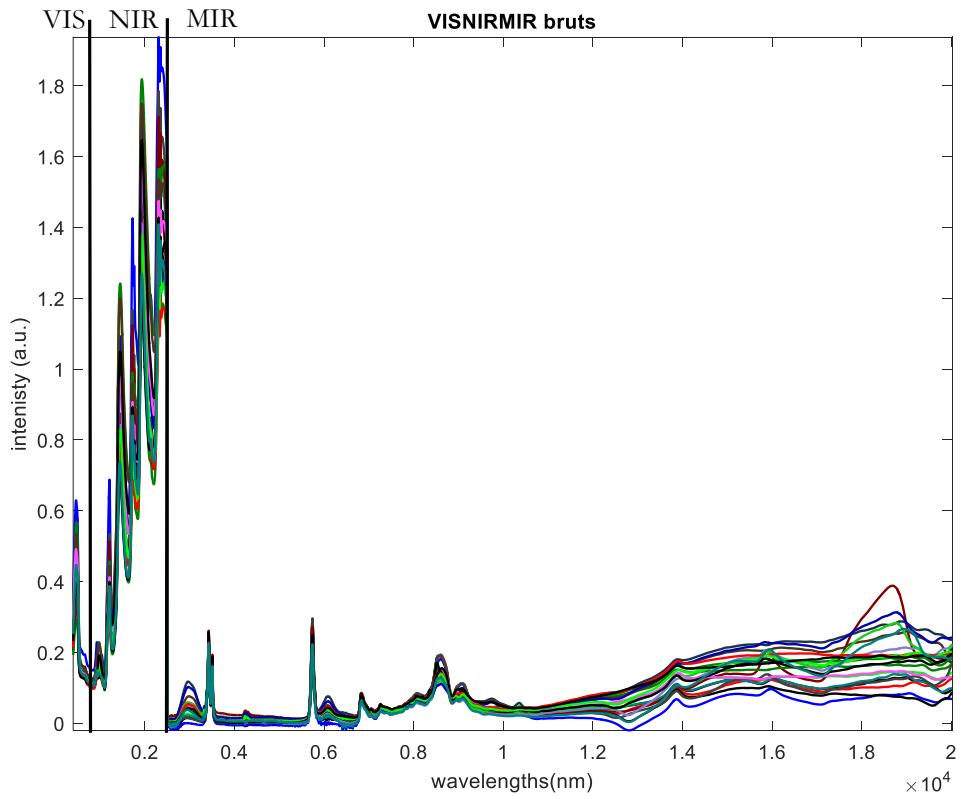


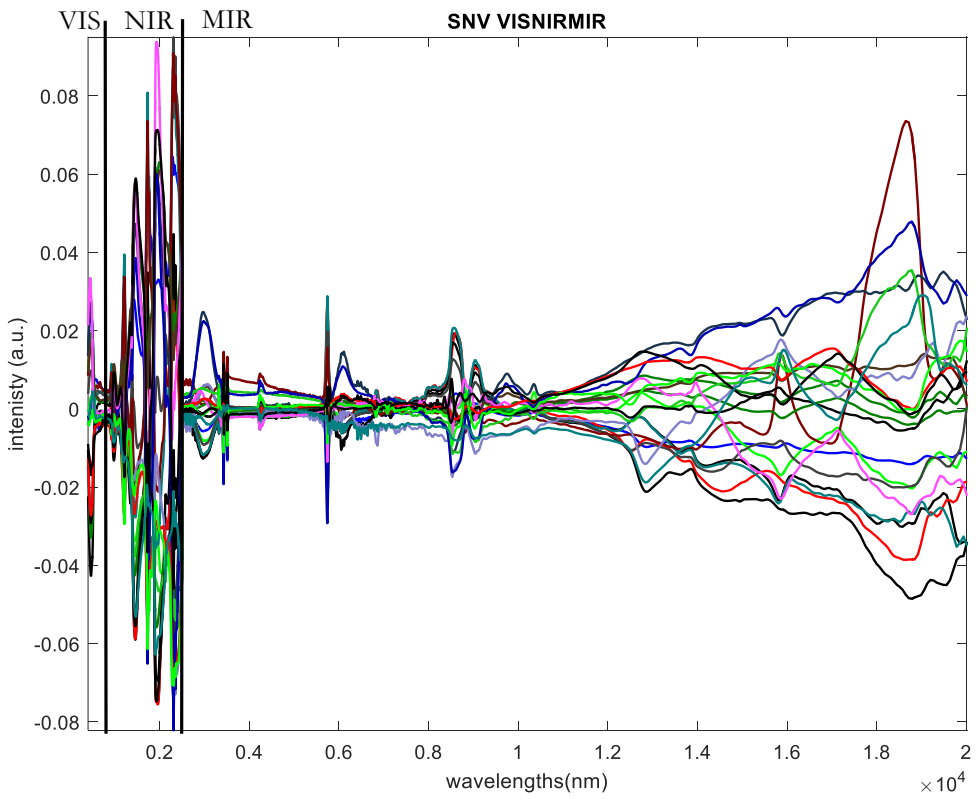
VIS et NIR acquis sur le même appareil
Avec unité en *longueur d'onde*

MIR acquis sur un autre appareil
Avec *nombre d'onde*

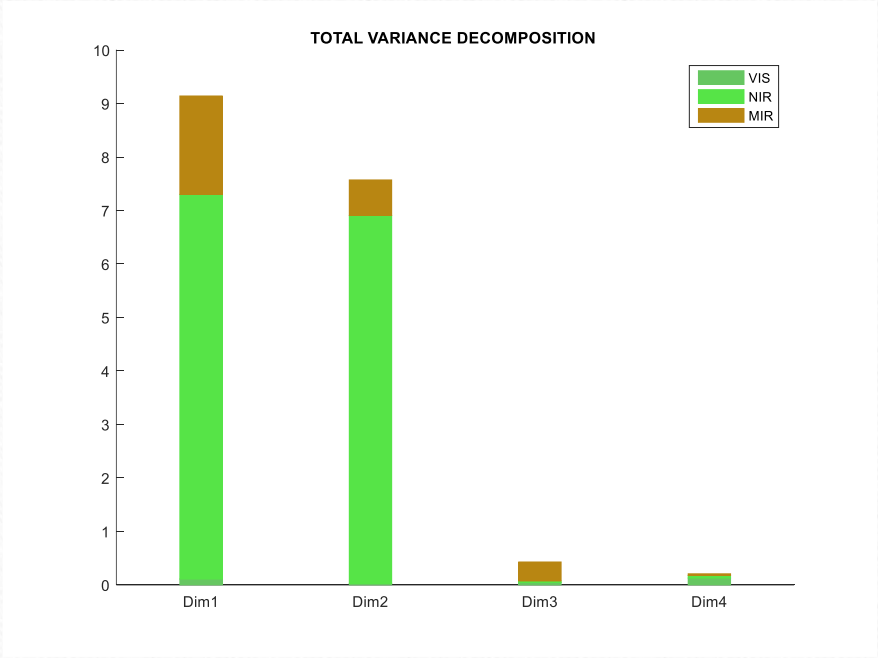
1 cm=1/nm*10^7

SNV (3blocs concaténés)



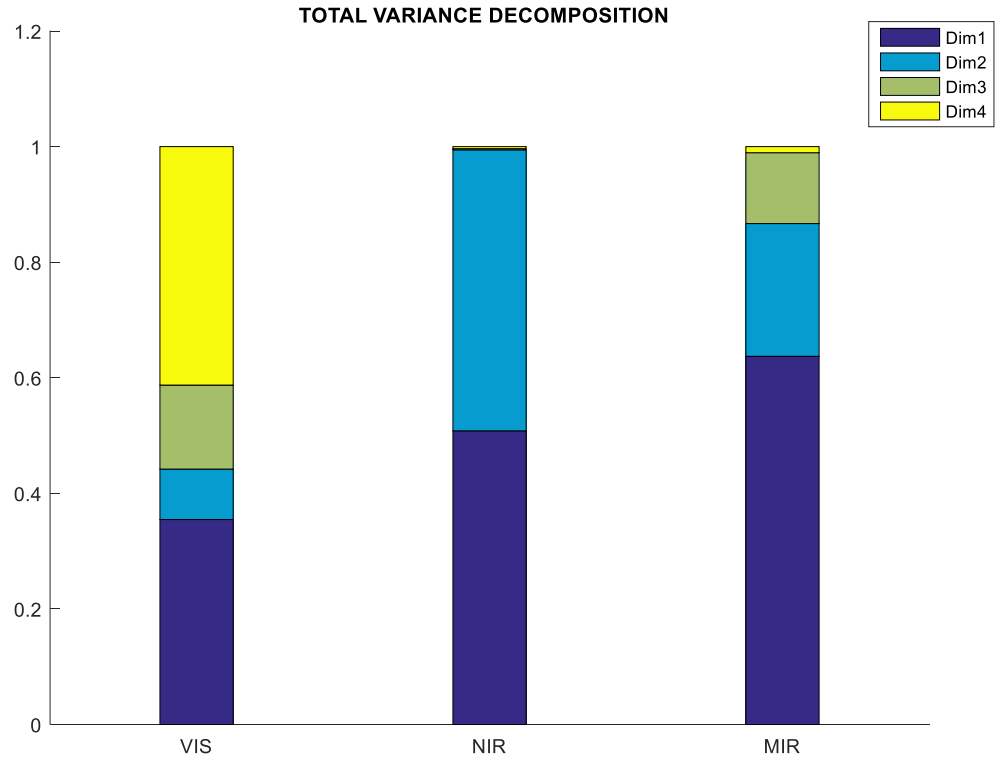


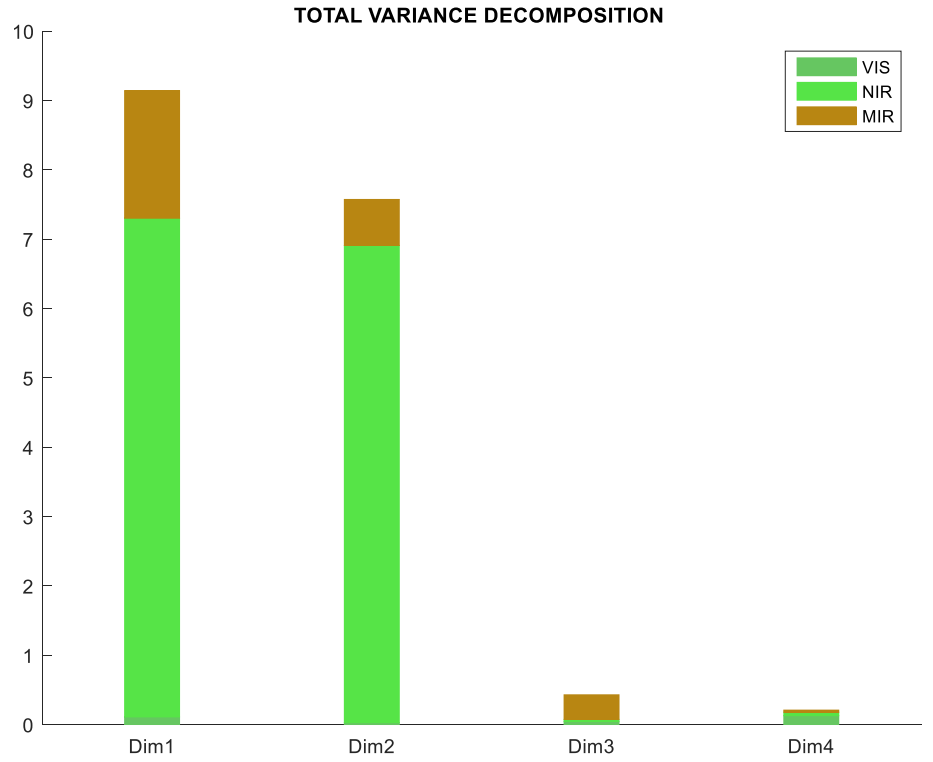
Analyse MB 3 blocs sans pondération



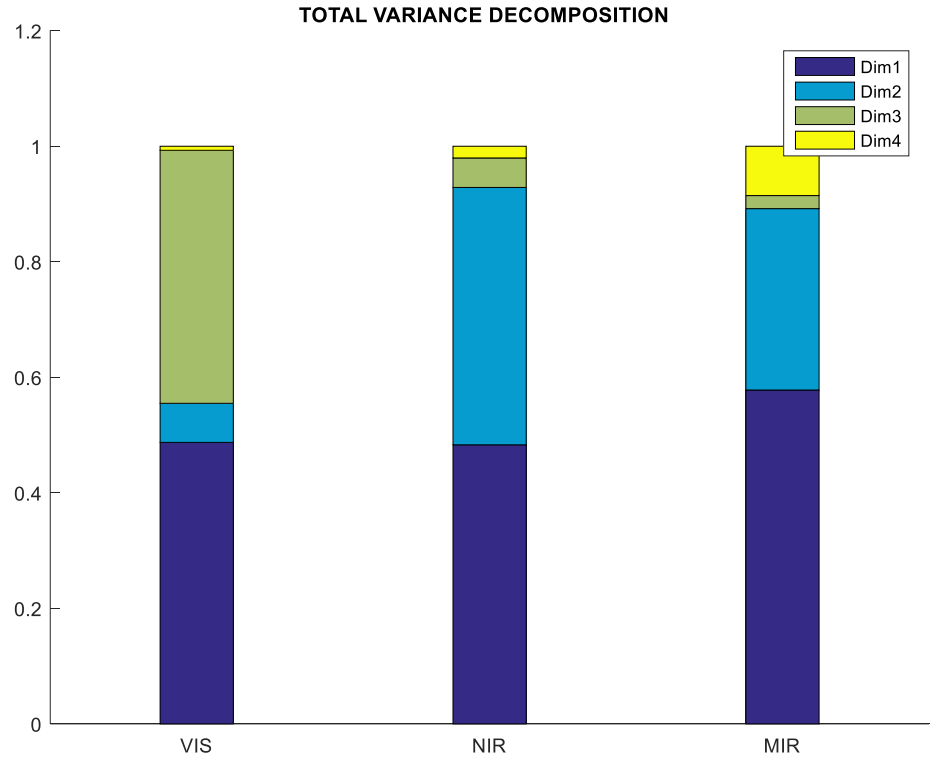
	variance	%
Visible	0.3412	2%
NIR	14.2518	81%
MIR	3.0154	17%

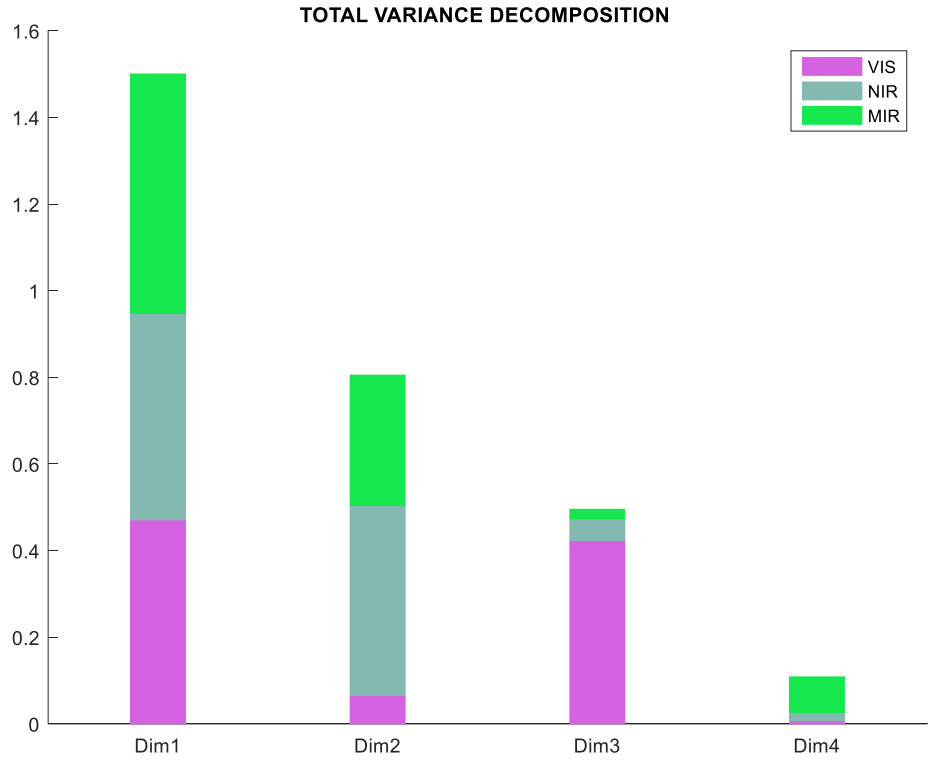
Intéressant ?



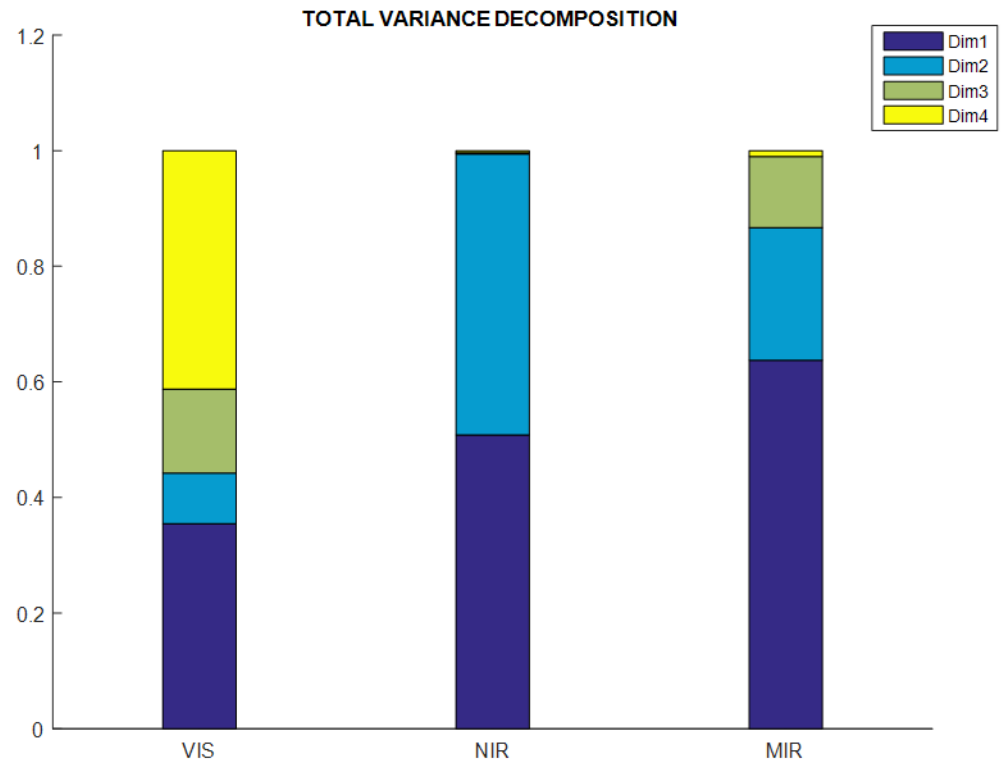


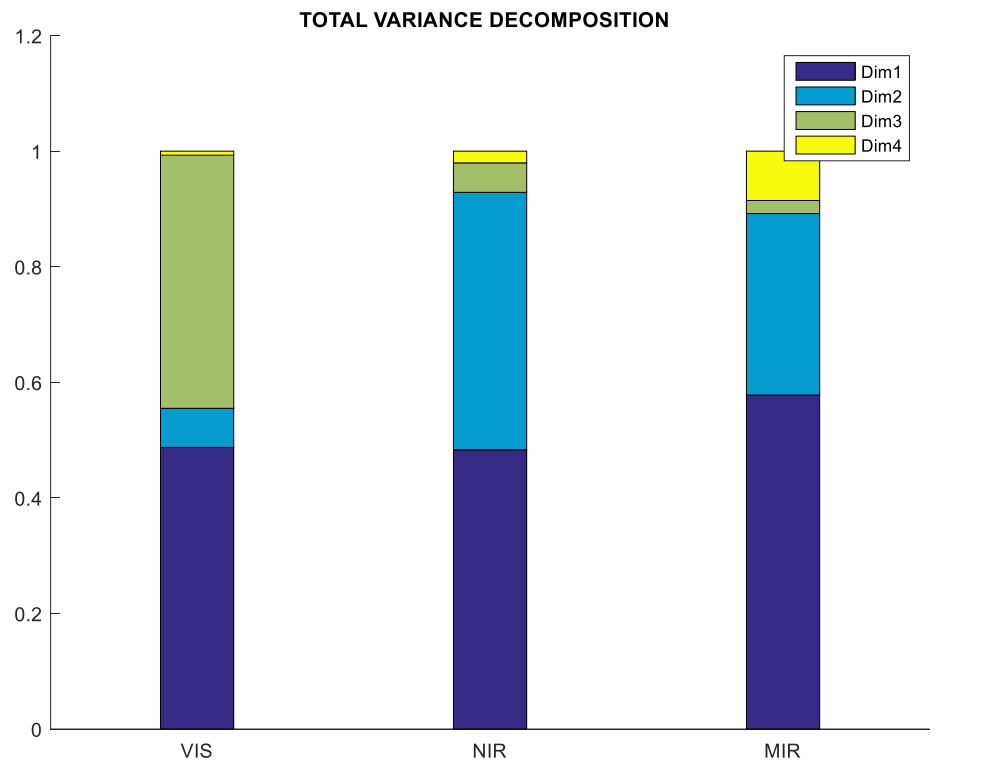
Variance totale de chaque bloc= 1





Quel choix ?

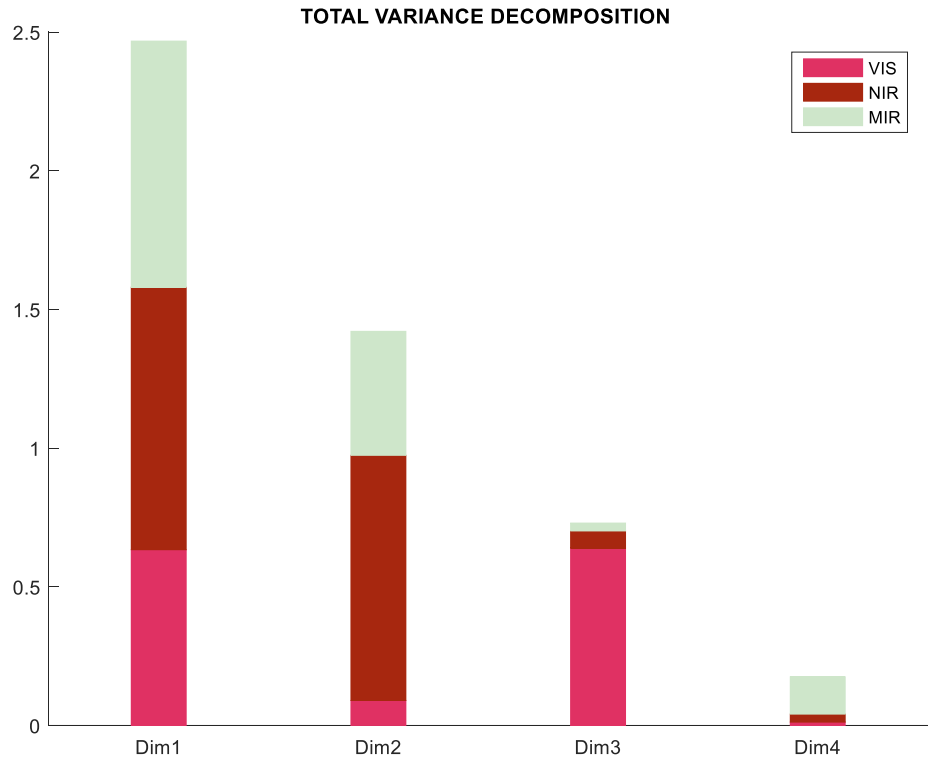




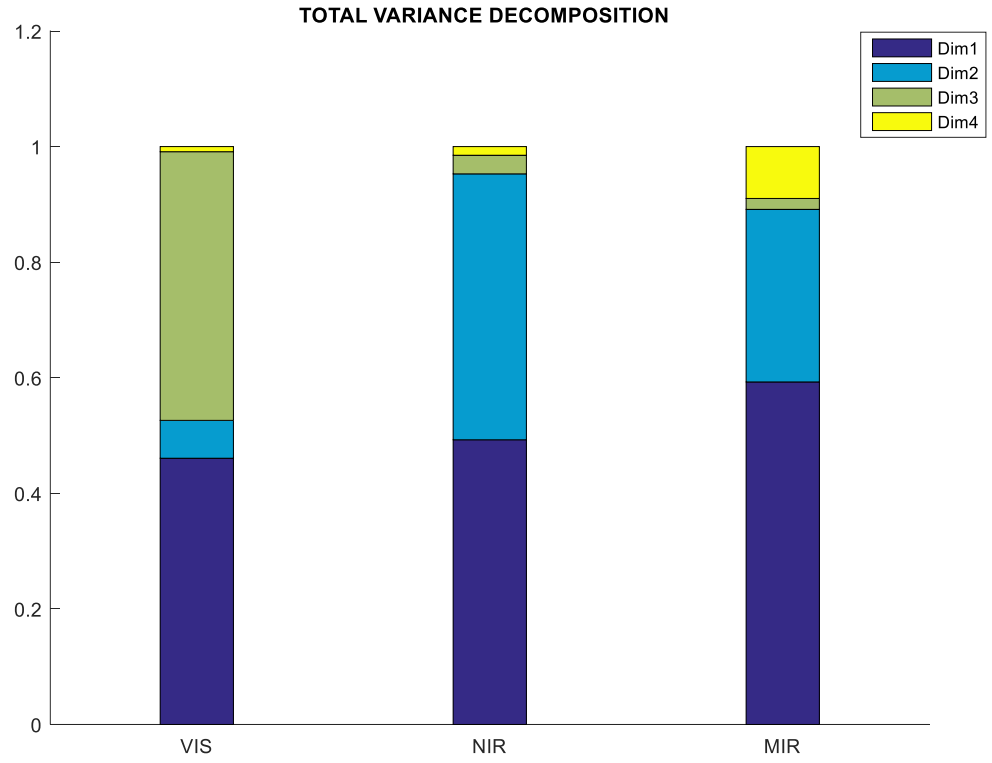
variance totale de chaque bloc= 1

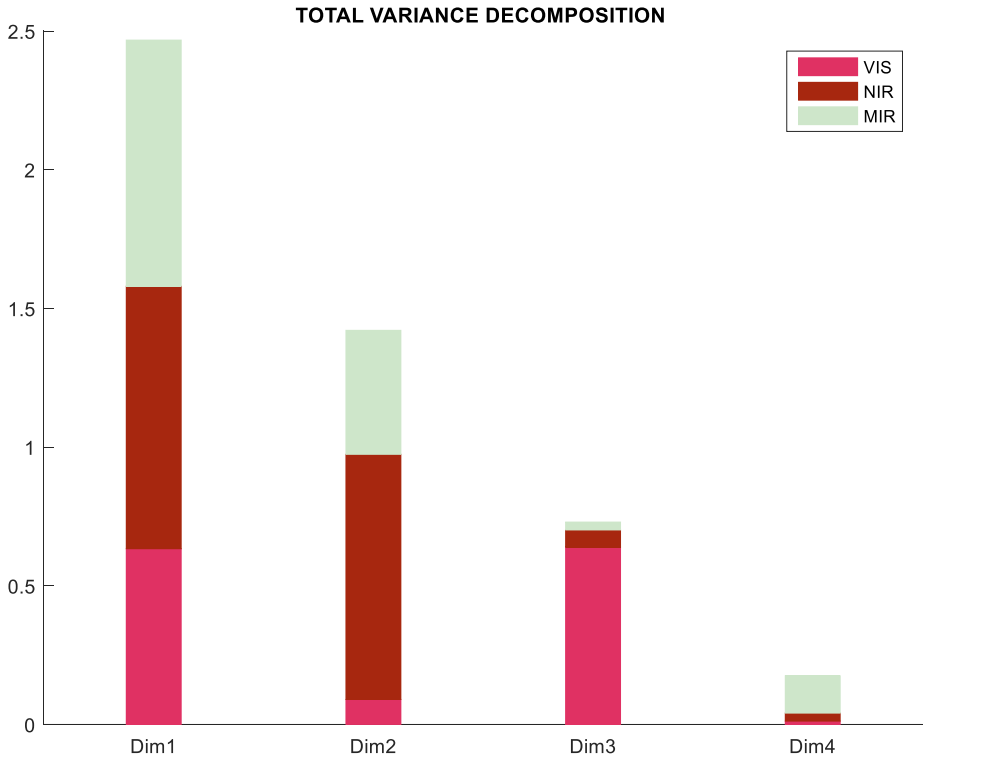
Variance CP1 de chaque bloc = 1

	variance	%
Visible	1.4259	29%
NIR	1.9439	39%
MIR	1.5660	32%

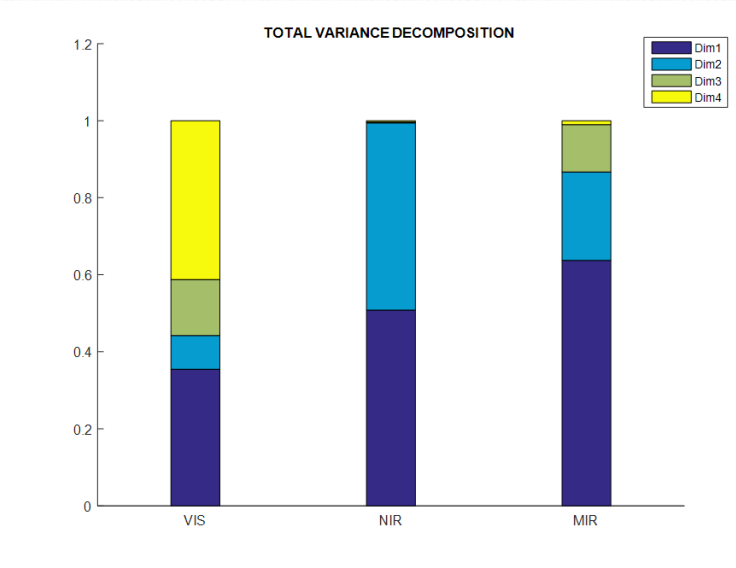


Variance CP1 de chaque bloc = 1

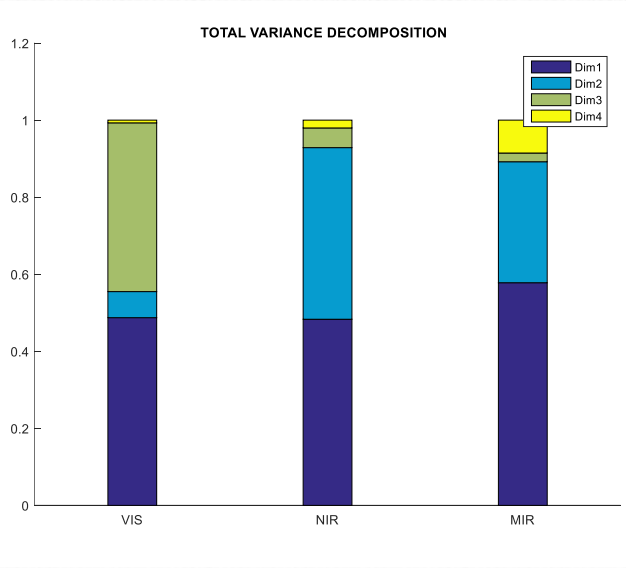




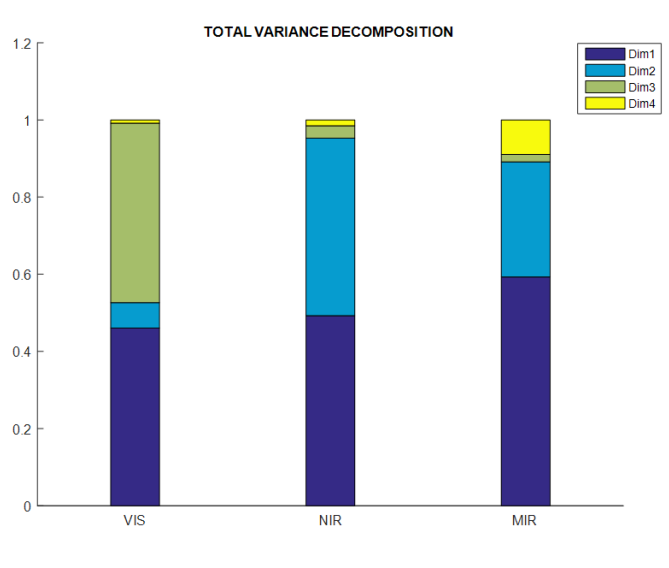
Parcimonie



sans



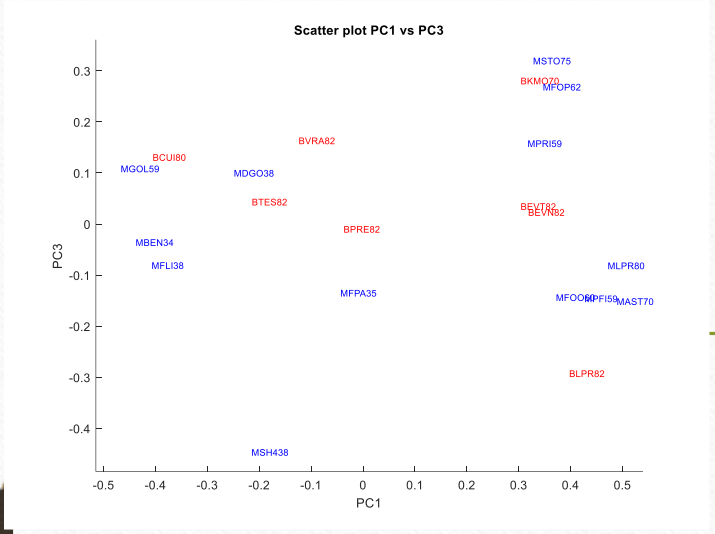
variance totale de chaque bloc = 1

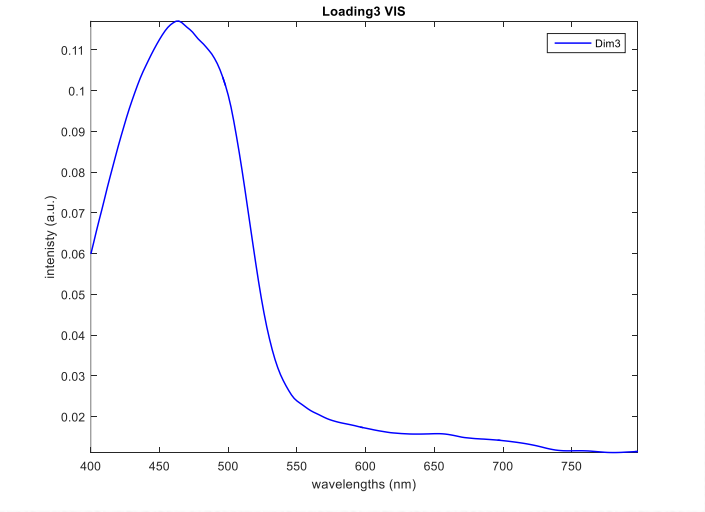


Variance CP1de chaque bloc= 1

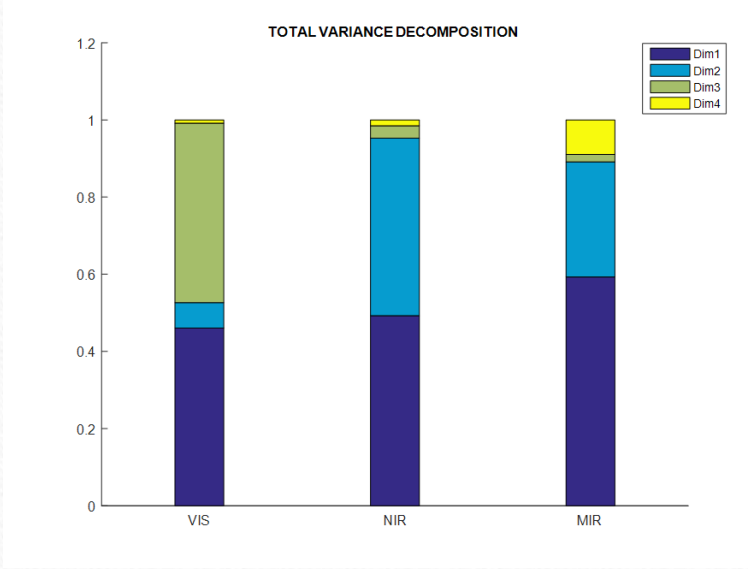
Analyse MB des 3 blocs avec pondération CP1 = 1

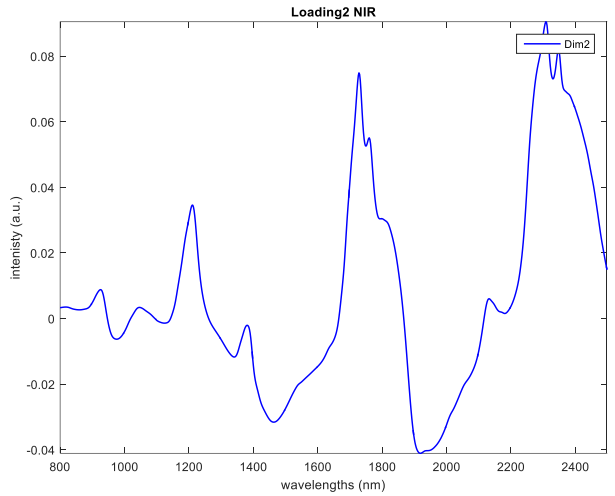
Interprétation dimension 3

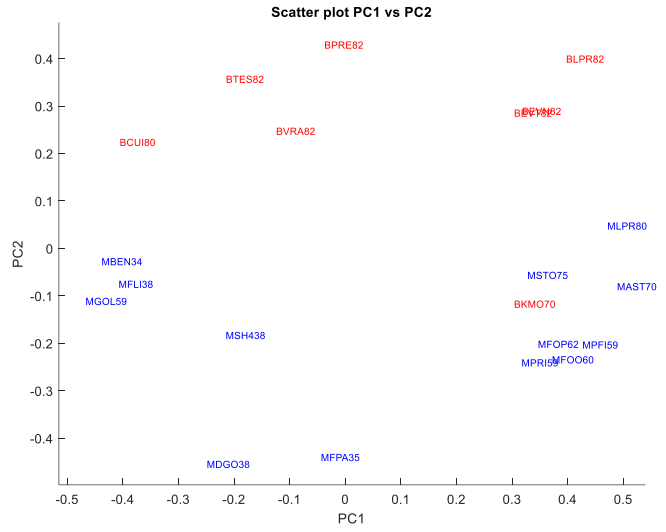


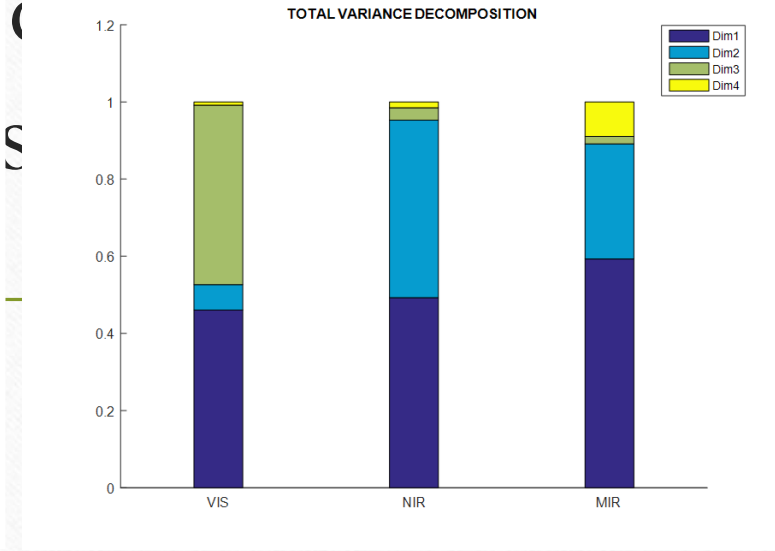


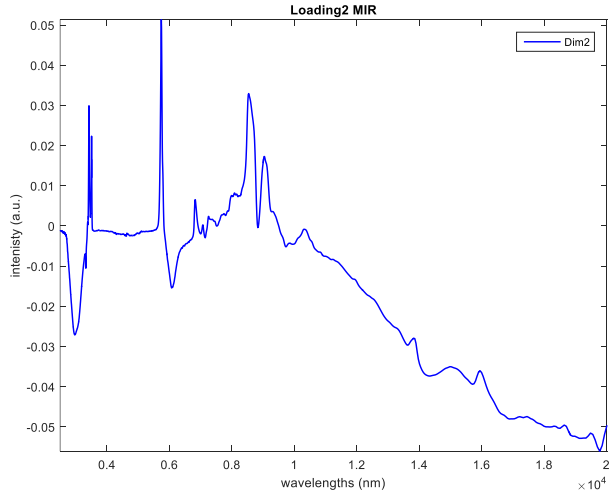
Le bloc visible contribue essentiellement à la construction de la dimension 3. Il correspond au spectre du β-carotène, qui est naturellement présent dans le beurre sous la forme « cis ».
(Dimension spécifique au bloc VIS)



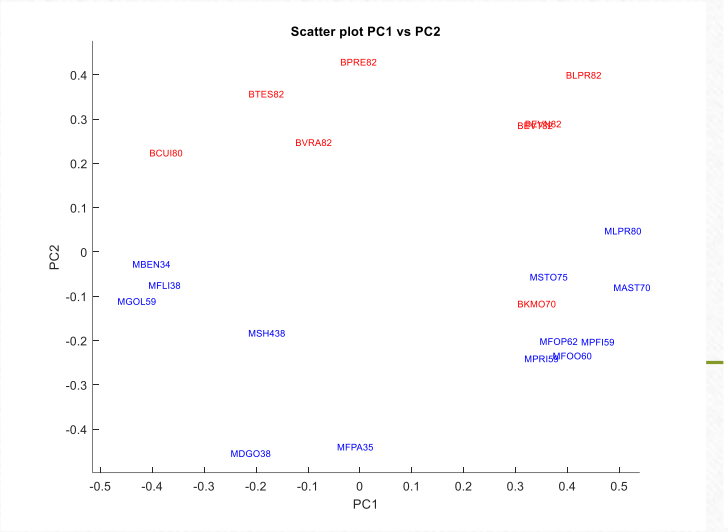


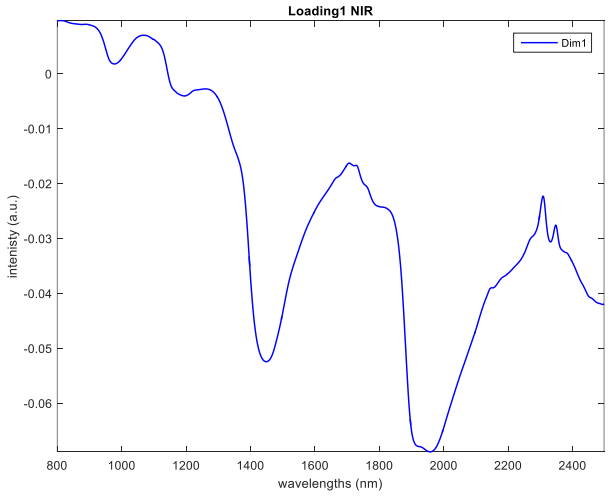




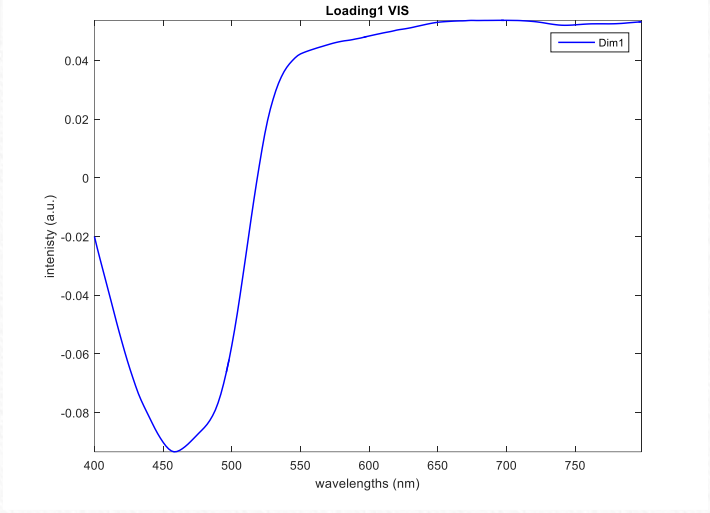


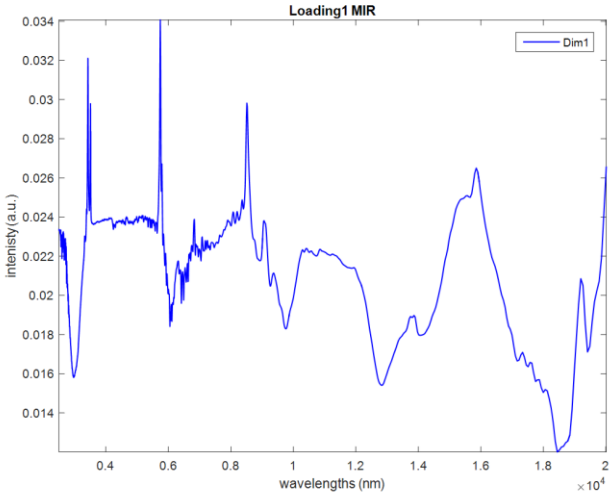
La dimension 2 est synthétisée à partir des blocs NIR et MIR. Le loading du NIR correspond à la teneur en MG qui croît avec l'axe, avec des valeurs négatives caractéristiques de l'eau.
Le loading du MIR a la même interprétation.
Les deux loadings vont dans le même sens, en accord avec les scores.
(dimesion commune entre NIR et MIR)

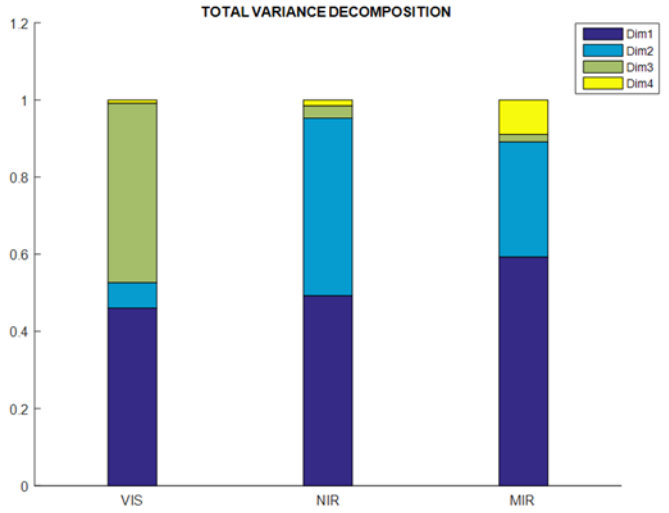




La dimension 1.
difficile à interpréter ?
.



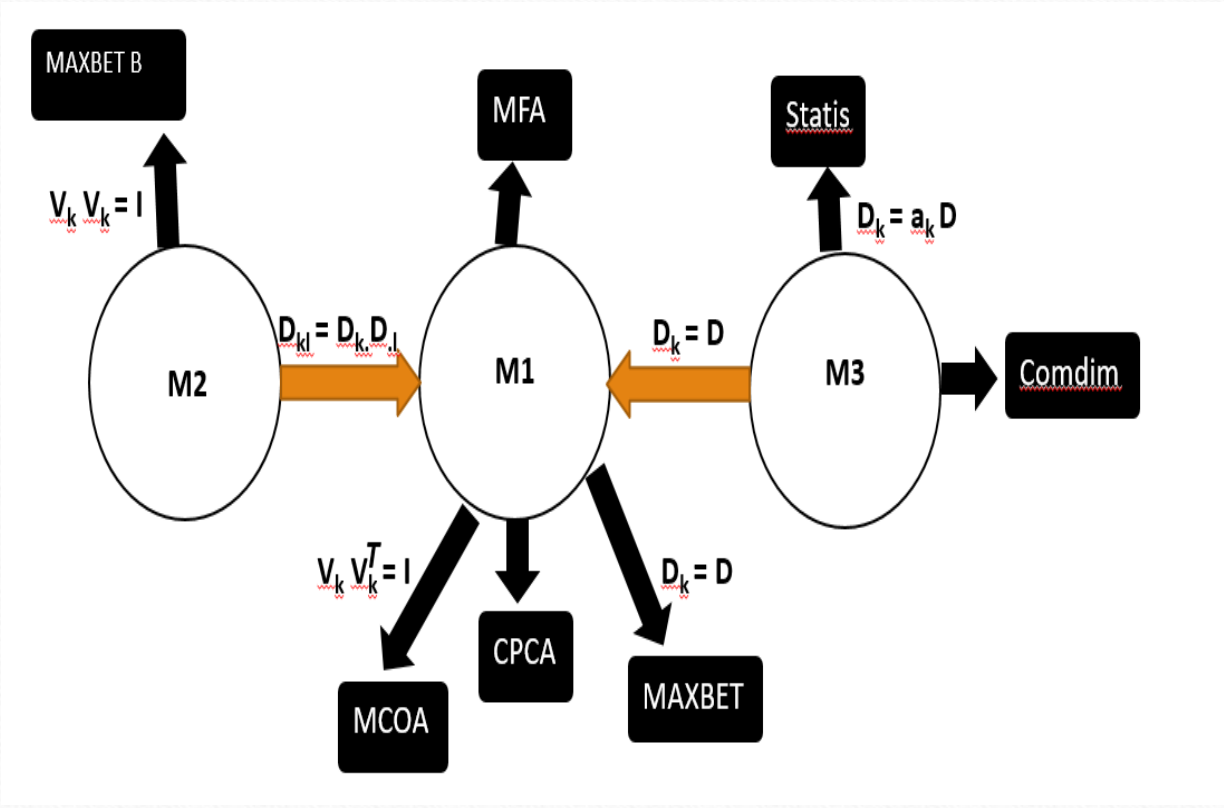




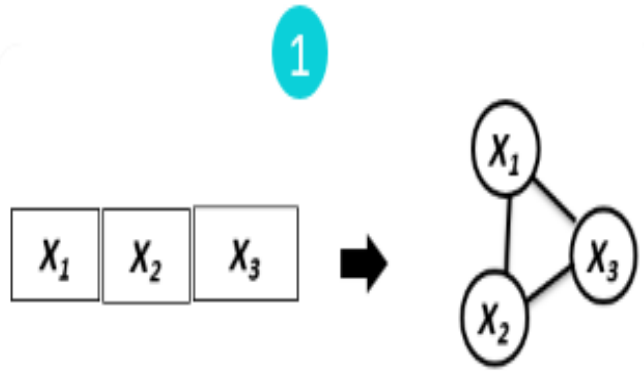
PARTIE 4.
Aller plus loin

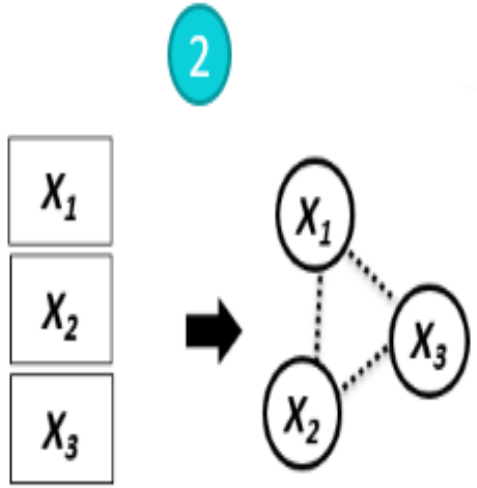
- Se former
  - Chemomics (session 4)
  - Chemoocs
- Approfondir
  - Un consortium dédié MIMS
  - Session spéciale Chimiométrie 2024
- Développer
  - Approche en réseaux de tableaux (NetPCA)

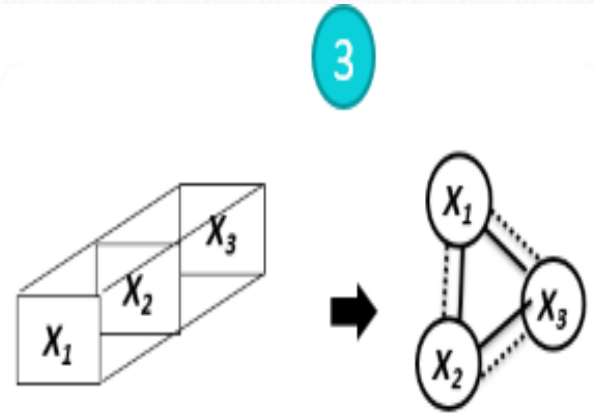
Consortium MIMS

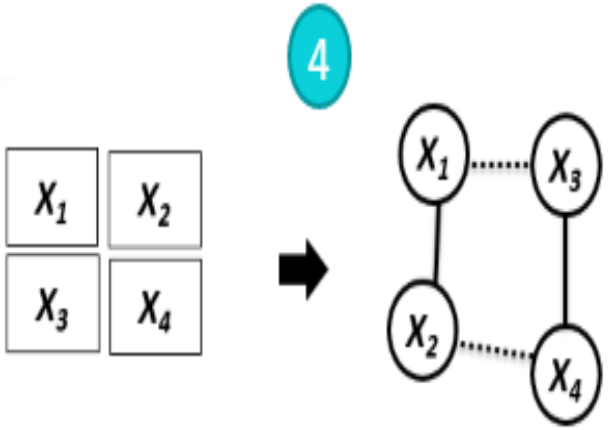


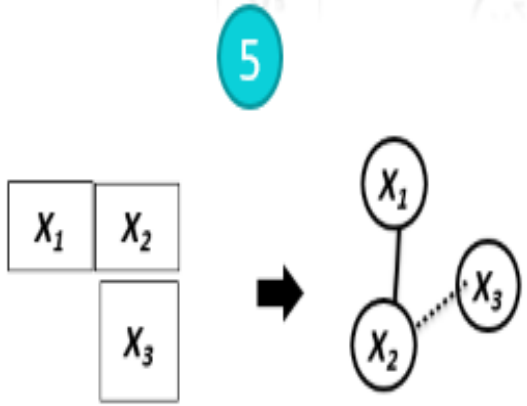
Métaprogramme DIGIT-BIO
« Biologie Numérique pour explorer et prédire le vivant »

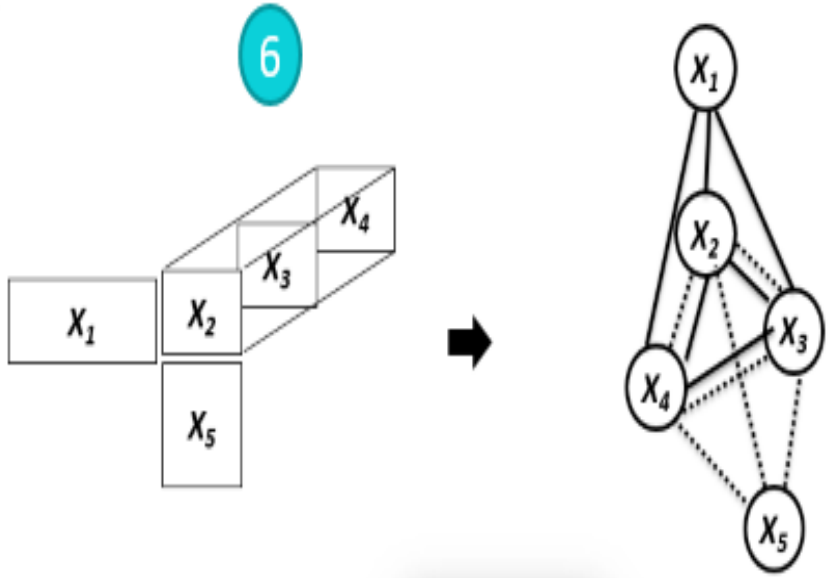

1
X1
X2
X3
X1
X2
X3


2
X1
X2
X3
X1
X2
X3


3
X1
X2
X3
X1
X2
X3

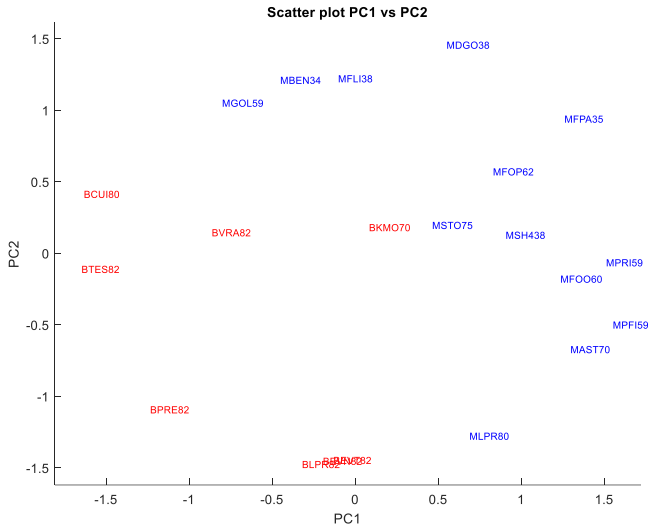

4
X1
X2
X3
X4
X1
X3
X2
X4

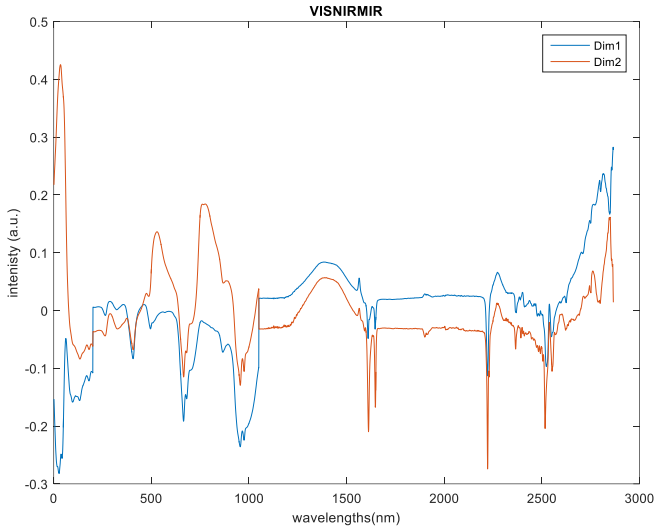

5
X1
X2
X3
X1
X3
X2

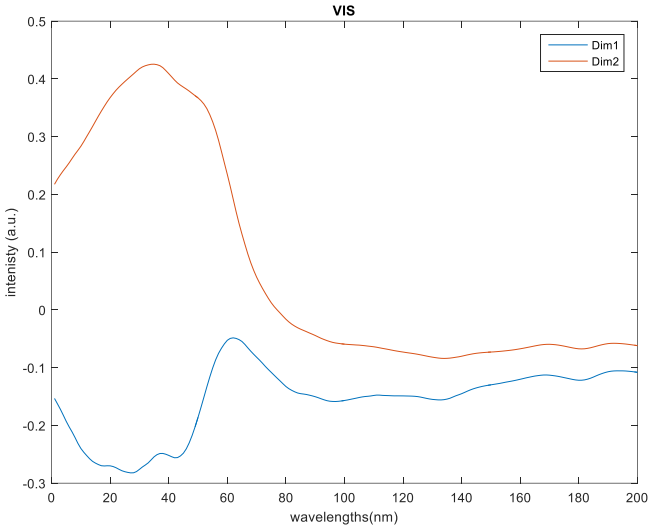

6
X1
X2
X3
X4
X5
X1
X2
X3
X4
X5

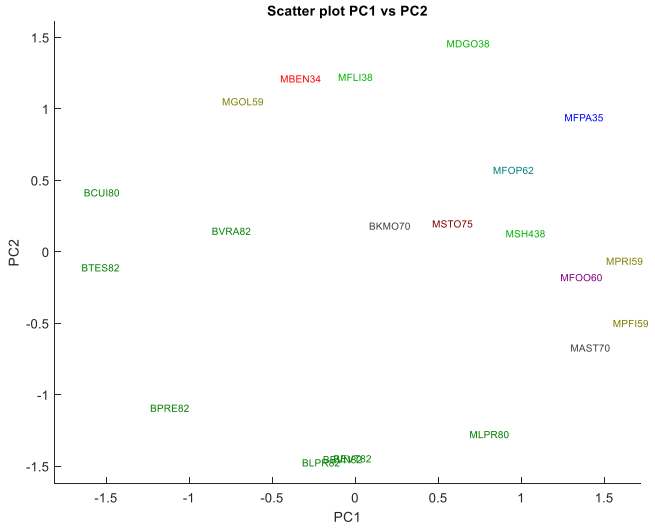

DISCUSSION
TiME

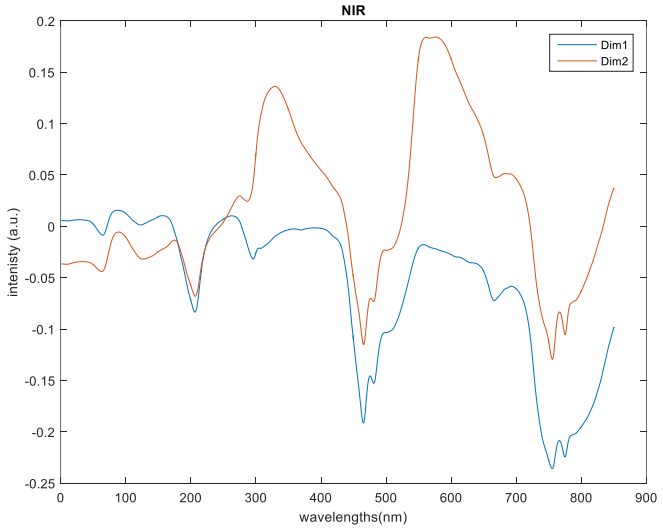
Prétraitements **snv par bloc**

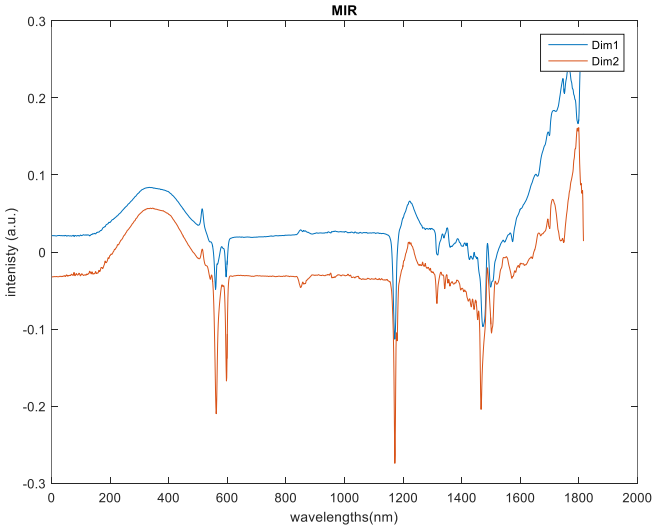

Scatter plot PC1 vs PC2
MDGO38
MBEN34
MFLI38
MGOL59
MFPA35
MFOP62
BCUI80
BVRA82
BKMO70
MSTO75
MSH438
MPRI59
BTES82
MFOO60
MPFI59
MAST70
BPRE82
MLPR80
PC1
PC2


VISNIRMIR
Dim1
Dim2
intenisty (a.u.)
wavelengths(nm)


VIS
Dim1
Dim2
intenisty (a.u.)
wavelengths(nm)

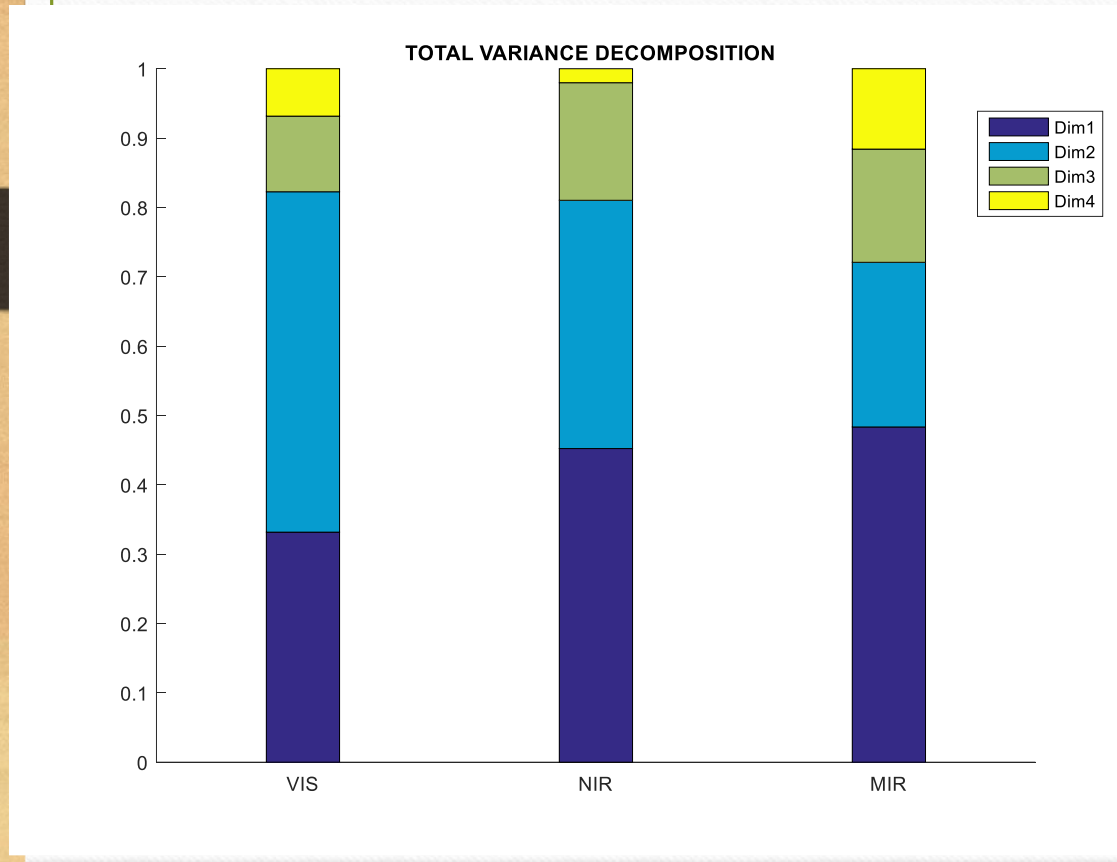

Scatter plot PC1 vs PC2
MDGO38
MBEN34
MFLI38
MGOL59
MFPA35
MFOP62
BCUI80
BVRA82
BKMO70
MSTO75
MSH438
MPRI59
BTES82
MFOO60
MPFI59
MAST70
BPRE82
MLPR80
PC1
PC2

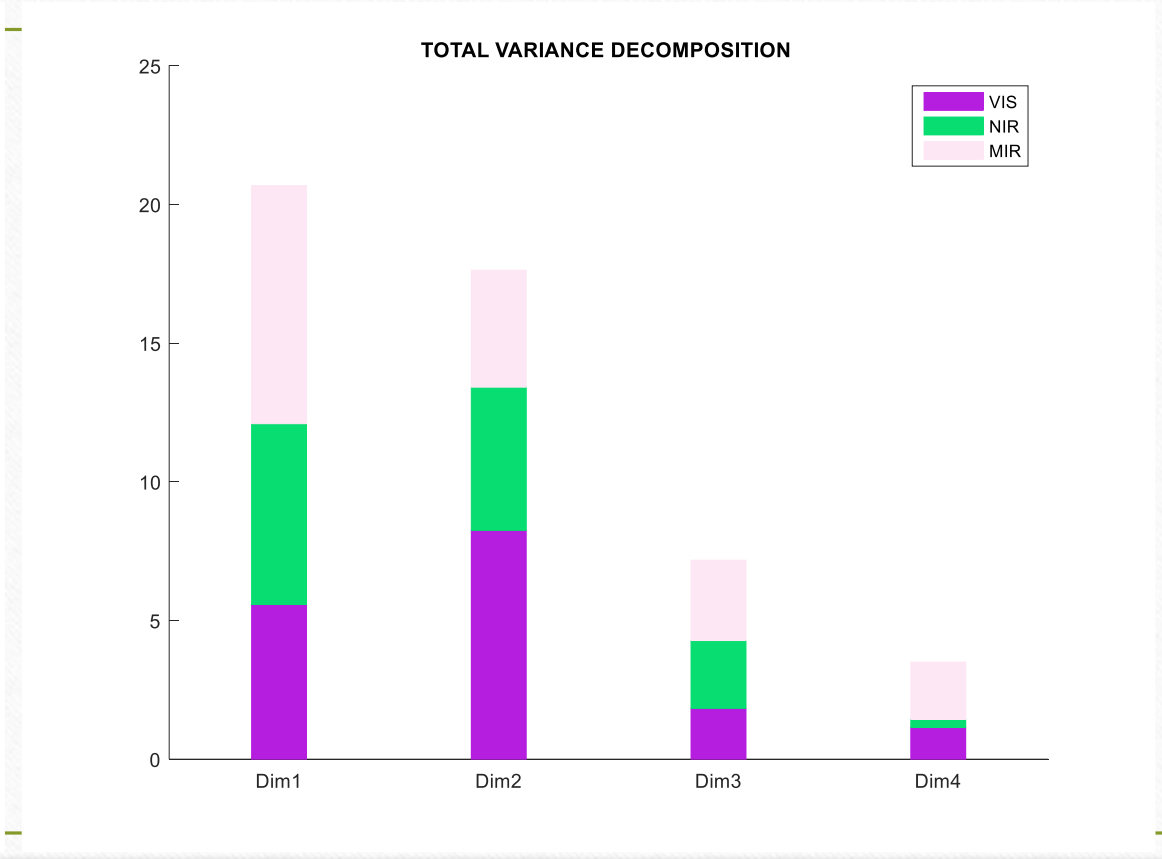

NIR
Dim1
Dim2
intenisty (a.u.)
wavelengths(nm)

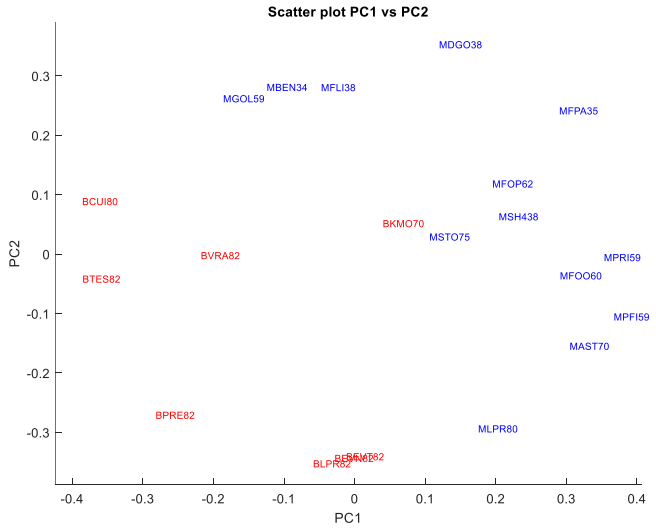

MIR
Dim1
Dim2
intenisty (a.u.)
wavelengths(nm)

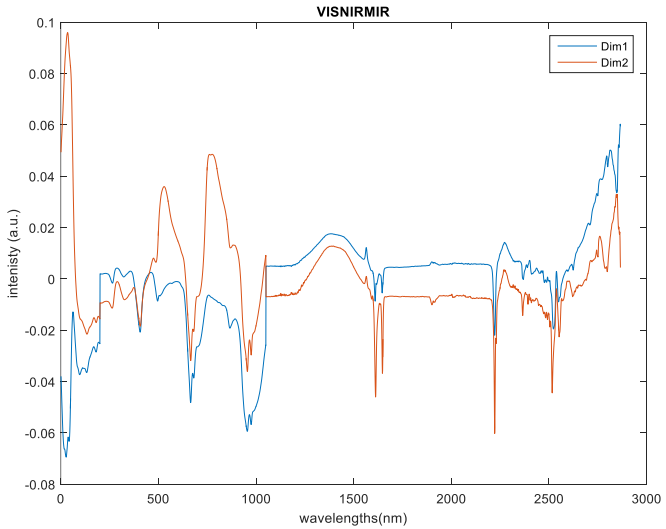
Variance expliqu

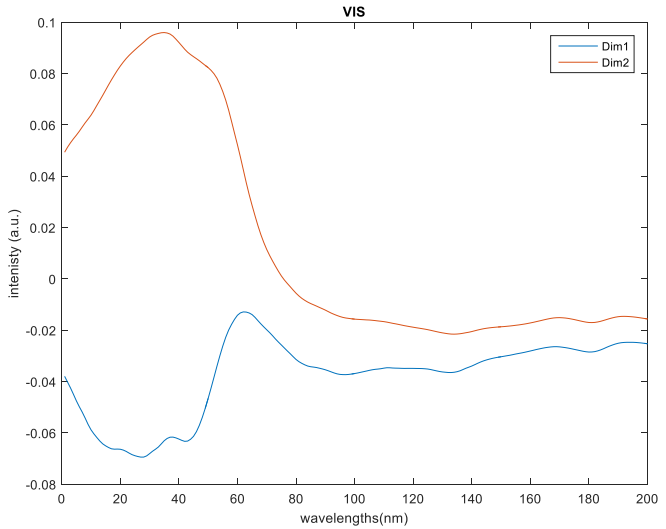
	Inertie Totale	Inertie Totale %
Visible	18.0463	33.0368
NIR	15.2891	27.9892
MIR	21.2896	38.9741

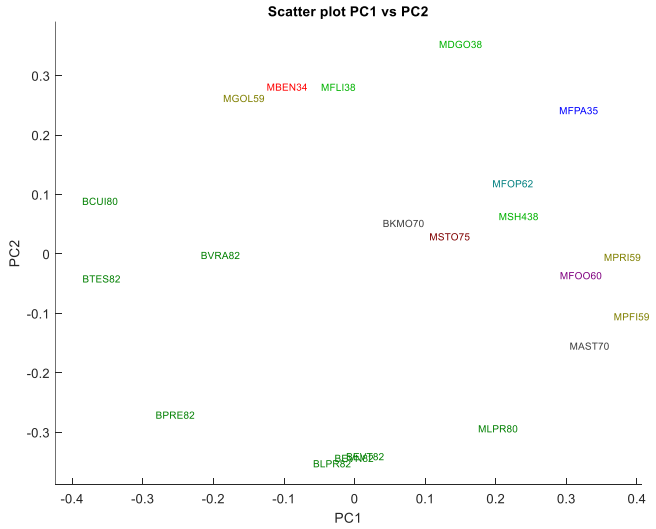


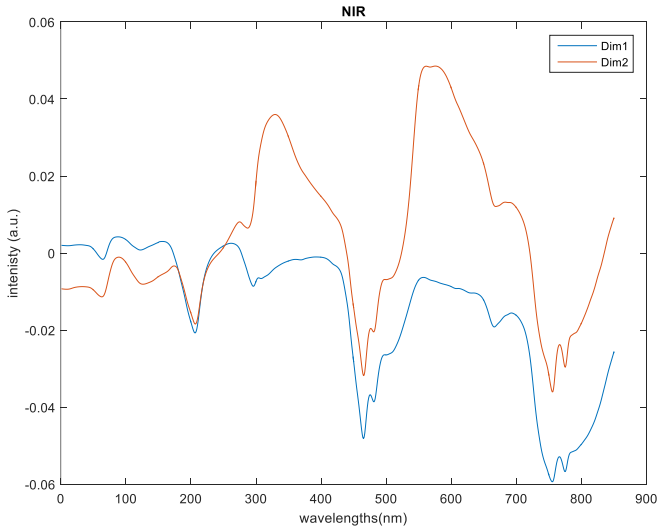


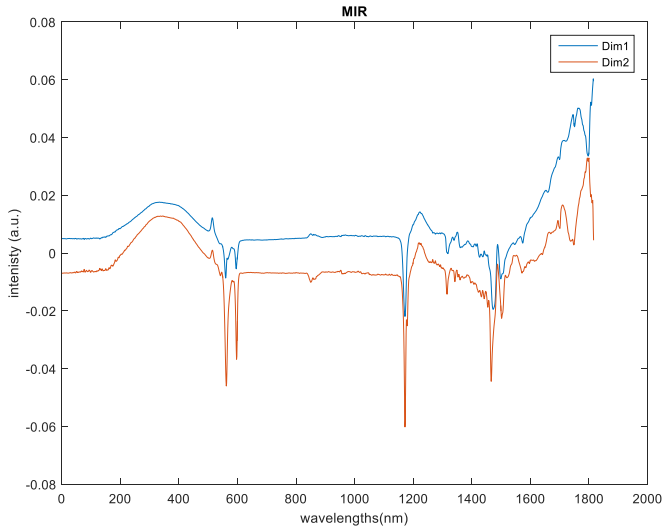

Scatter plot PC1 vs PC2
MDGO38
MBEN34
MFLI38
MGOL59
MFPA35
MFOP62
BCUI80
MSH438
BKMO70
MSTO75
BVRA82
MPRI59
BTES82
MFOO60
MPFI59
MAST70
BPRE82
MLPR80
BLPR82
PC2
PC1


VISNIRMIR
Dim1
Dim2
intenisty (a.u.)
wavelengths(nm)


VIS
Dim1
Dim2
intenisty (a.u.)
wavelengths(nm)

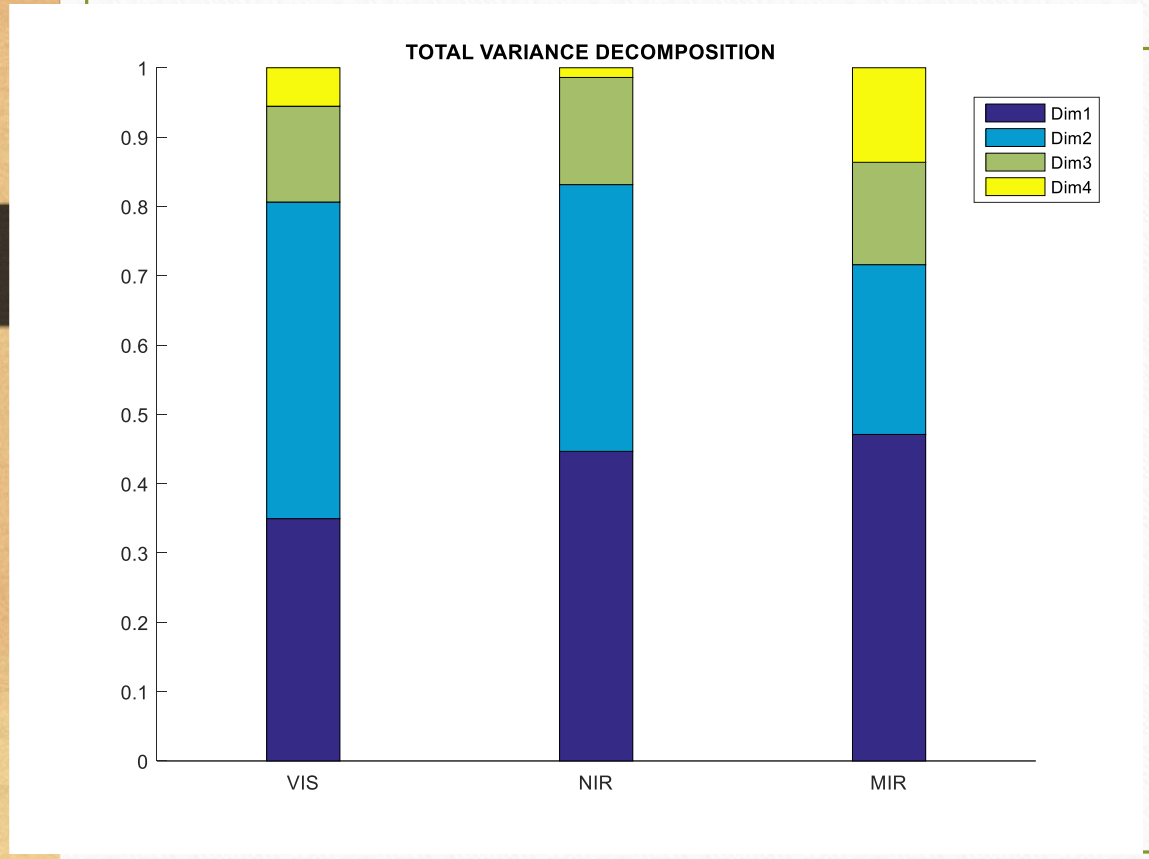

Scatter plot PC1 vs PC2
MDGO38
MBEN34
MFLI38
MGOL59
MFPA35
MFOP62
BCUI80
MSH438
BKMO70
MSTO75
BVRA82
MPRI59
BTES82
MFOO60
MPFI59
MAST70
BPRE82
MLPR80
BLPR82
PC2
PC1

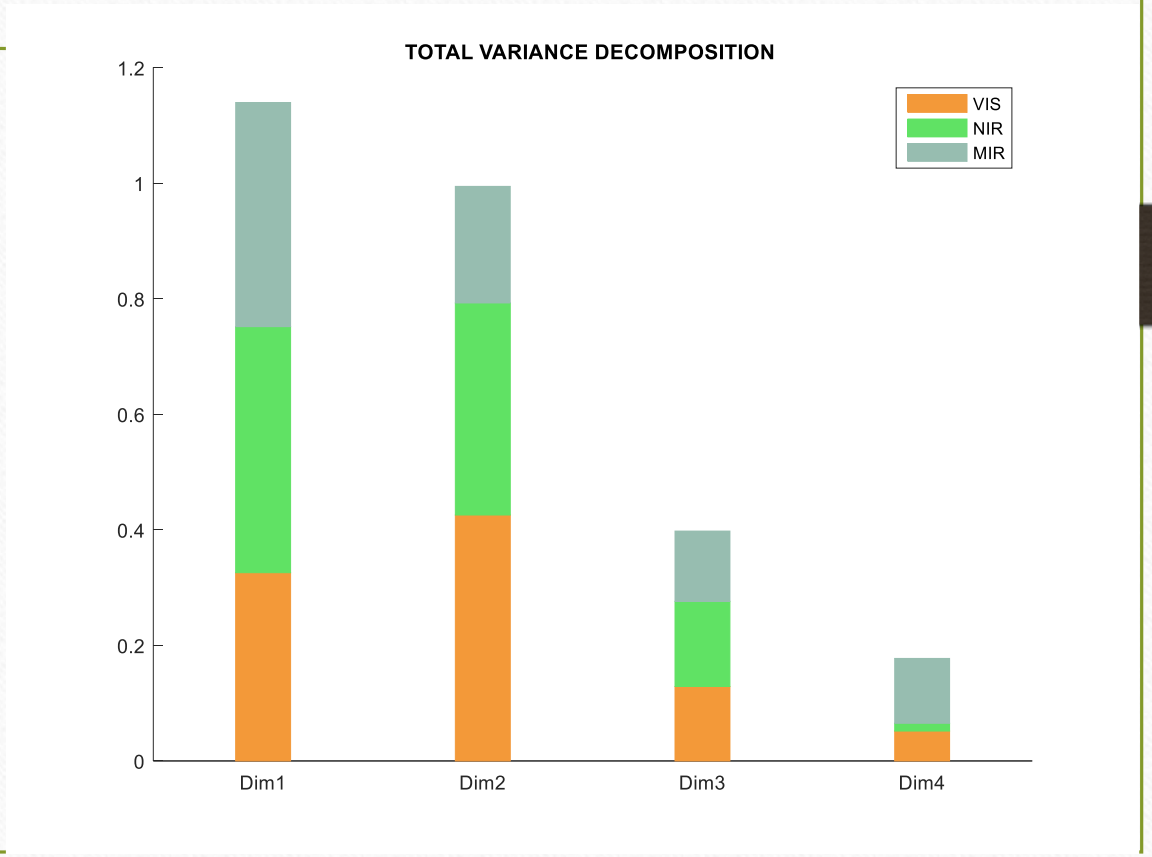

NIR
Dim1
Dim2
intenisty (a.u.)
wavelengths(nm)

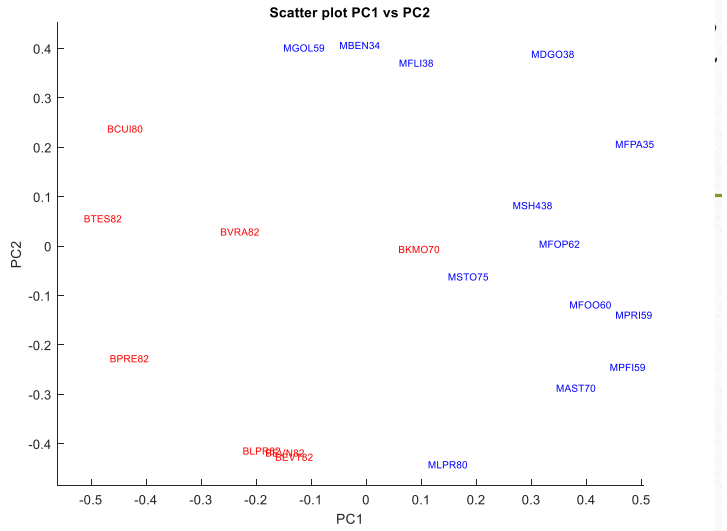

MIR
Dim1
Dim2
intenisty (a.u.)
wavelengths(nm)

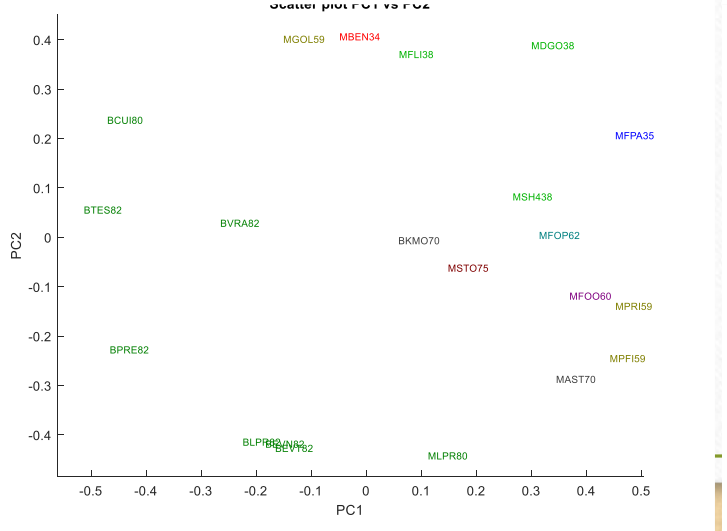
Variance expliqu

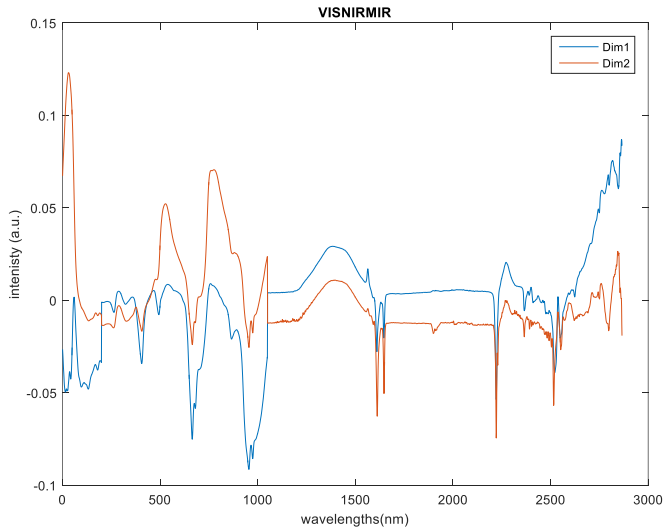
	Inertie Totale	Inertie Totale %
Visible	1.0000	33.3333
NIR	1.0000	33.3333
MIR	1.0000	33.3333

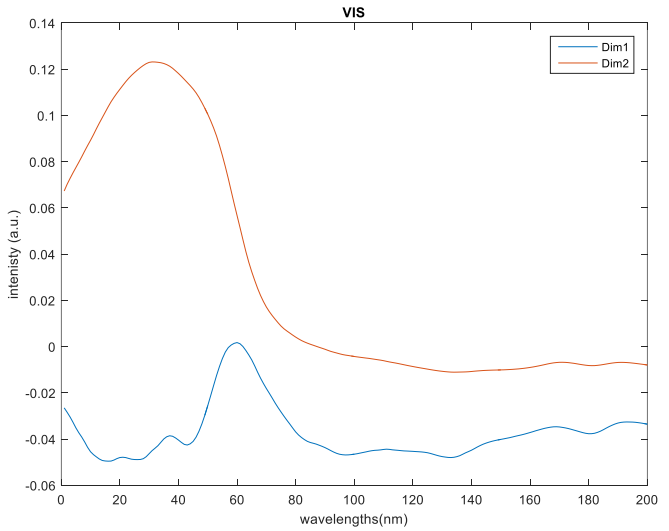


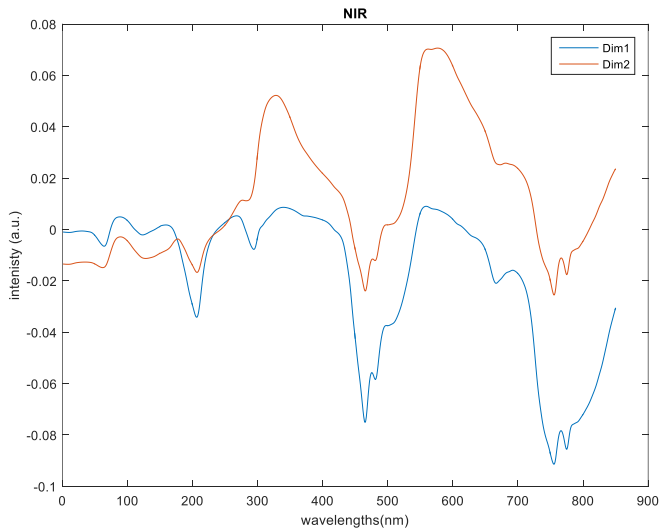


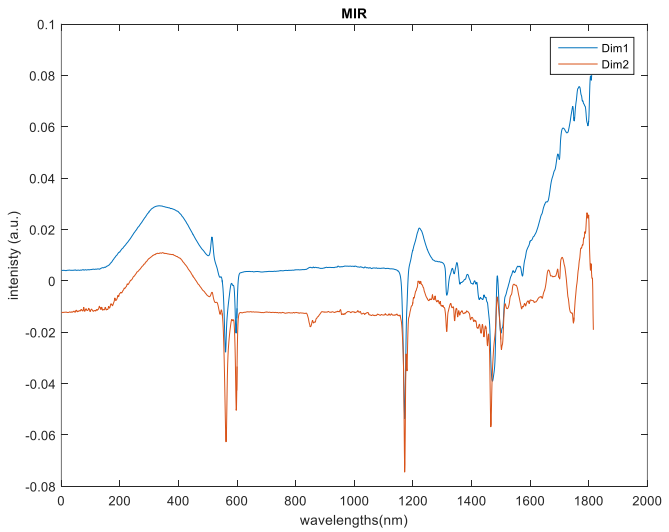

Scatter plot PC1 vs PC2
MGOL59
MBEN34
MFLI38
MDGO38
BCUI80
MFPA35
MSH438
BTES82
BVRA82
BKMO70
MFOP62
MSTO75
MFOO60
MPRI59
BPRE82
MPFI59
MAST70
MLPR80
PC1
PC2


MGOL59
MBEN34
MFLI38
MDGO38
BCUI80
MFPA35
MSH438
BTES82
BVRA82
BKMO70
MFOP62
MSTO75
MFOO60
MPRI59
BPRE82
MPFI59
MAST70
MLPR80
PC1
PC2


VISNIRMIR
Dim1
Dim2
intenisty (a.u.)
wavelengths(nm)


VIS
Dim1
Dim2
intenisty (a.u.)
wavelengths(nm)


NIR
Dim1
Dim2
intenisty (a.u.)
wavelengths(nm)


MIR
Dim1
Dim2
intenisty (a.u.)
wavelengths(nm)

Variance expliqu

	Inertie Totale	Inertie Totale %
Visible	1.4224	26.8526
NIR	1.8737	35.3727
MIR	2.0010	37.7747