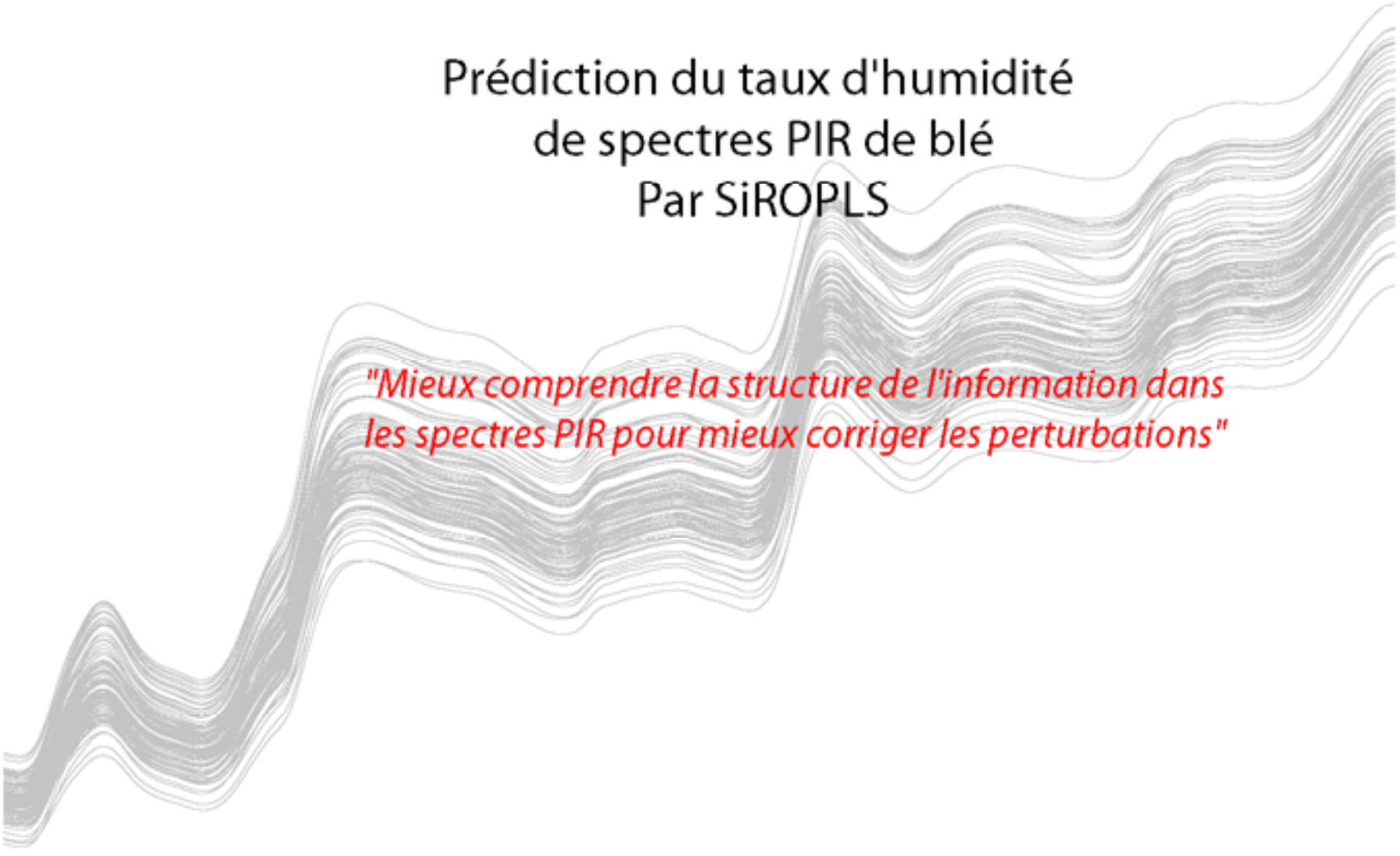
Traitement du signal et généralisation de la régression PLS pour la modélisation de spectres PIR

Thomas VERRON

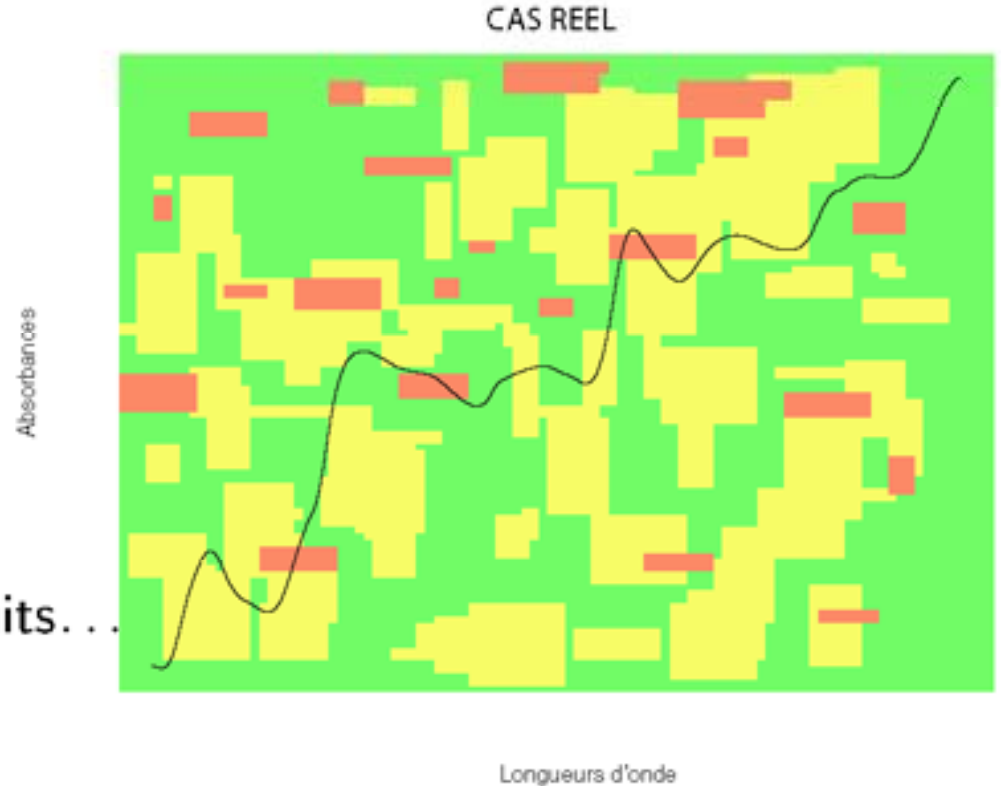
28 novembre 2005



Prédiction du taux d'humidité de spectres PIR de blé Par SiROPLS

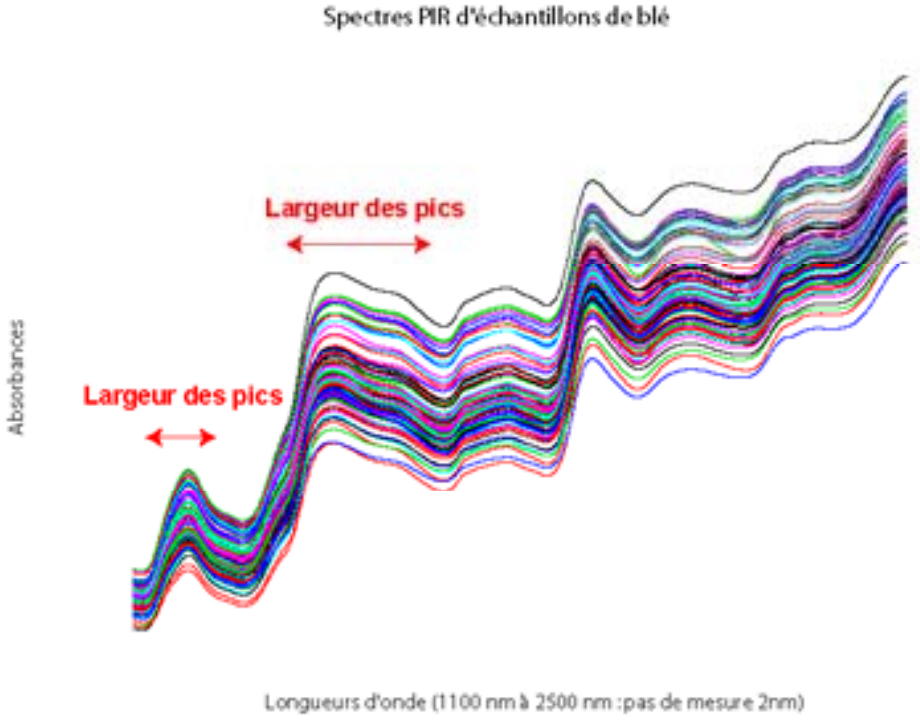
"Mieux comprendre la structure de l'information dans les spectres PIR pour mieux corriger les perturbations"

- Chimiques provenant du constituant recherché,
- Chimiques provenant de la matrice,
- Physiques provenant de la matrice physique : dispersion de la lumière, bruits...



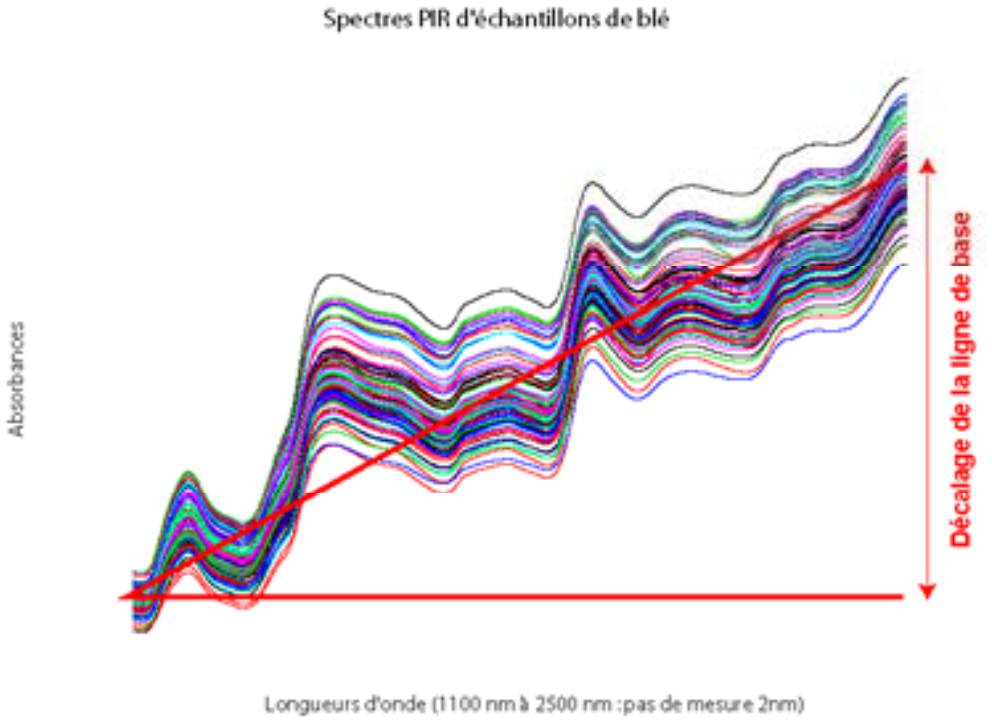
Le mélange d'informations chimiques et physiques engendre :

- ✗ des chevauchements de pics,



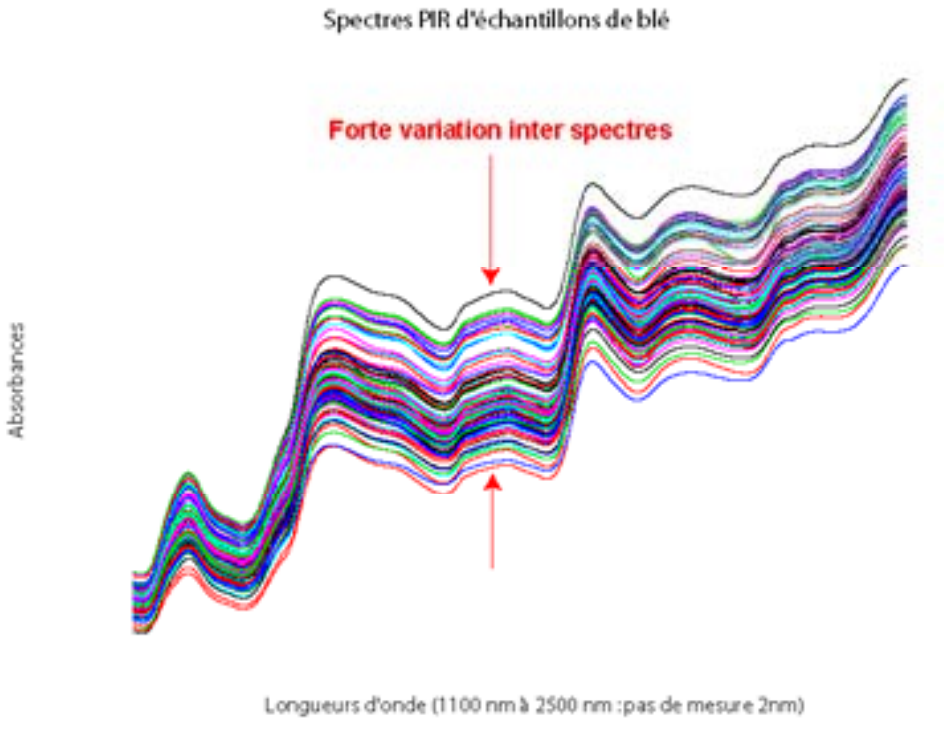
Le mélange d'informations chimiques et physiques engendre :

- ✗ des chevauchements de pics,
- ✗ un décalage de la ligne de base,



Le mélange d'informations chimiques et physiques engendre :

- ✗ des chevauchements de pics,
- ✗ un décalage de la ligne de base,
- ✗ des variations non spécifiques entre les spectres PIR.



Le mélange d'informations chimiques et physiques engendre :

- ✘ des chevauchements de pics,
- ✘ un décalage de la ligne de base,
- ✘ des variations non spécifiques entre les spectres PIR.

⇨ Spectres complexes et embrouillés.

Le mélange d'informations chimiques et physiques engendre :

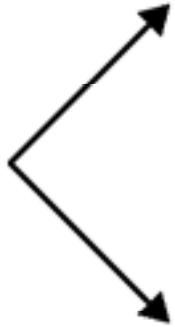
- ✘ des chevauchements de pics,
- ✘ un décalage de la ligne de base,
- ✘ des variations non spécifiques entre les spectres PIR.

⇨ Spectres complexes et embrouillés.

⇩

Méthodes de correction

Deux types d'approches

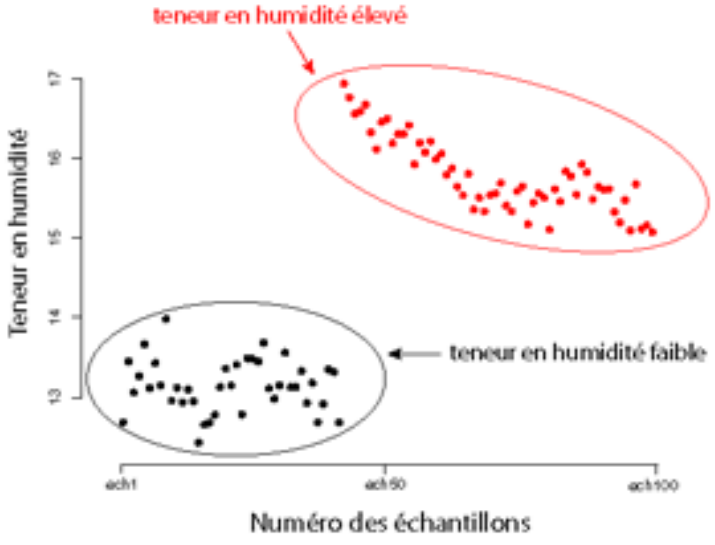


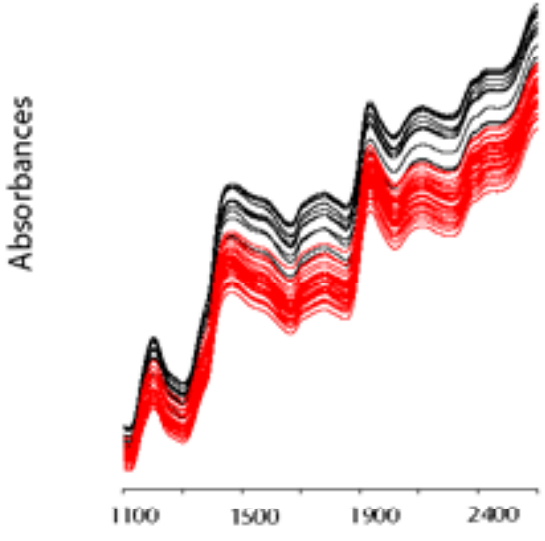
Les approches "Physiques"

- Multiple Scatter Correction
- Standard Normal Variate
- Les dérivées

Les approches "Chimiques"

- Indirect : OSC de Wold...
- Direct : OSC de Fearn, DOSC...
- O-PLS



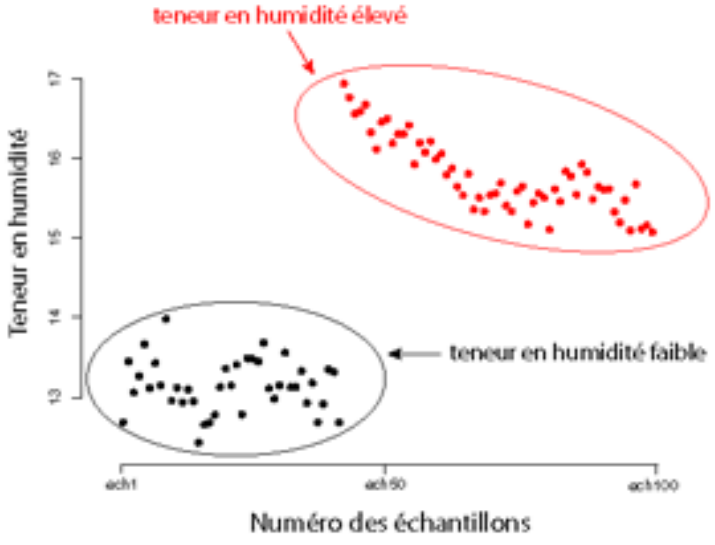


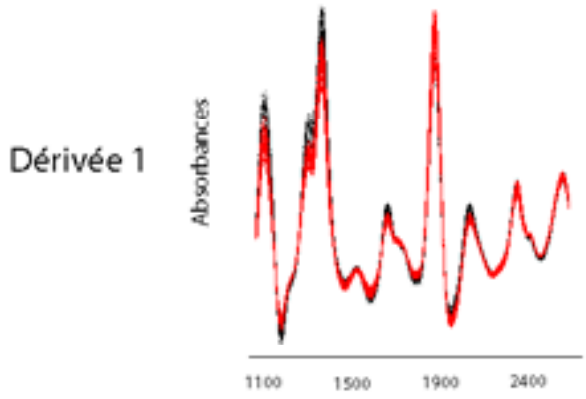
Les approches "Physiques"

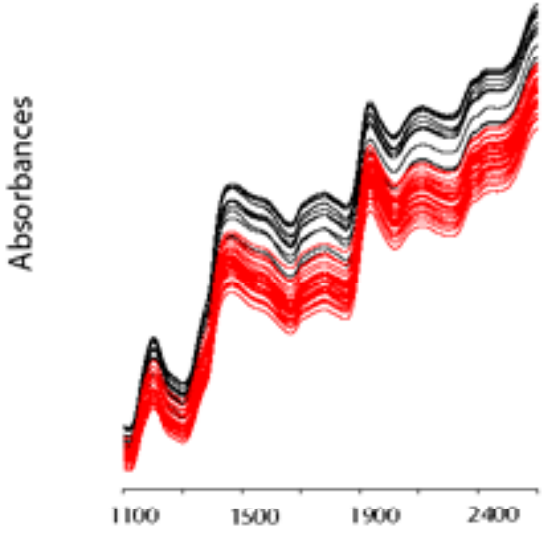
- Multiple Scatter Correction
- Standard Normal Variate
- Les dérivées

Les approches "Chimiques"

- Indirect : OSC de Wold...
- Direct : OSC de Fearn, DOSC...
- O-PLS

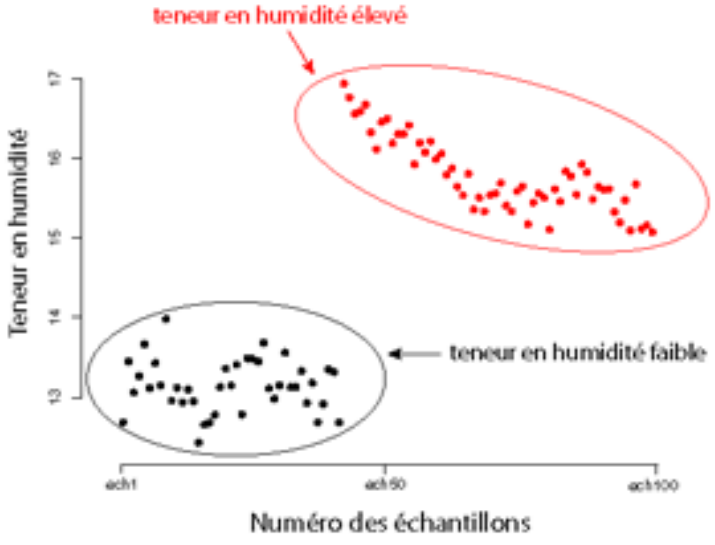


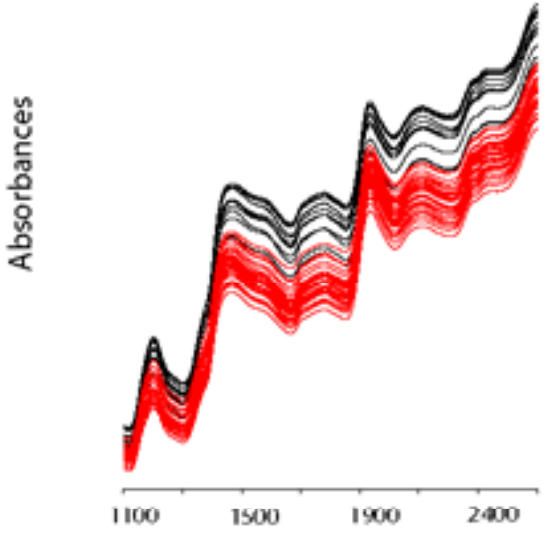




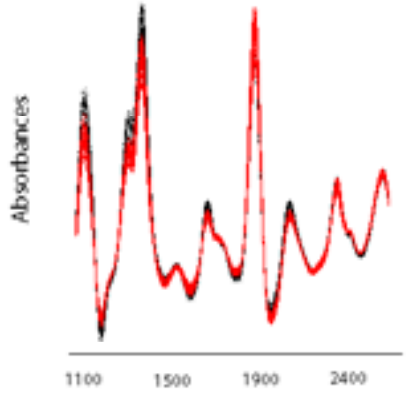
Les approches "Chimiques"

- Indirect : OSC de Wold...
- Direct : OSC de Fearn, DOSC...
- O-PLS

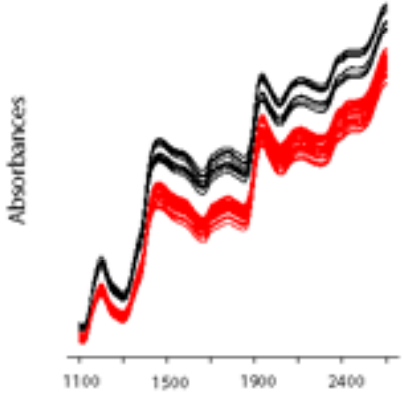




Dérivée 1



O-PLS



Décomposer

Séparer les différentes sources d'informations pour mettre en évidence les caractéristiques du signal pertinentes et non évidentes dans les spectres.

Décomposer

Séparer les différentes sources d'informations pour mettre en évidence les caractéristiques du signal pertinentes et non évidentes dans les spectres.

Les transformées (décompositions)

Traitements du signal

- Ondelettes
- Transformées de Fourier
- ⋮

Décomposition polynomiale

- Bsplines
- ⋮

Décomposer

Séparer les différentes sources d'informations pour mettre en évidence les caractéristiques du signal pertinentes et non évidentes dans les spectres.

Les transformées (décompositions)	Des méthodes corrections
Traitements du signal	
Ondelettes →	WILMA Seuillage de coefficient WOSC...
Transformées de Fourier →	Débruitage Compression...
⋮	
Décomposition polynomiale	
Bsplines →	Compression Régression non linéaire...
⋮	

Décomposer

Séparer les différentes sources d'informations pour mettre en évidence les caractéristiques du signal pertinentes et non évidentes dans les spectres.

Reconstruire

Reconstruire des spectres ne contenant que l'information pertinente par rapport au problème d'optimisation.

Les transformées (décompositions)

Traitements du signal

- Ondelettes
- Transformées de Fourier
- ⋮

Décomposition polynomiale

- Bsplines
- ⋮

Décomposer

Séparer les différentes sources d'informations pour mettre en évidence les caractéristiques du signal pertinentes et non évidentes dans les spectres.

Reconstruire

Reconstruire des spectres ne contenant que l'information pertinente par rapport au problème d'optimisation.

Les transformées (décompositions)

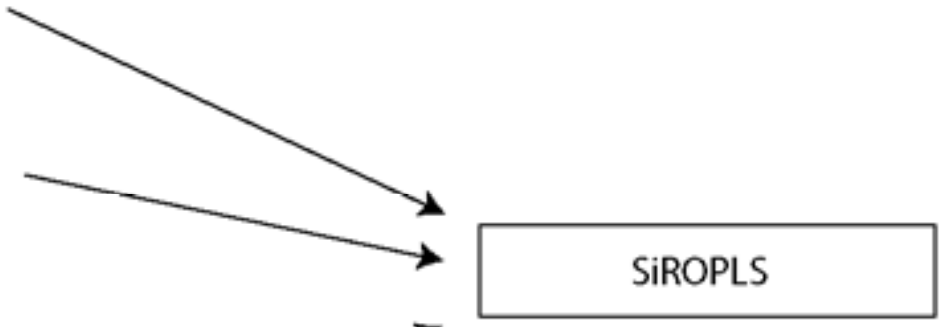
Traitements du signal

- Ondelettes
- Transformées de Fourier
- ⋮

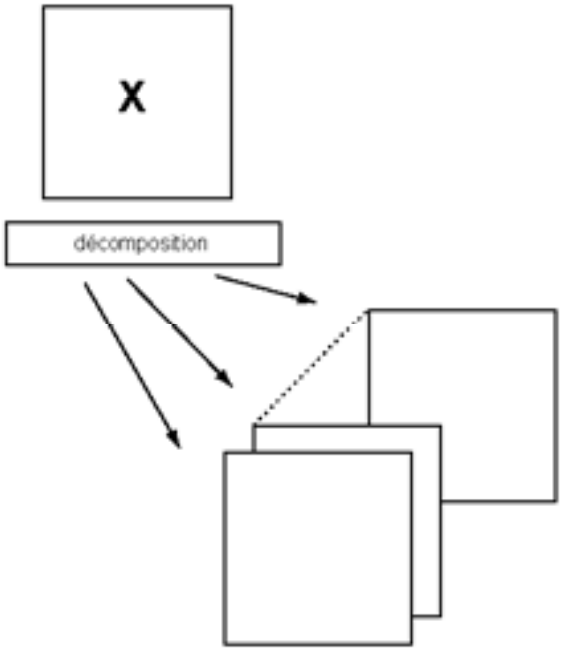
Décomposition polynomiale

- Bsplines
- ⋮

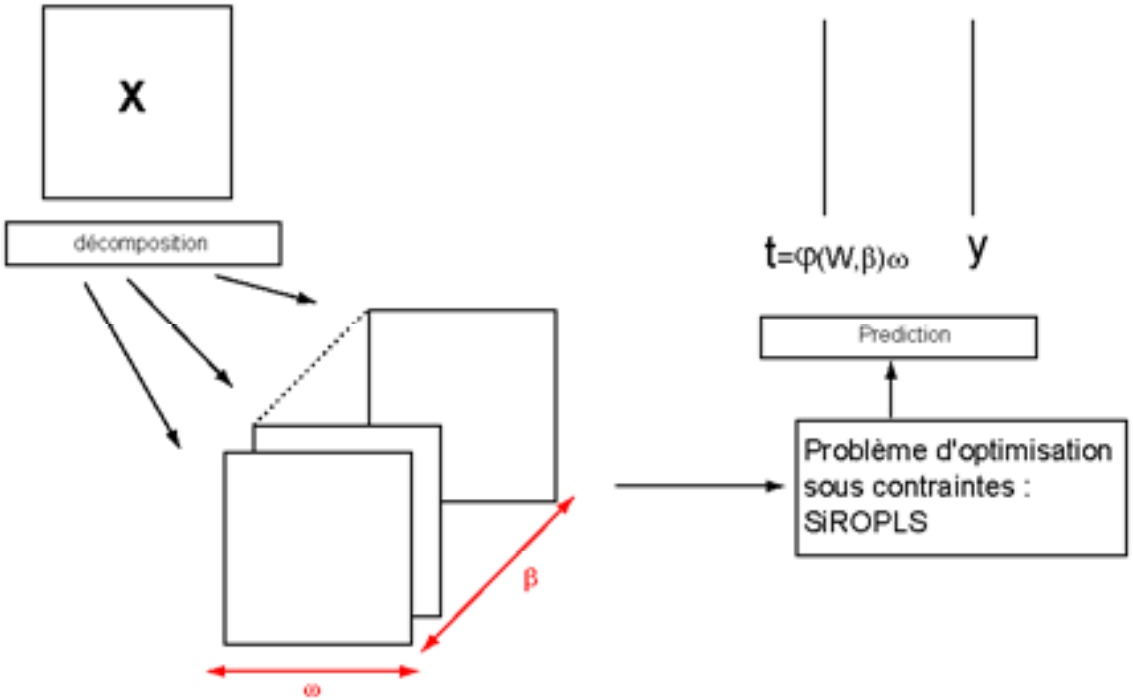
Notre Approche



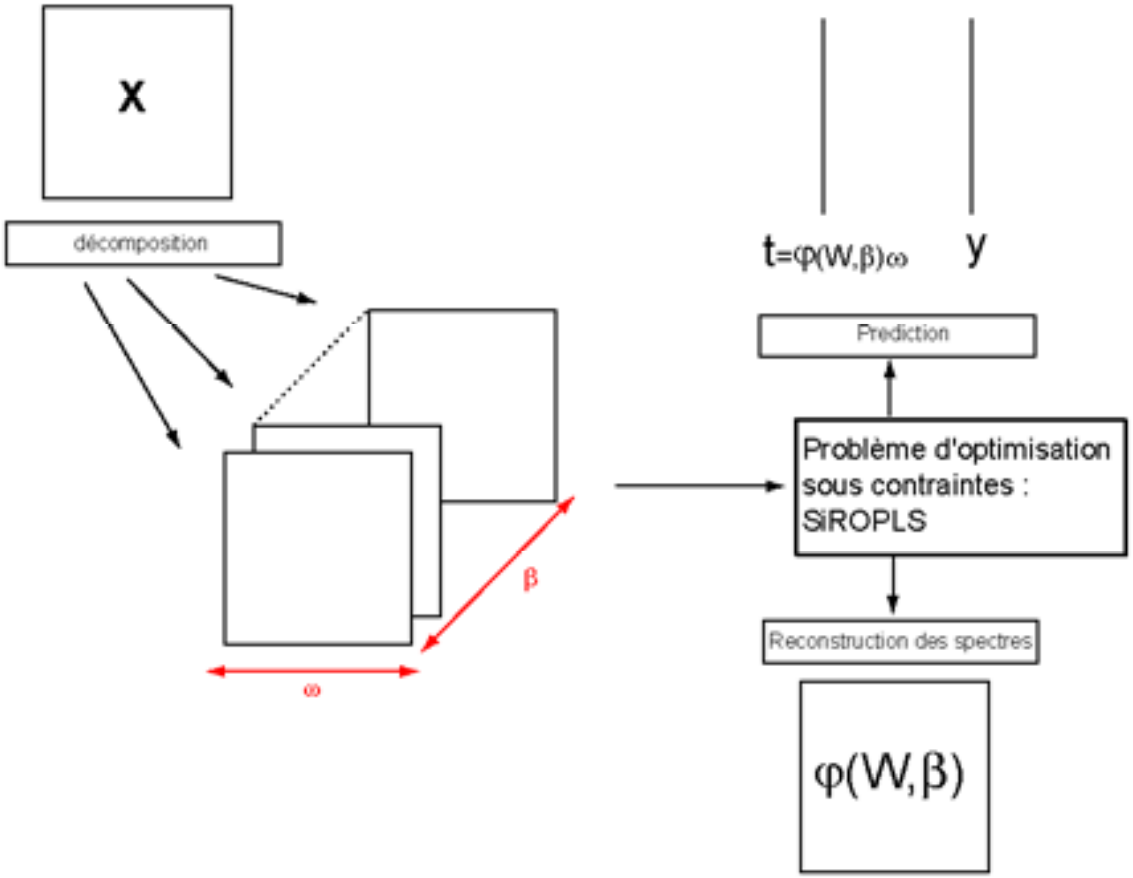
méthode SiROPLS



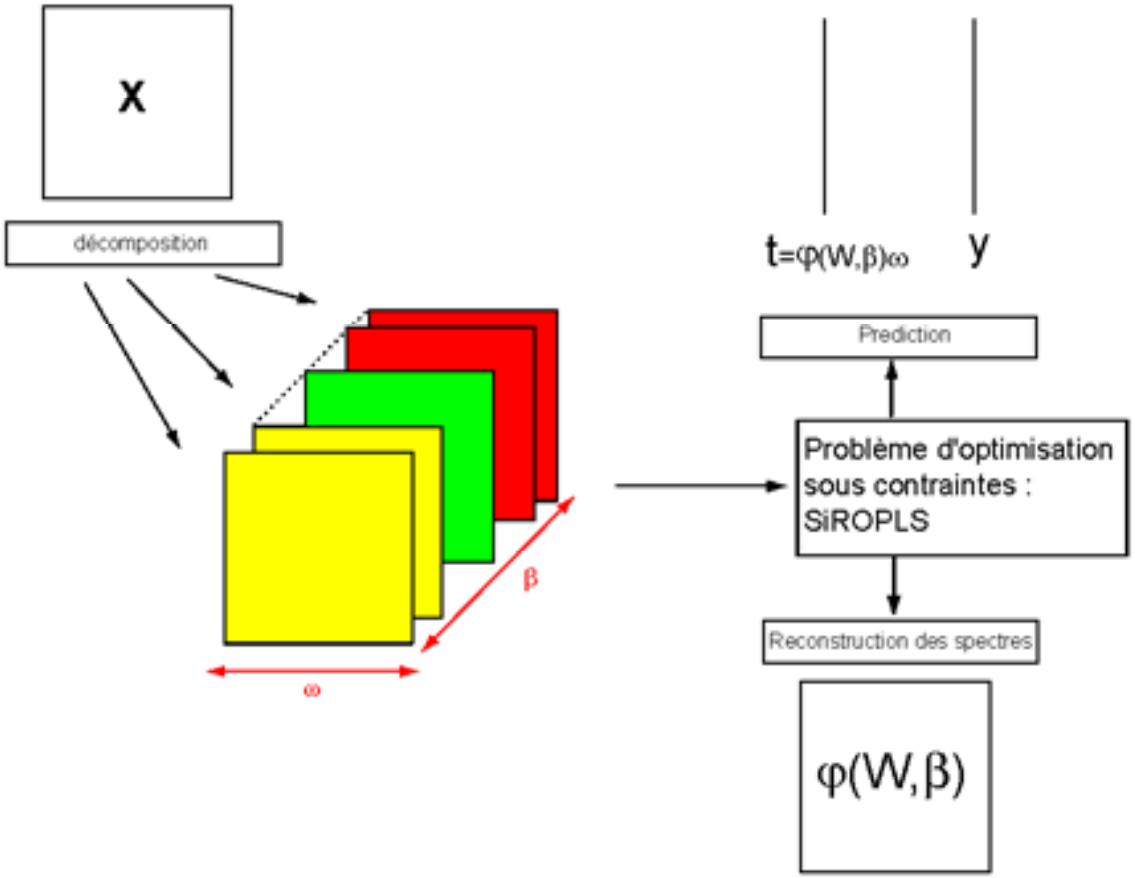
méthode SiROPLS



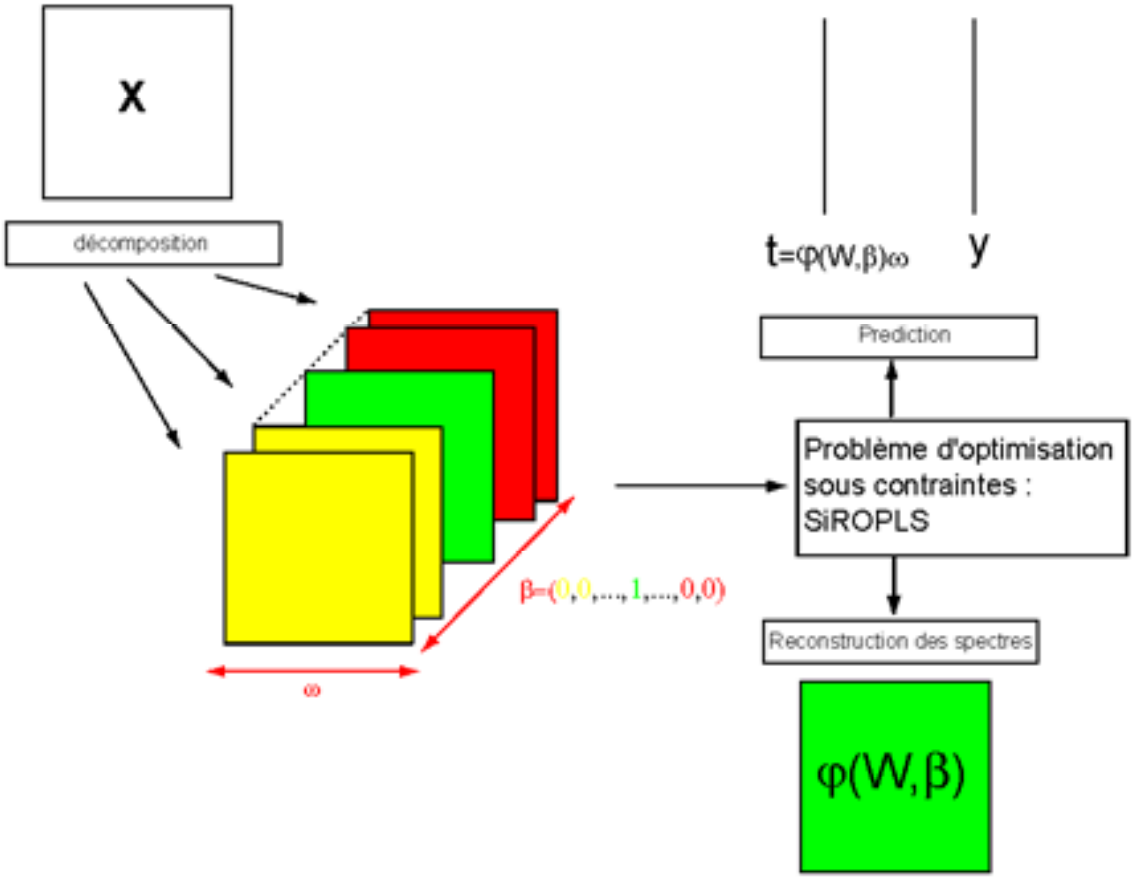
méthode SiROPLS



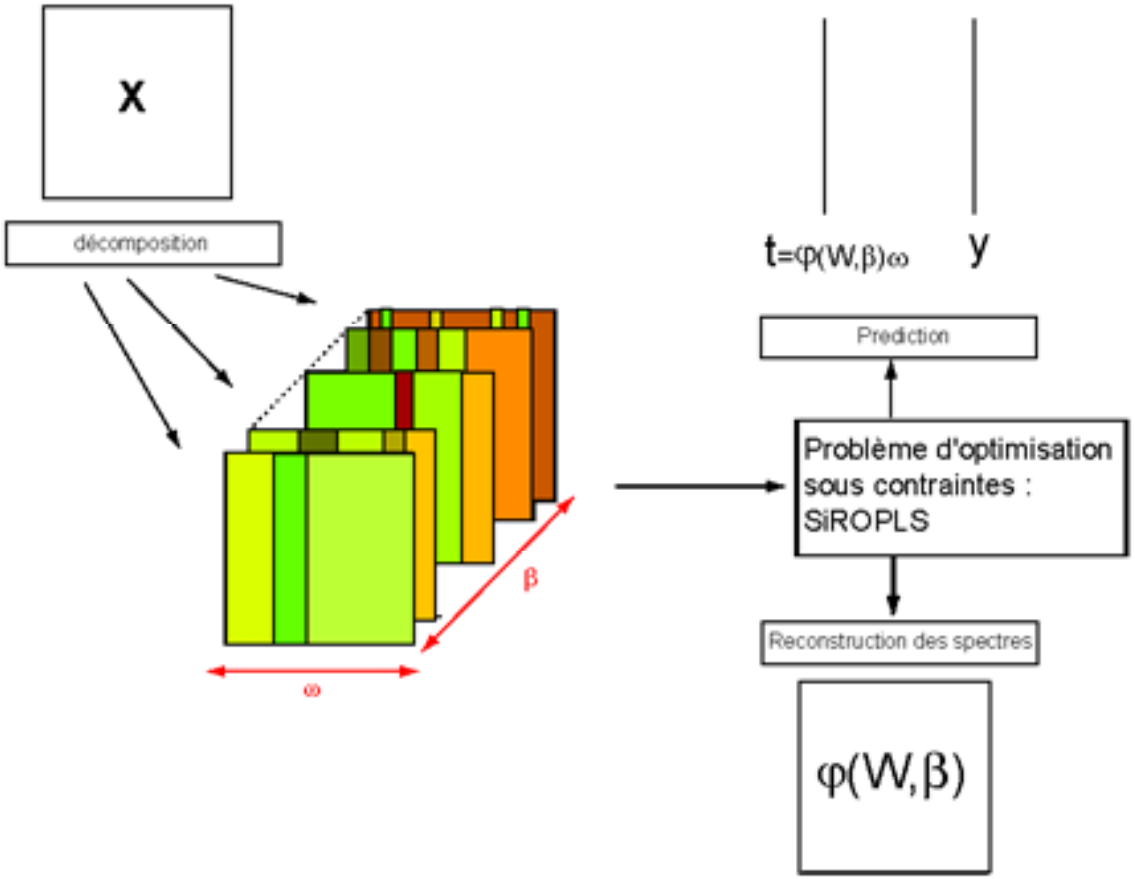
méthode SiROPLS : le cas idéal



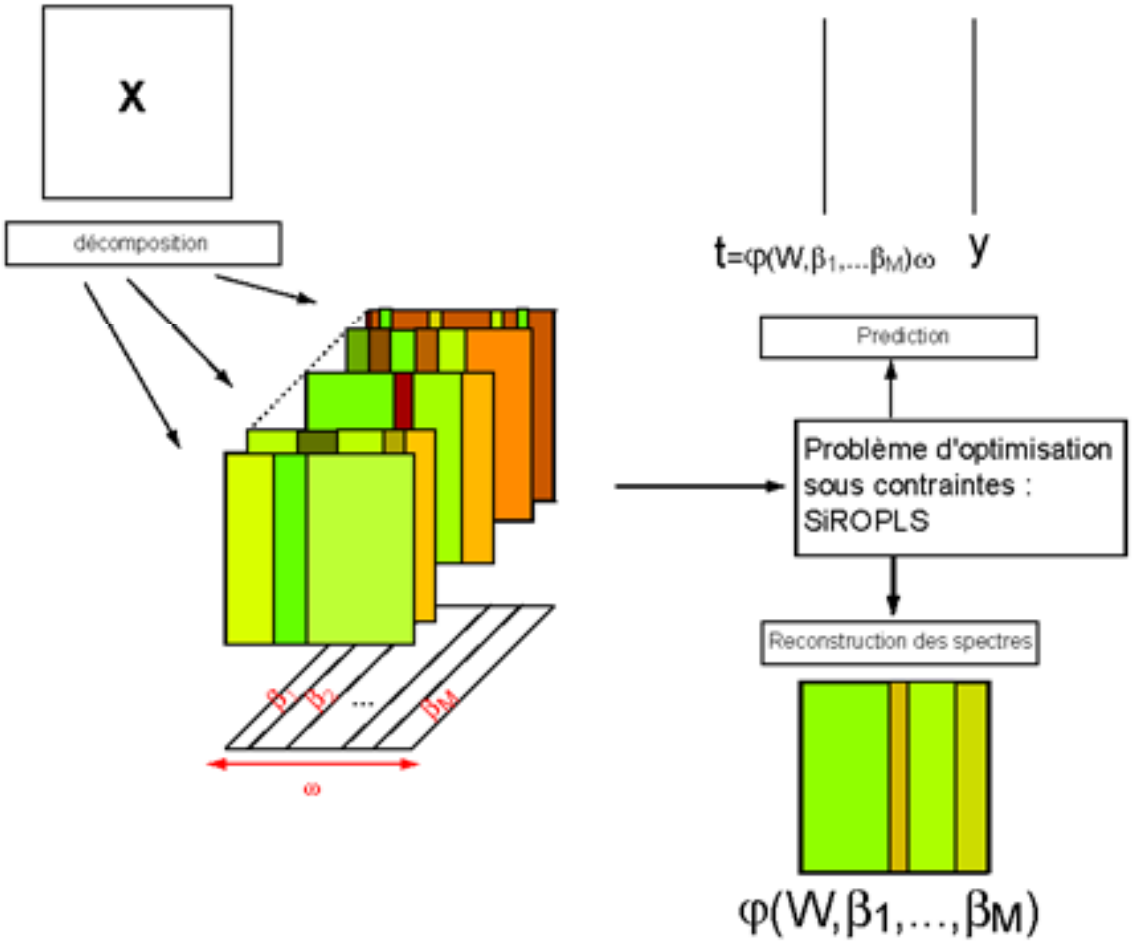
méthode SiROPLS : le cas idéal



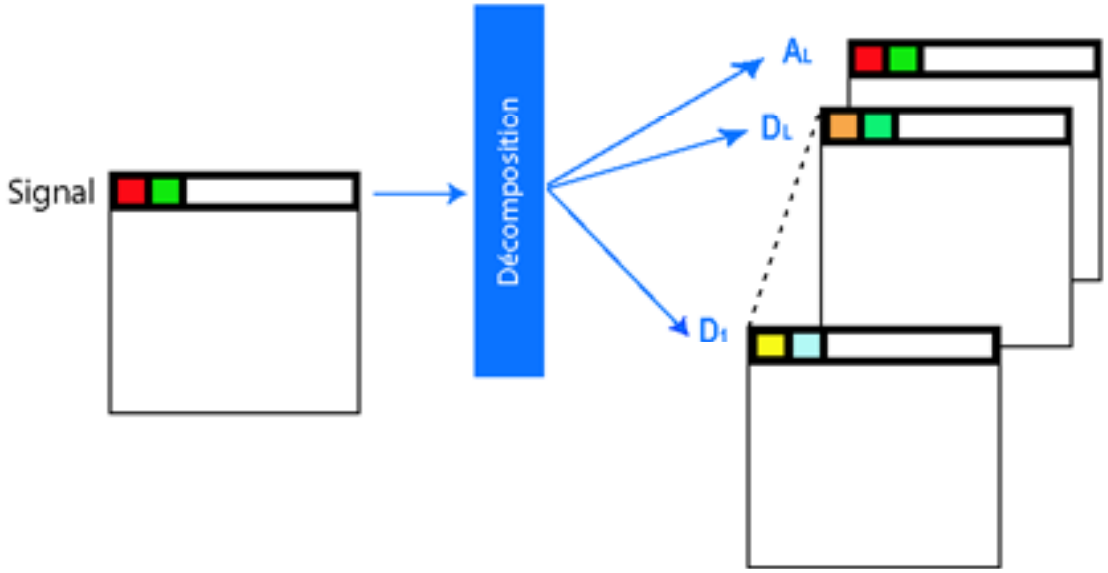
méthode SiROPLS : le cas réel



méthode SiROPLS : le cas réel

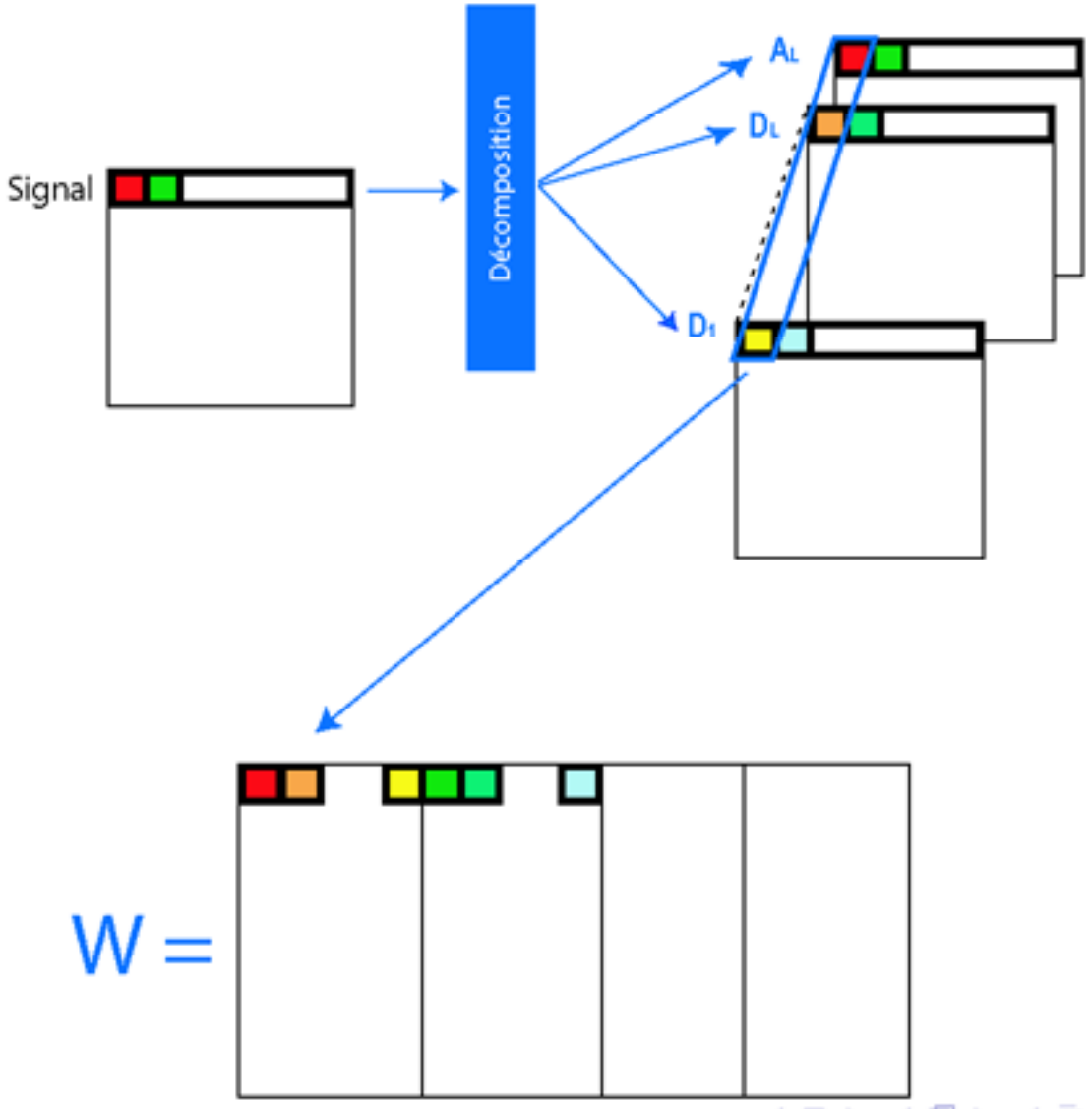


Construction de la matrice W

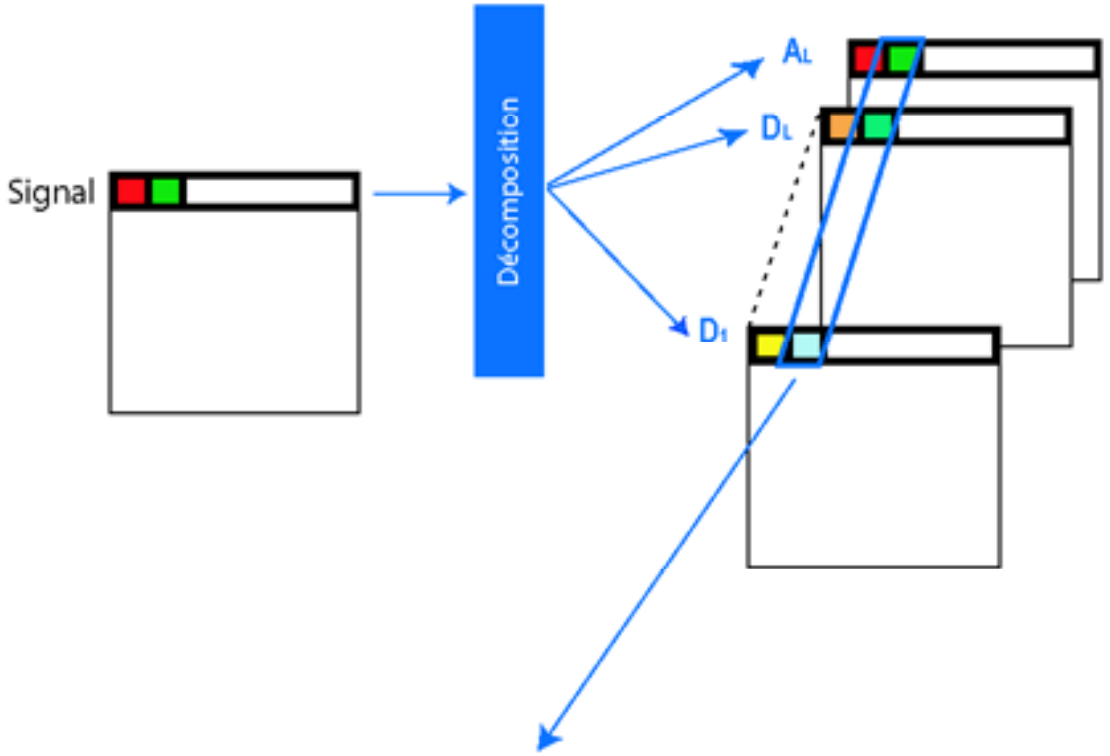


$$W = $$

Construction de la matrice W

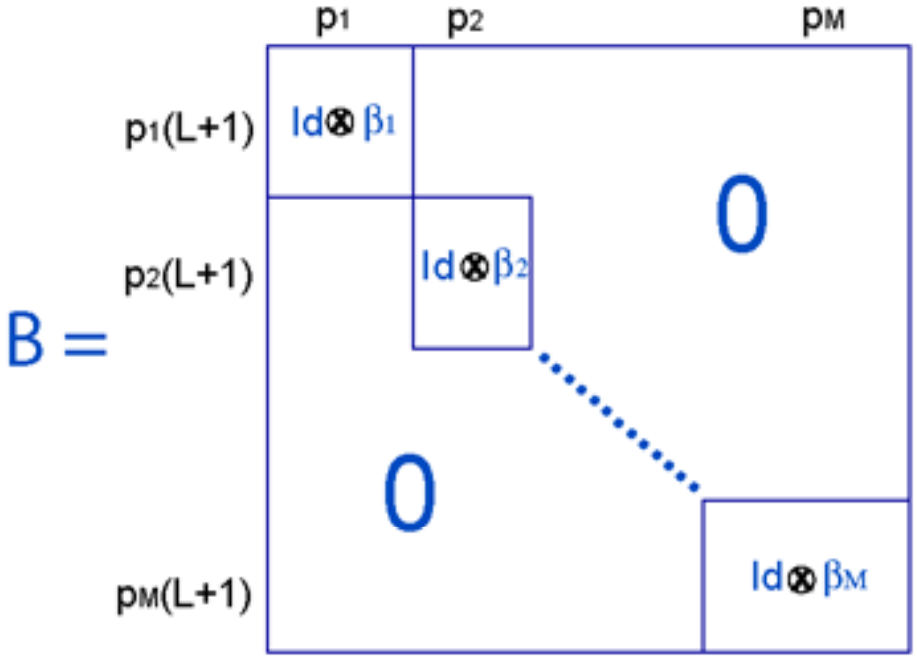


Construction de la matrice W



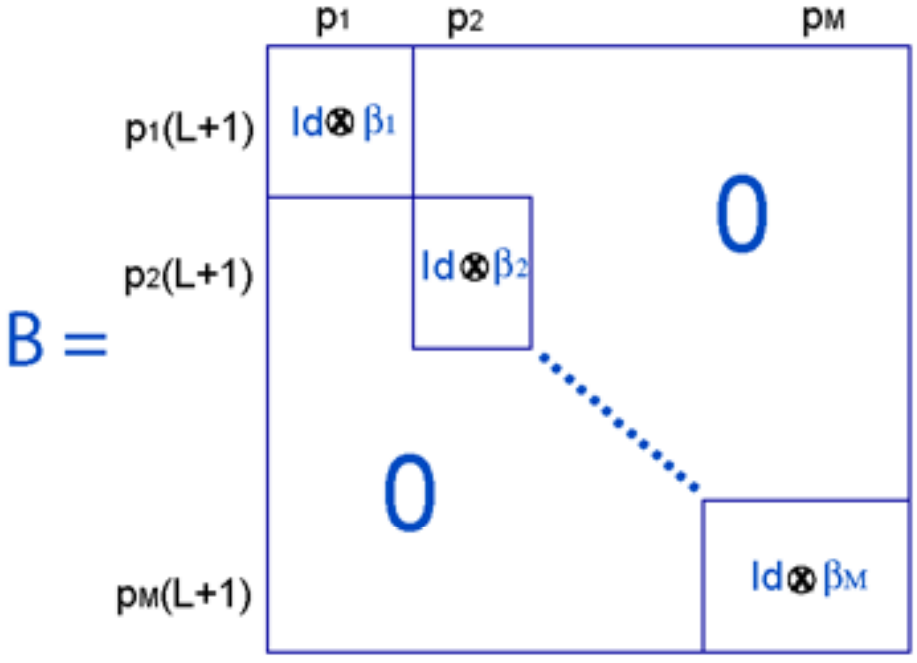
$$W = \begin{bmatrix} & & & & \\ & & & & \\ & & & & \\ & & & & \end{bmatrix}$$

La matrice B contenant les vecteurs β



- $\sum_{j=1}^{M} p_j = p$
- $\varphi(W, \beta_1, \ldots, \beta_M) = WB$
- Si $M = 1$ ⇛ $WB = W(Id \otimes \beta_1)$
- Si $\beta_1 = (1, \ldots, 1)'$, ..., $\beta_M = (1, \ldots, 1)'$ ⇛ $WB = X$.

La matrice B contenant les vecteurs β



- $\sum_{j=1}^{M} p_j = p$
- $\varphi(W, \beta_1, \ldots, \beta_M) = WB$
- Si $M = 1$ ➠ $WB = W(Id \otimes \beta_1)$
- Si $\beta_1 = (1, \ldots, 1)'$, ..., $\beta_M = (1, \ldots, 1)'$ ➠ $WB = X$.

Algorithme SiROPLS

Définition

A la k^{eme} étape, on cherche les vecteurs β_m^k et w_k qui maximisent :

$$cov(t_k, y^{(k)})_D = y^{(k)'} DWBQw_k$$

Sous les contraintes

- $\|w_k\|_Q^2 = 1$
- $\|W_m(Id \otimes \beta_m^k)\|_D^2 = \|X_m^{(k-1)}\|_D^2 \qquad (m = 1, \ldots, M)$
- $t_k' D t_i = 0 \qquad (i = 1, \ldots, k-1)$

Propriétés de l'algorithme

Convergence

L'algorithme génère une série croissante et positive de la fonction objectif. La covariance entre la composante t et le vecteur y converge.

Modèle

Le modèle SiROPLS peut s'écrire en fonction des données initiales :

$$y = W\gamma \qquad \text{avec} \qquad \gamma = r\frac{t' y^{k-1}}{t^T t}$$

r est obtenu en utilisant la transformation de Dayal et MacGregor, et vérifie : $t = P^{\perp}_{(t_1,\ldots,t_{k-1})} WBw = Wr$.

Propriétés de l'algorithme

Convergence

L'algorithme génère une série croissante et positive de la fonction objectif. La covariance entre la composante t et le vecteur y converge.

Modèle

Le modèle SiROPLS peut s'écrire en fonction des données initiales :

$$y = W\gamma \qquad \text{avec} \qquad \gamma = r\frac{t'y^{k-1}}{t^T t}$$

r est obtenu en utilisant la transformation de Dayal et MacGregor, et vérifie : $t = P^{\perp}_{(t_1,\ldots,t_{k-1})} WBw = Wr$.

Les paramètres contrôler par l'utilisateur

- Choix de la décomposition (la transformation et ses paramètres).
- Choix du nombre de niveaux de la décomposition.
- Découpage : nombre et longueur des intervalles pour les β.

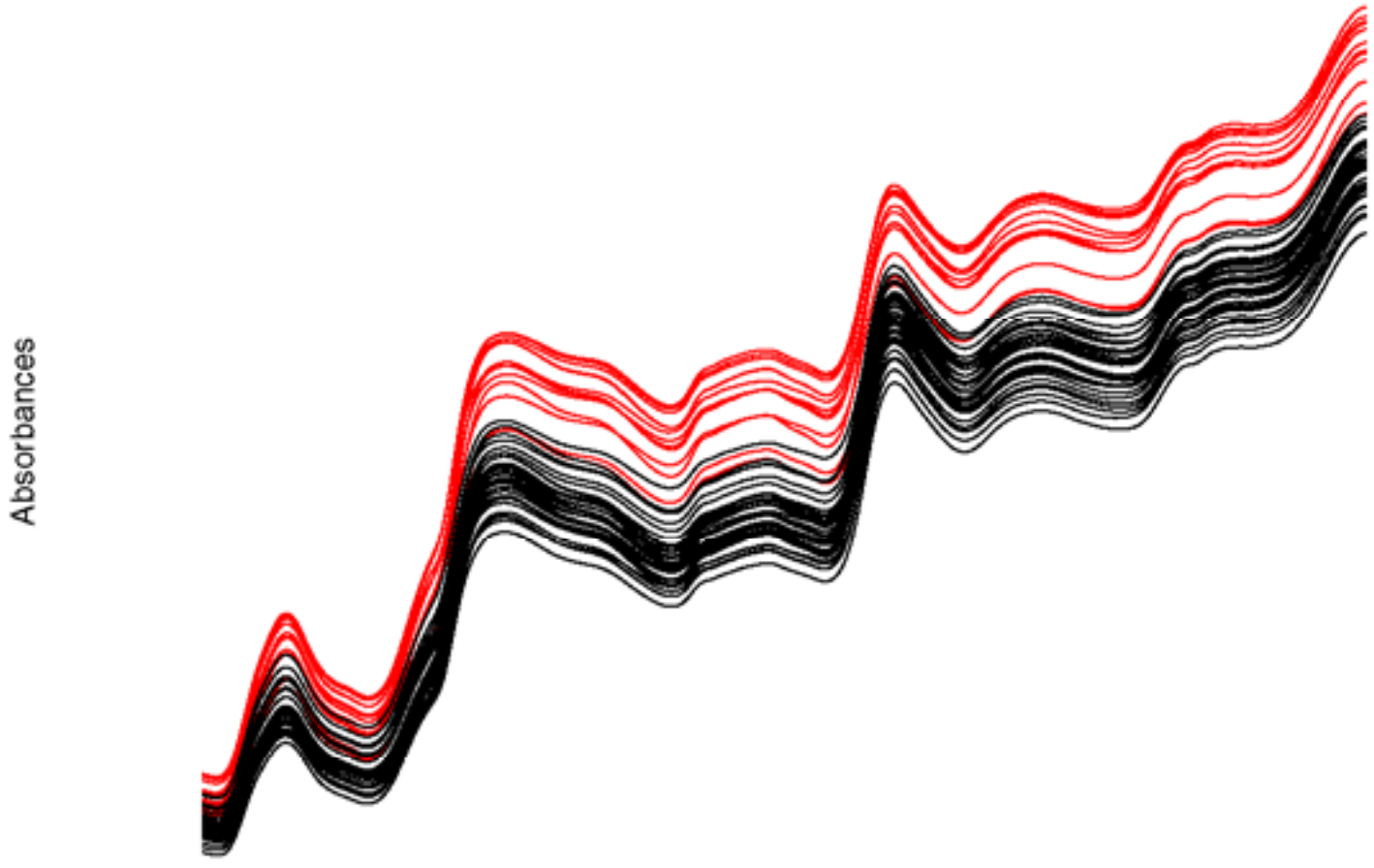
Application : Jeu de données : 100 spectres PIR de blé (données de Kalivas)

1. ondelettes
   - Extension du signal : lissage.
   - Famille d'ondelette Daubechies : db8
   - Nombre de niveaux de décomposition : 8
   - Découpage : fac.tol= 0.55
2. Transformées de Fourier
   - Nombre de niveaux de décomposition : 8
   - Découpage : fac.tol= 0.65
3. B-splines
   - degré : 3
   - nombre noeuds : 12
   - Nombre de niveaux de décomposition : 8
   - Découpage : fac.tol= 0.7

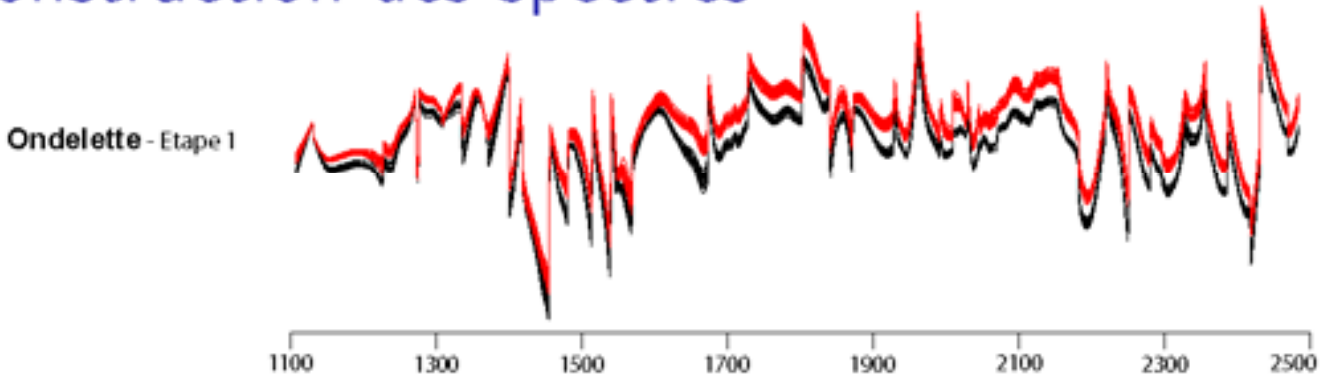
méthodes	RMSEPm	nb comp	méthodes	RMSEPm	nb comp
PCR	0.273	4	Poly-PCR	0.282	3
PCR_TLS	0.296	4	Spline-PLS	0.366	4
PCRS	0.282	3	KNN	1.205	-
PCRS_TLS	0.305	3	LWR (K=50)	0.271	4
PLS	0.271	4	WILMA MLR V db8	0.226	-
Stepwise MLR1 and 5 GA	0.269	5	WILMA MLR R db4	0.240	-
GA-FT	0.272	4	WILMA MLR C coif5	0.235	-
UVE-PCR	0.276	4	WILMA MLR V db9	0.225	-
UVE-PCRS	0.275	3	WILMA PLS R db7	0.238	3
UVE-PLS	0.271	4	WILMA PLS R db1	0.234	3
RCE-PLS	0.288	4	WILMA PLS R sym4	0.245	3
NL-PCR	0.284	6	WILMA PLS C db9	0.250	3
NL-PCRS	0.306	5	SiROPLS ondelettes	0.222	3
NL-UVE-PCR	0.272	4	SiROPLS TF	0.240	3
NL-UVE-PCRS	0.271	3	SiROPLS splines	0.227	3

méthodes	RMSEPm	nb comp	méthodes	RMSEPm	nb comp
PCR	0.273	4	Poly-PCR	0.282	3
PCR_TLS	0.296	4	Spline-PLS	0.366	4
PCRS	0.282	3	KNN	1.205	-
PCRS_TLS	0.305	3	LWR (K−50)	0.271	4
PLS	0.271	4	WILMA MLR V db8	0.226	-
Stepwise MLR1 and 5 GA	0.269	5	WILMA MLR R db4	0.240	-
GA-FT	0.272	4	WILMA MLR C coif5	0.235	-
UVE-PCR	0.276	4	WILMA MLR V db9	0.225	-
UVE-PCRS	0.275	3	WILMA PLS R db7	0.238	3
UVE-PLS	0.271	4	WILMA PLS R db1	0.234	3
RCE-PLS	0.288	4	WILMA PLS R sym4	0.245	3
NL-PCR	0.284	6	WILMA PLS C db9	0.250	3
NL-PCRS	0.306	5	SiROPLS ondelettes	0.222	3
NL-UVE-PCR	0.272	4	SiROPLS TF	0.240	3
NL-UVE-PCRS	0.271	3	SiROPLS splines	0.227	3

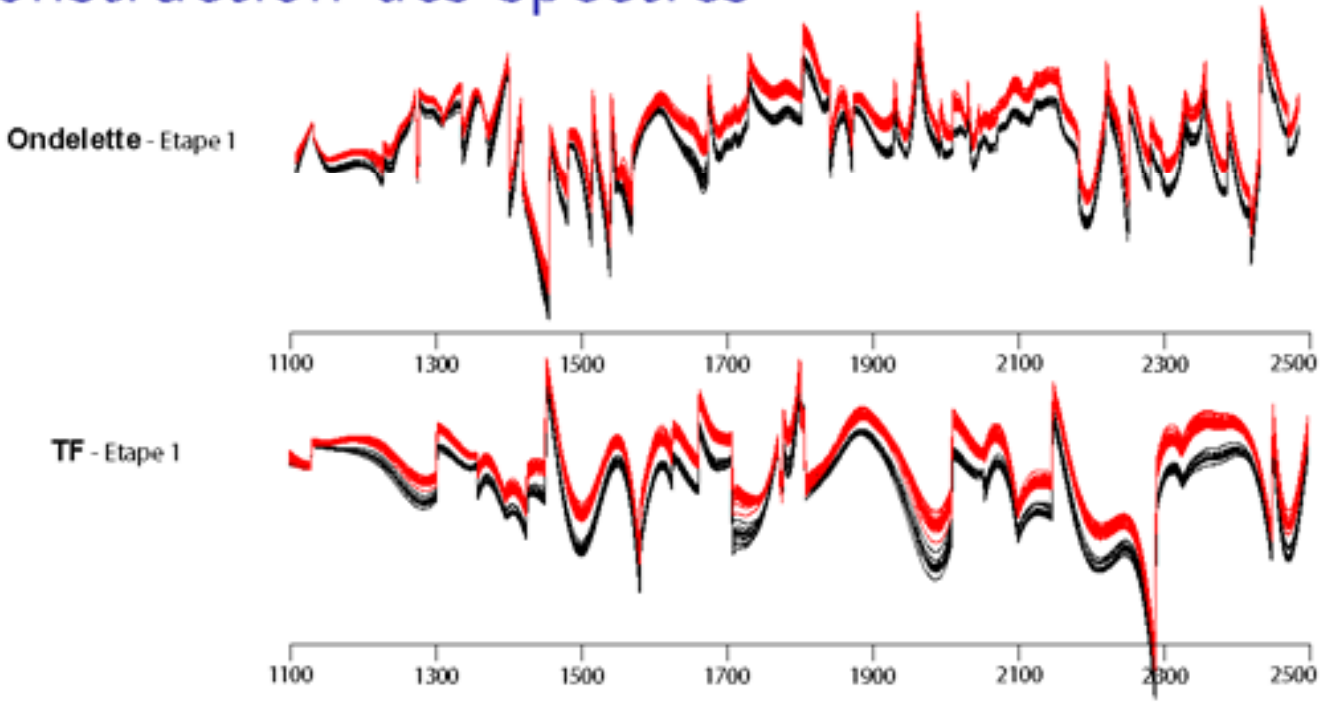
Spectres PIR de blé



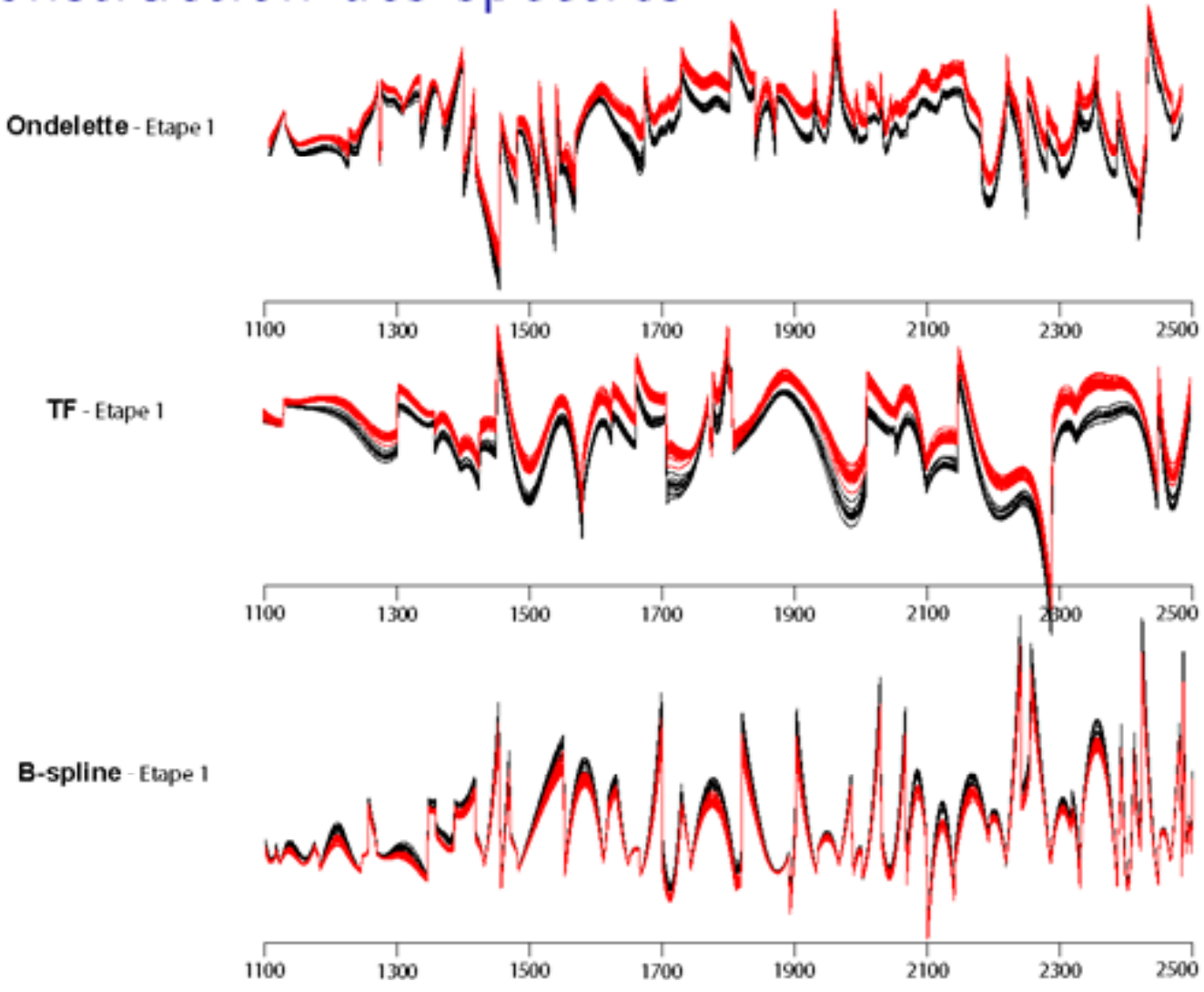
Reconstruction des spectres



Reconstruction des spectres



Reconstruction des spectres



Conclusions

Avantages :

- Une méthode de reconstruction générale.
- Une correction auto-adaptative incluse dans le critère de modélisation.
- Amélioration de la prédiction.
- Diminution du nombre de composantes du modèle optimal.
- Peu de paramètres à fixer.
- Possibilité de visualiser les spectres corrigés.

Inconvénients :

- Temps de calcul important.
- Choix de paramètres délicats.

Conclusions

Avantages :

- Une méthode de reconstruction générale.
- Une correction auto-adaptative incluse dans le critère de modélisation.
- Amélioration de la prédiction.
- Diminution du nombre de composantes du modèle optimal.
- Peu de paramètres à fixer.
- Possibilité de visualiser les spectres corrigés.

Inconvénients :

- Temps de calcul important.
- Choix de paramètres délicats.