

---

# LIBS et chimiométrie

*Journées LIBS 2009*  
*Bordeaux – 18 mai 2009*

Jean-Baptiste Sirven

CEA Saclay  
Laboratoire Réactivité des Surfaces et Interfaces

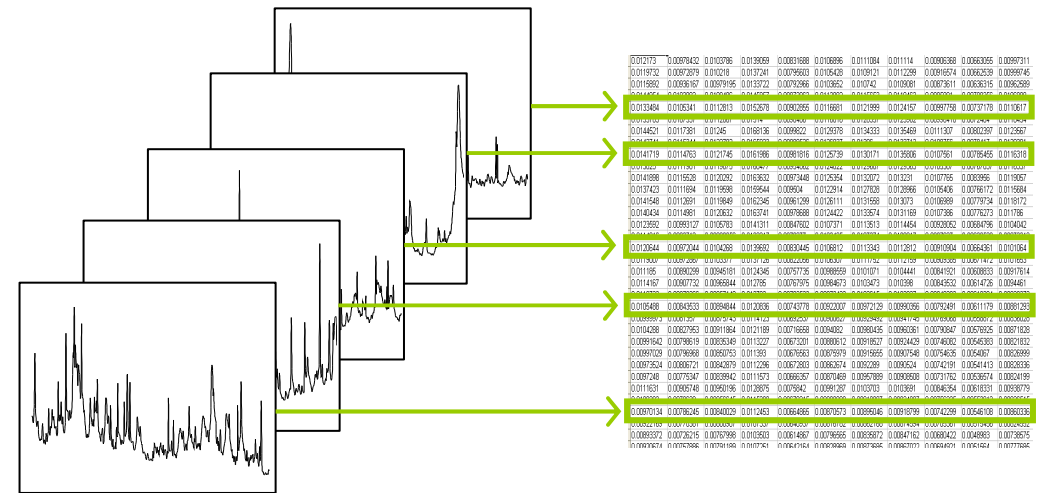
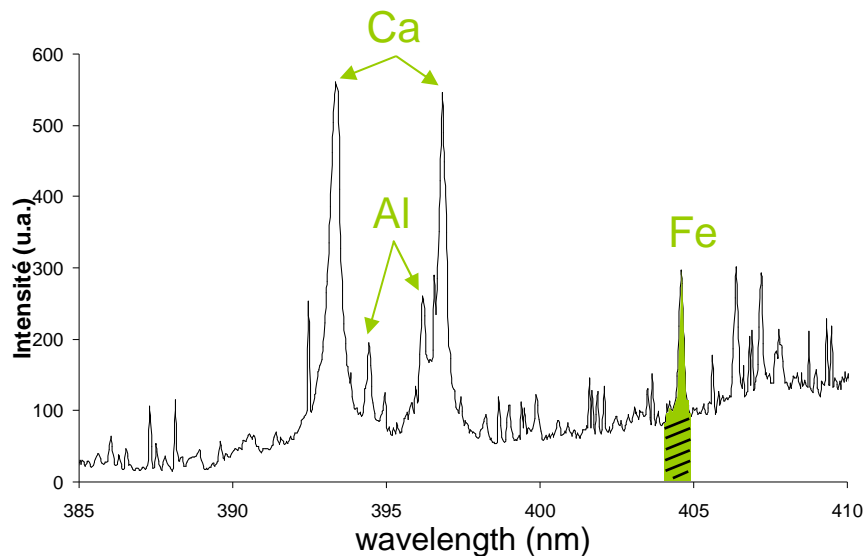


- ✚ Les spectres LIBS contiennent beaucoup d'information :
  - **Chimique** : présence d'éléments, de molécules, concentrations...
  - **Physique** : Bremsstrahlung, température / densité électroniques, auto-absorption...
  - **Instrumentale** : bruit, réponse du spectromètre, dérives éventuelles (température, longueur d'onde...)...
  
- ✚ ... en particulier celle qui nous intéresse (?)
- ✚ La **chimiométrie** est la mise en œuvre d'un ensemble de techniques visant à extraire cette information des données expérimentales
  
- ✚ Une **analyse chimiométrique** fait intervenir :
  - Un instrument (+ protocole de mesure) → sensibilité, reproductibilité
  - Une base de données → représentativité, mesures de référence
  - Un modèle → justesse, précision
  
- ✚ Les défaillances de l'un de ces composants peuvent être compensées par les 2 autres (dans une certaine mesure)

# Traitement des données



- Comment se présentent les données ?
- Quelles sont les informations recherchées ?
  - Analyse qualitative** : identification / classification des échantillons
  - Analyse quantitative** : mesure de la composition élémentaire

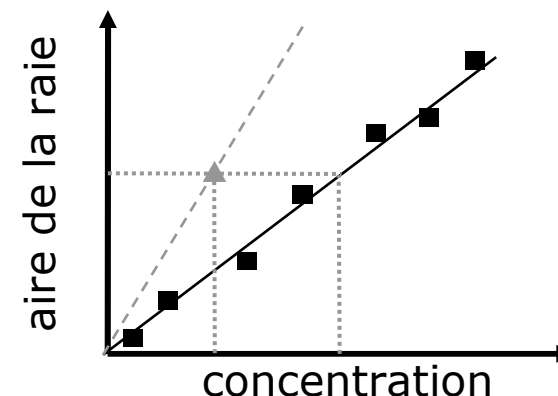
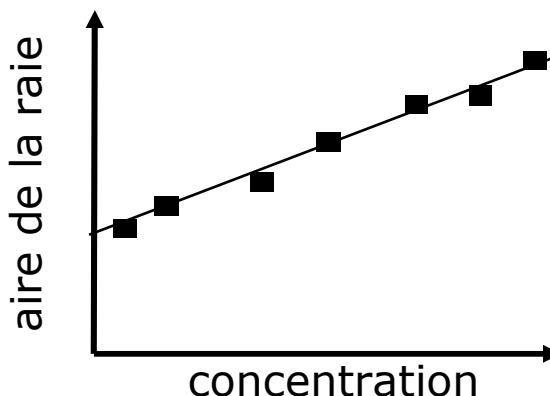
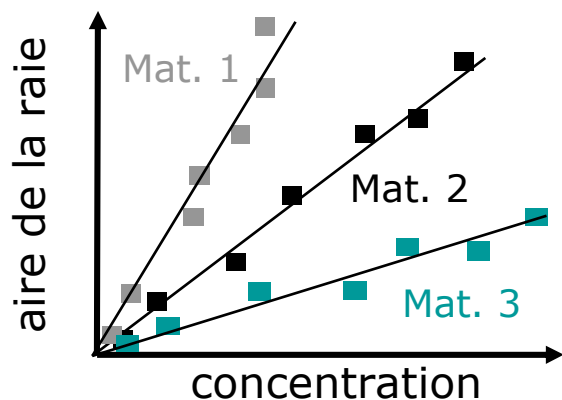


- Traitement mathématique / statistique des données expérimentales → analyse **multivariée**

# Apport des méthodes multivariées



- Les méthodes univariées pour l'analyse qualitative et quantitative :
  - Sont sensibles aux **effets de matrice**
  - Sont soumises aux **interférences spectrales**
  - Ne permettent pas de détecter les **échantillons aberrants**



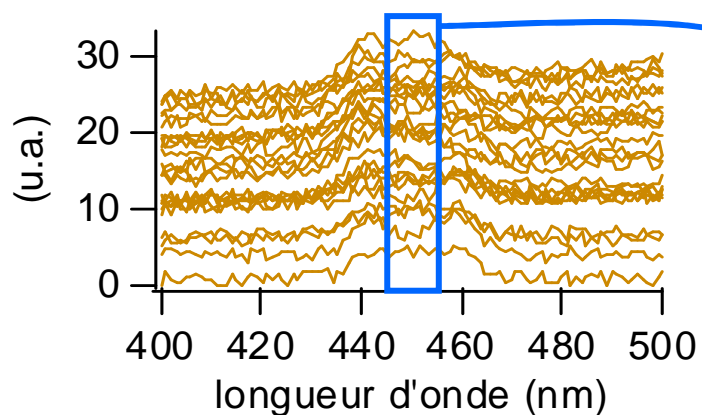
- Apport des **méthodes multivariées** :
  - Prise en compte d'une plus grande quantité de données expérimentales → effet de moyenne
  - Outils pour en extraire l'information recherchée et évaluer sa qualité

# Analyse univariée VS multivariée

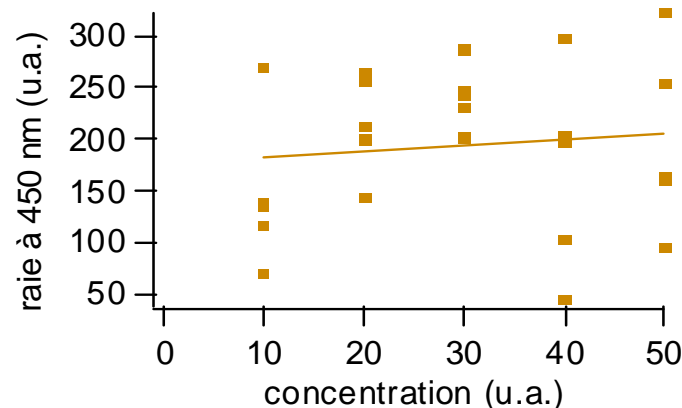


- Simulation de données spectrales inexploitable par une droite d'étalonnage  $signal = f(concentration)$

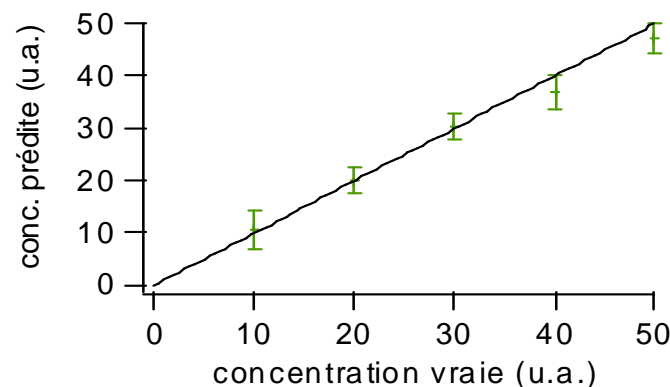
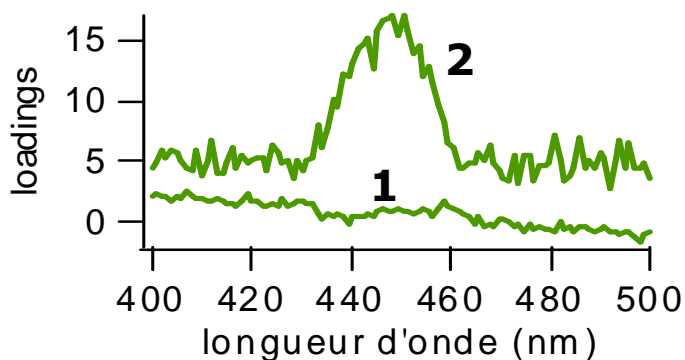
**Spectres expérimentaux**



**« Droite d'étalonnage »**



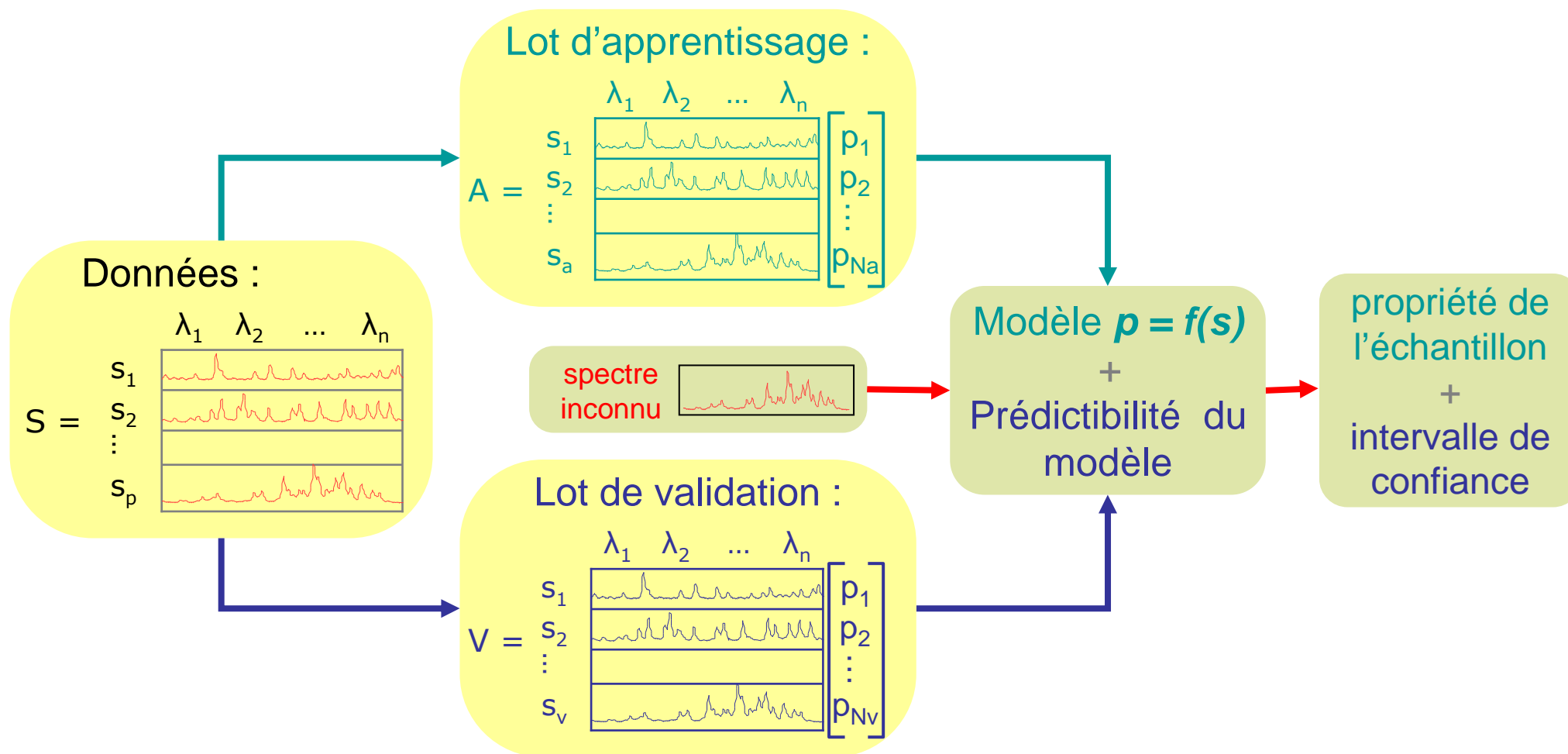
**Régression PLS**



# Méthodes supervisées $\Rightarrow$ prédiction



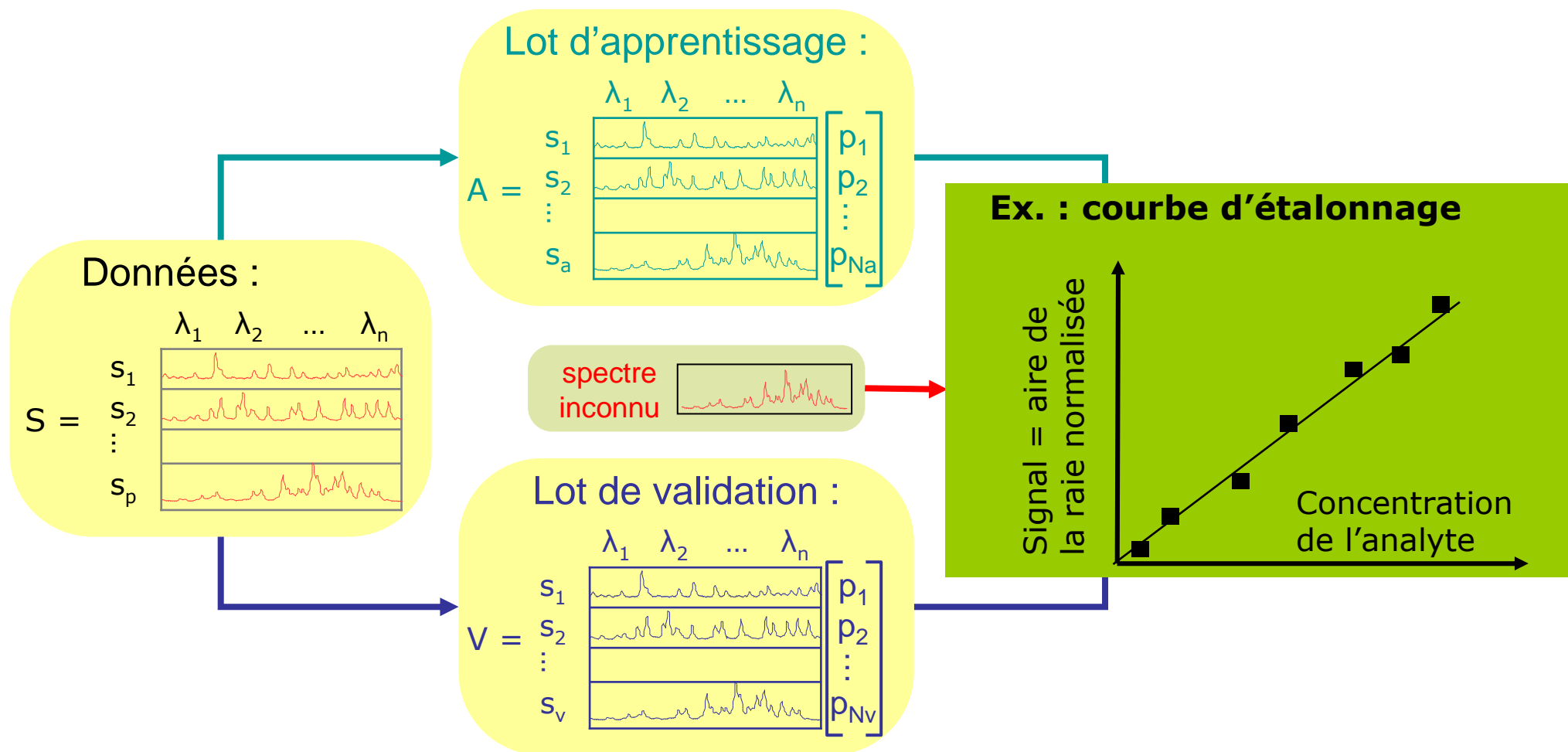
- On cherche à **prédire** une propriété de l'échantillon **p** (qualitative ou quantitative) à partir d'un **modèle**



# Méthodes supervisées $\Rightarrow$ prédiction



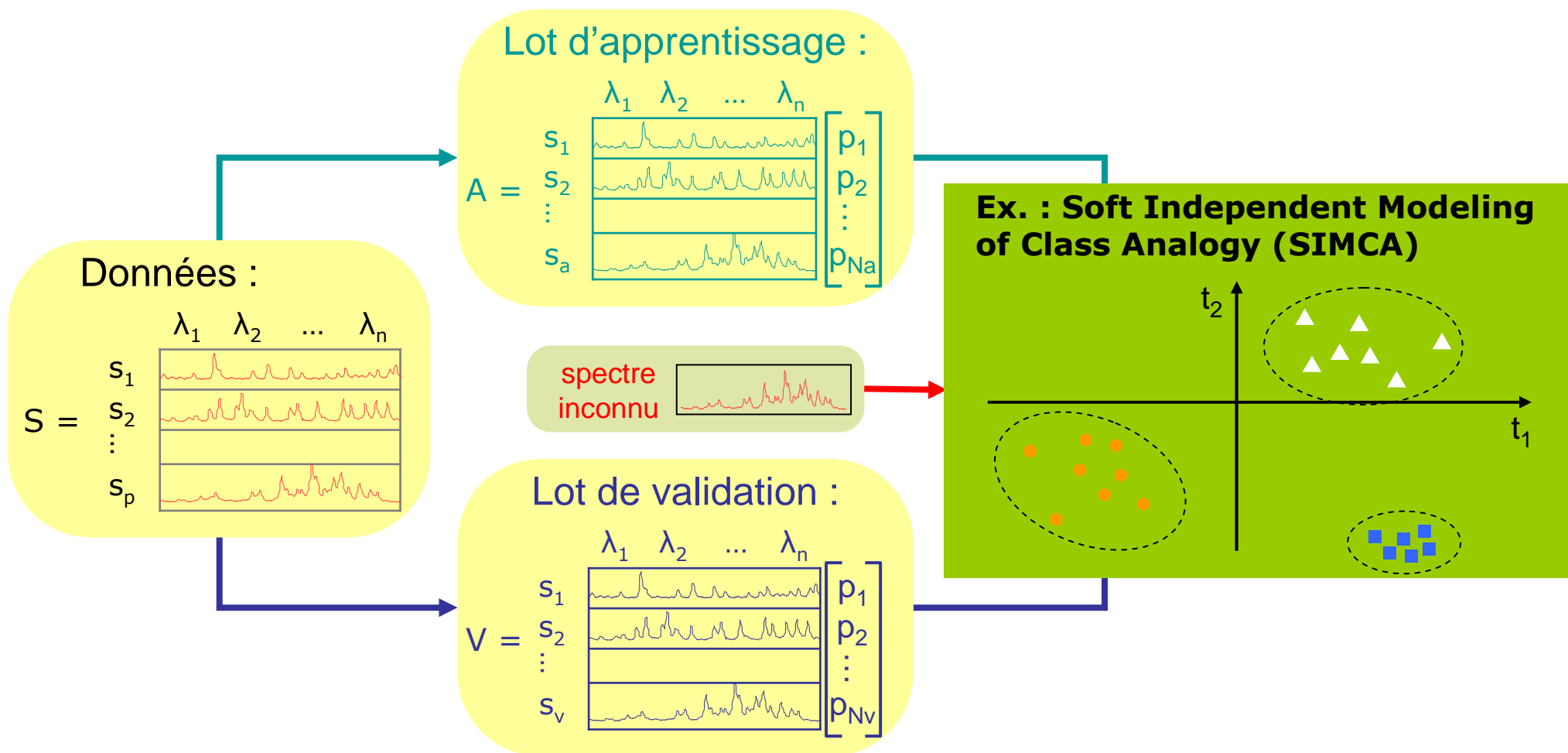
- On cherche à **prédire** une propriété de l'échantillon **p** (qualitative ou quantitative) à partir d'un **modèle**



# Méthodes supervisées ⇒ prédiction



- On cherche à **prédire** une propriété de l'échantillon **p** (qualitative ou quantitative) à partir d'un **modèle**



---

# 1. Quelques méthodes multivariées

---

# 1.1. Groupement (*clustering*)



✚ Analyse qualitative **non supervisée**

✚ Définir un **critère de distance** entre les observations

■ Ex. : distance euclidienne

$$d_{ij} = \sqrt{\sum_p (I_i(\lambda_p) - I_j(\lambda_p))^2}$$

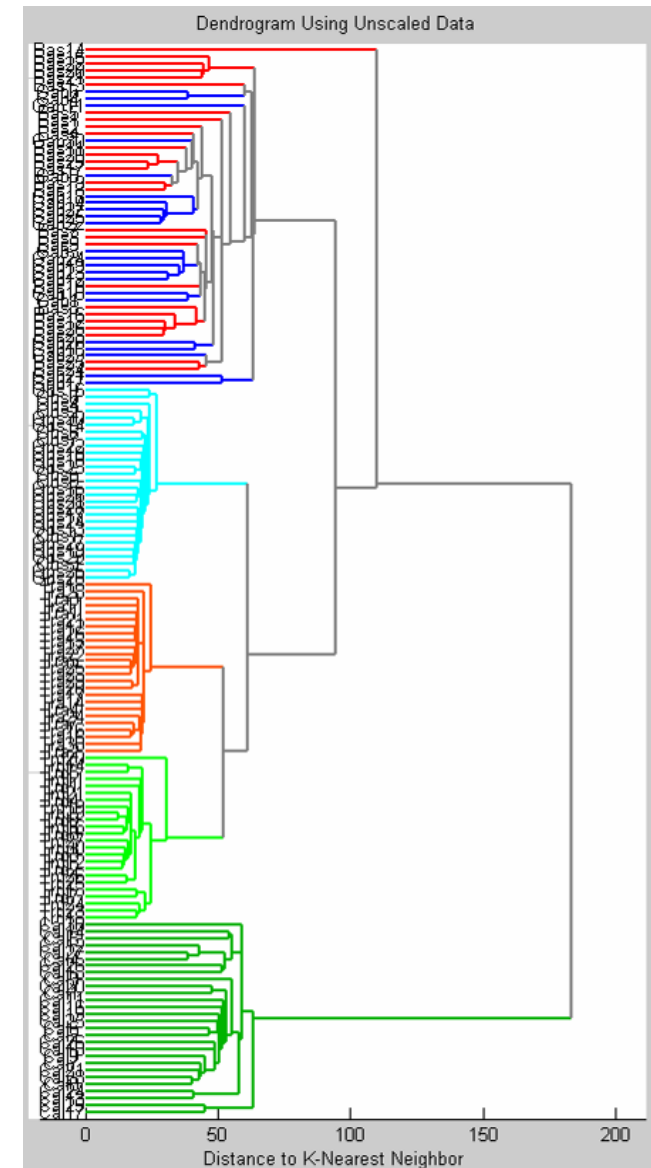
✚ Calculer la matrice des distances

✚ Le premier groupe est composé des 2 observations les plus proches

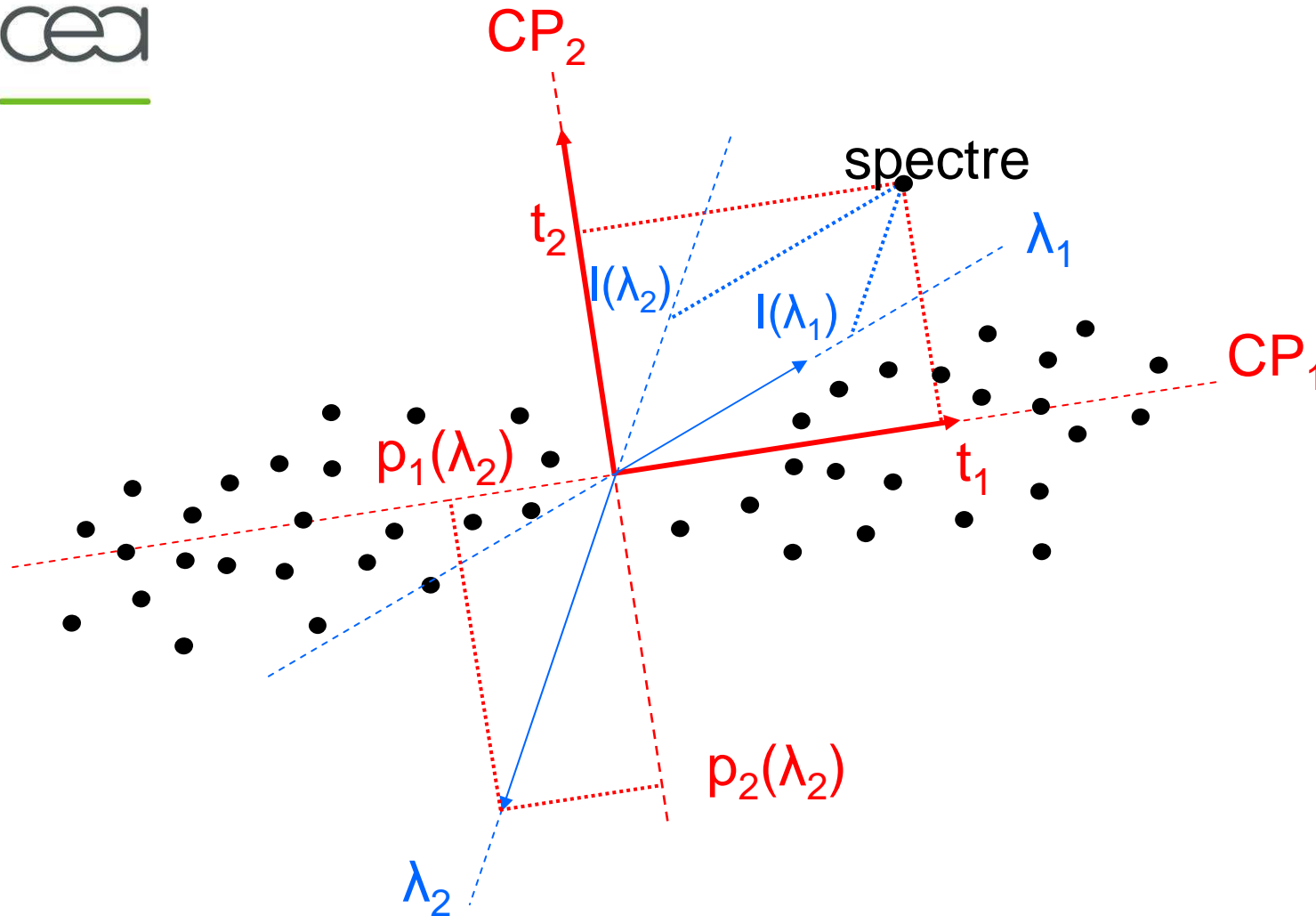
✚ Définir un **critère de proximité** entre les groupes

■ Ex. : plus proche voisin

✚ Recalculer la matrice des distances et itérer



# 1.2. Analyse en Composantes Principales (1)

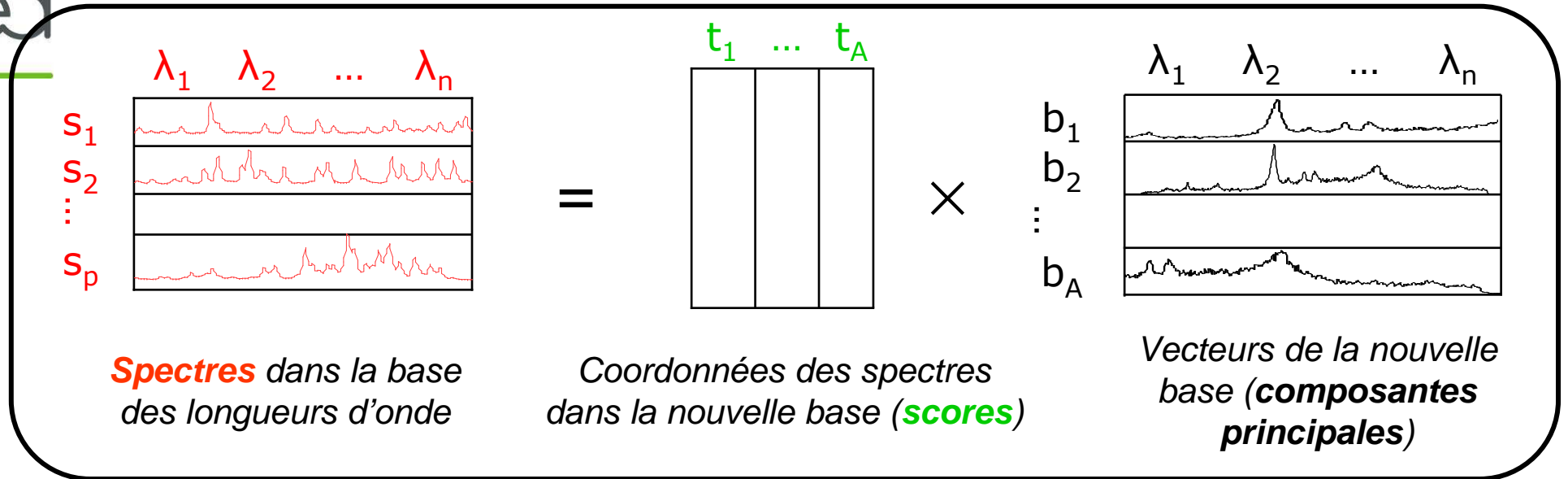


$$S = \sum_1^n I(\lambda_i) \cdot \vec{\lambda}_i$$

↓ ACP

$$S = \sum_1^A t_i \cdot \vec{CP}_i$$

# 1.2. Analyse en Composantes Principales (2)

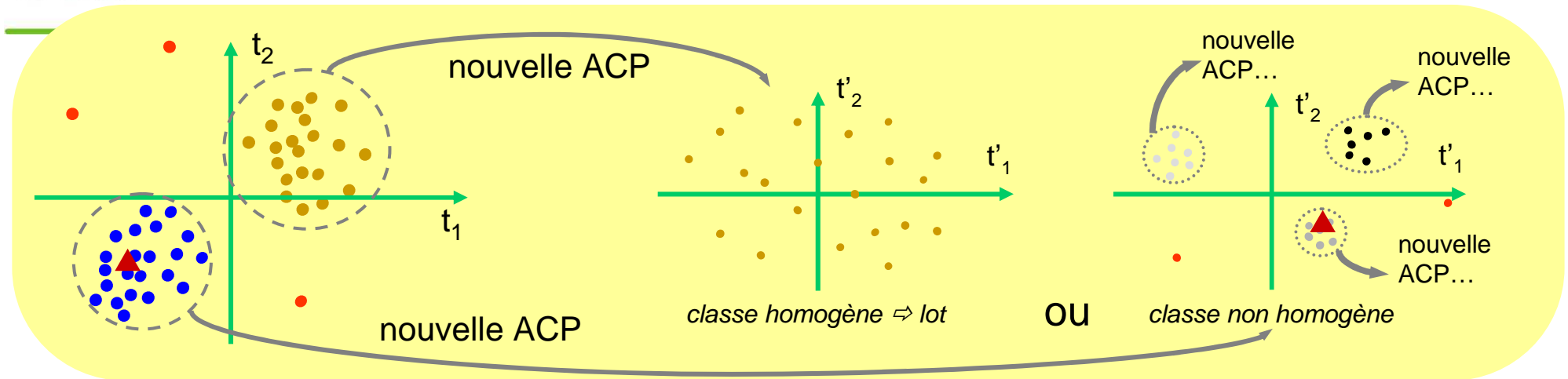


- Les **composantes principales** mettent en évidence les régions des spectres ayant la plus grande variance
- **Réduction de la dimensionnalité des données ( $A \ll n$ )** : les premières composantes principales contiennent la quasi-totalité de l'information contenue dans les données
  - ⇒ représentation graphique de l'**ensemble** des spectres
- **Interprétation physique** des composantes principales et des scores.

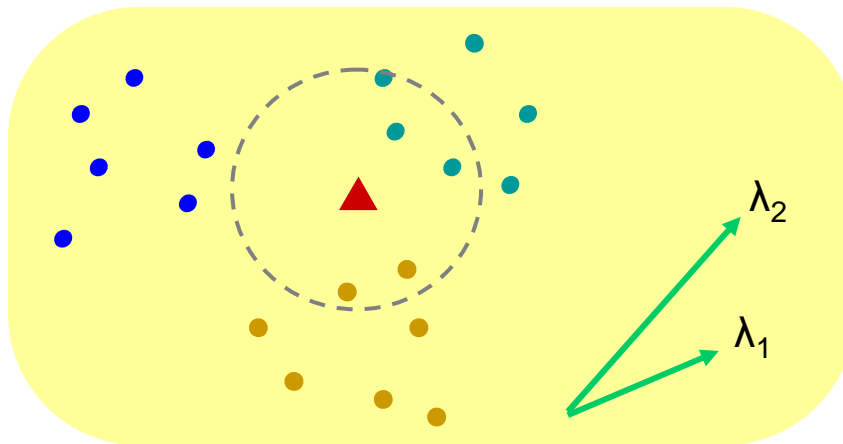
# 1.3. Méthodes non supervisées et prédiction



## Classification par ACP successives



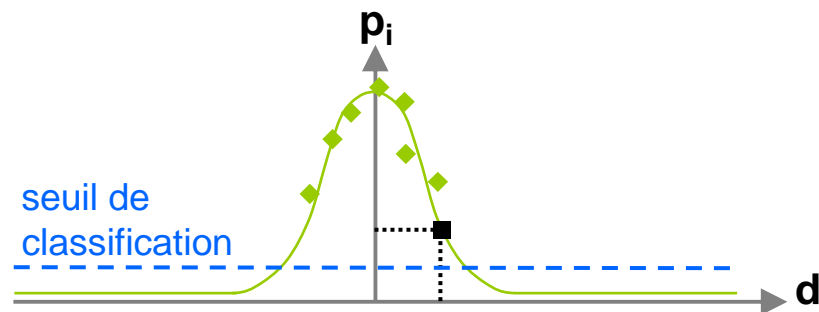
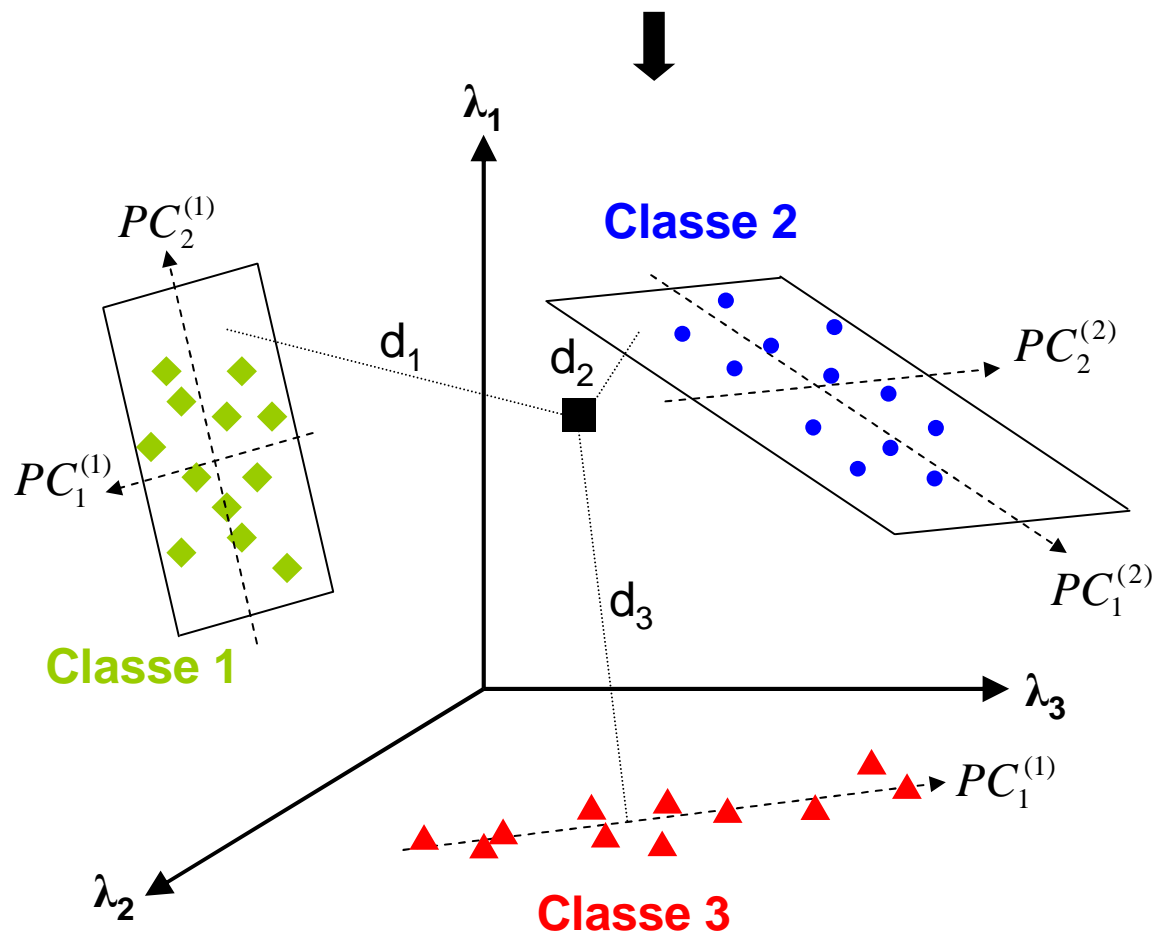
## K plus proches voisins (ex. avec K = 5)



## 1.4. Soft Independent Modeling of Class Analogy (SIMCA)



✓ **Apprentissage**  $\Rightarrow$  un modèle ACP est calculé pour chaque classe à partir du lot d'apprentissage.



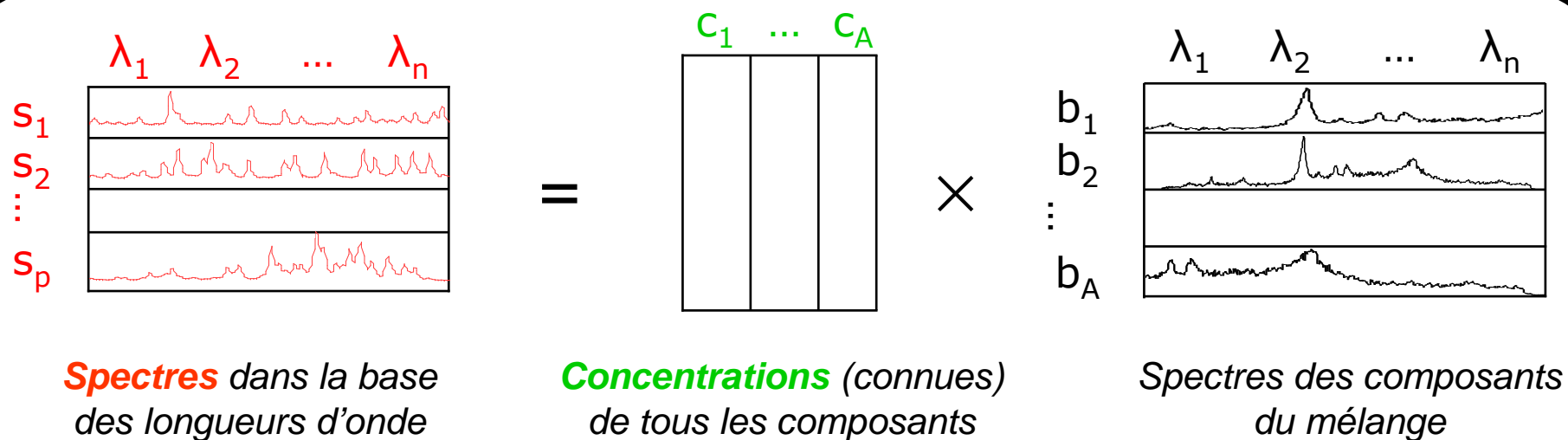
✓ **Prédiction**  $\Rightarrow$  on calcule la probabilité d'appartenance du spectre à chaque modèle  $i$  ( $p_i$ ). Le spectre est alloué à la classe pour laquelle  $p_i$  est maximum.

# 1.5. Régression multilinéaire (MLR)



Utile si :

- Tous les composants du mélange sont connus
- Leur concentration est parfaitement connue

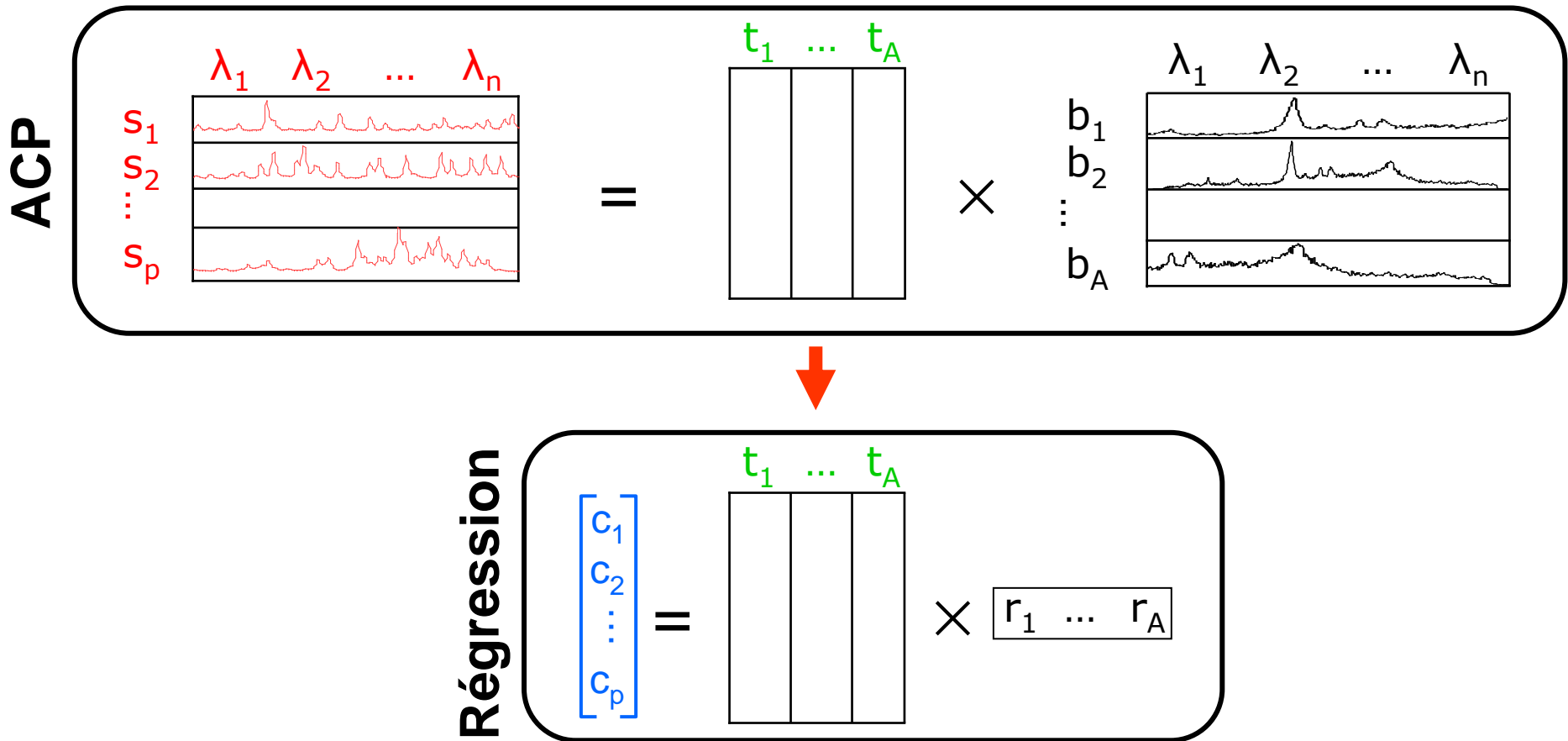


# 1.6. Régression sur composantes principales (PCR)



Utile si :

- Seuls quelques composants du mélange sont connus
- Leur concentration est parfaitement connue

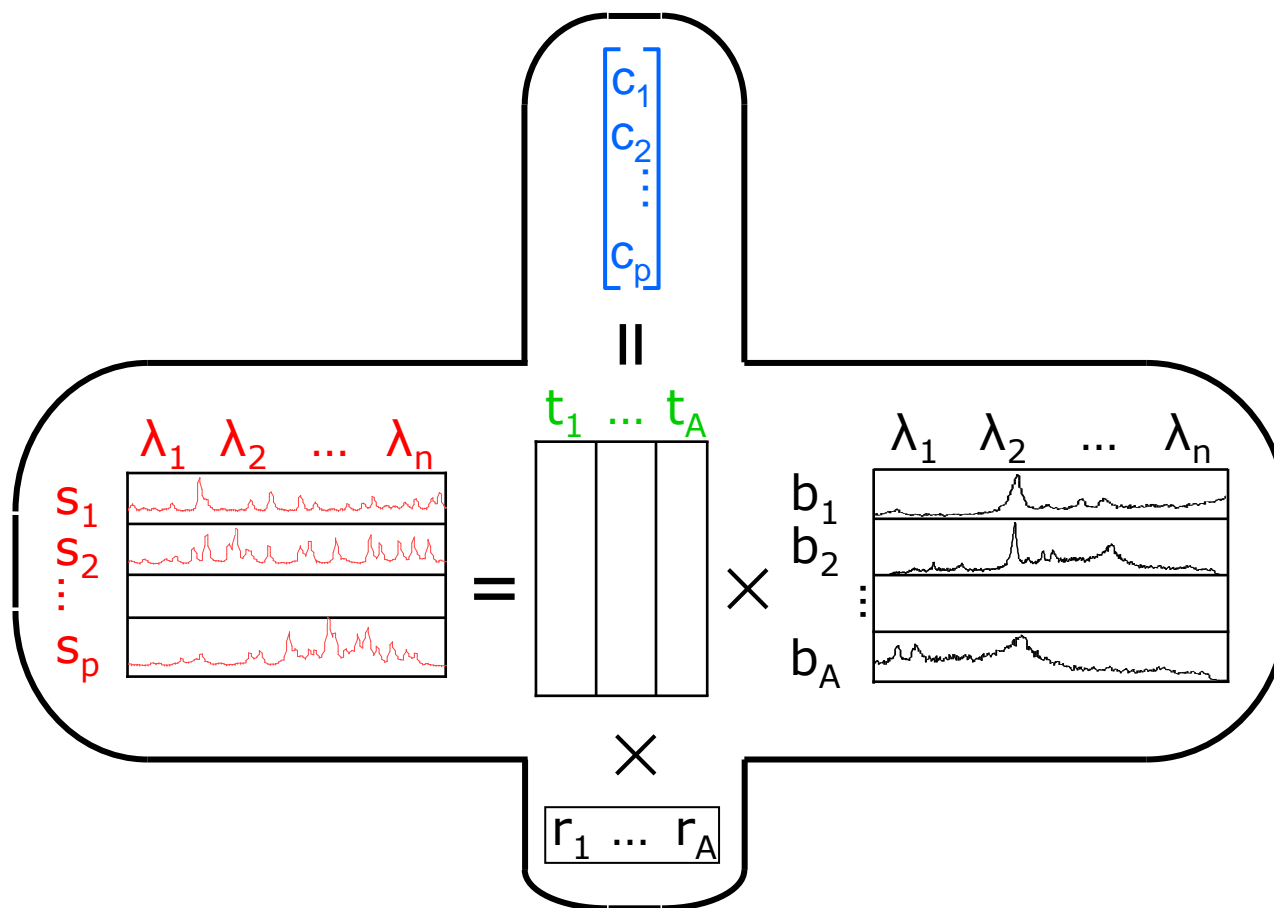


# 1.7. Régression PLS (Partial Least-Squares)



Utile si :

- Seuls quelques composants du mélange sont connus
- Leur concentration n'est pas parfaitement connue



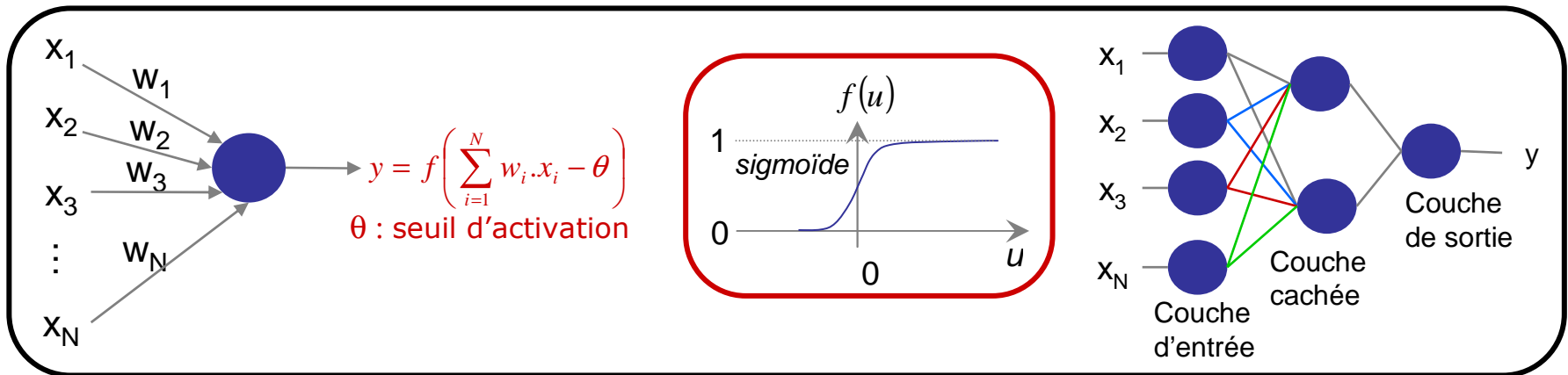
Les composantes de la régression PLS mettent en évidence les régions des spectres **corrélées à l'élément recherché** ( $\neq$  PCR).

Analyse discriminante par régression PLS (**PLS-DA**) : les concentrations sont remplacées par des variables binaires.

Les scores permettent de calculer une **probabilité d'appartenance** à chaque classe.

Toutes les classes sont modélisées **simultanément** ( $\neq$  SIMCA).

# 1.8. Réseaux de neurones



- ✚ Modélisation de **n'importe quelle relation** (linéaire ou non) entre les variables :
  - $x_i$  : intensités spectrales ou données dérivées des spectres
  - $y$  : concentration de l'analyte / variable muette (classe)
- ✚ **Étalonnage du réseau** : comparaison entre la valeur prédite et la valeur vraie de la concentration  $\Rightarrow$  calcul itératif des poids  $w_i = g(c_{pred} - c_{vraie})$ 
  - $\leftrightarrow$  **Fit** de la fonction entre les spectres et la concentration
- ✚ **Validation** : évaluation des performances prédictives du réseau **lorsque les poids sont fixés**
- ✚ Nombre de paramètres du réseau (3 couches / 1 sortie) :  $N_{param} = N_c(N_e + 1) + 1$

## **2. Analyse multivariée en LIBS**

---

# Spécificité des données LIBS



- ✚ Paramètres influençant le prétraitement des données, le choix et / ou le calcul du modèle :
  - **Fond spectral** (détecteur + Bremsstrahlung) → Soustraction éventuelle
  - **Bruit** → Filtrage, lissage
  - **Dynamique temporelle** → ?
  - **Pics définis par peu de pixels** → Sensibilité à une dérive en  $\lambda$
  - **Signaux faibles limités par le bruit de photons** → Fluctuations importantes
  - **Auto-absorption** → Non linéarité raie / concentration
  - **Redondance de l'information** (plusieurs raies par élément) → Variables colinéaires
  - **Données fortement multidimensionnelles** → Sélection des variables, compression des données
  - **Relative facilité à acquérir des données** → Constitution de plusieurs lots de spectres indépendants

# Choix de la méthode

---



- ✚ A-t-on des **connaissances préalables** sur les données ?
  - → méthode exploratoire / modélisation
  
- ✚ Cherche-t-on à évaluer une propriété **qualitative** ou **quantitative** ?
  - → méthode de classification / quantification
  
- ✚ Dispose-t-on de **données d'apprentissage** et de **données de test** ?
  - → méthode supervisée / non supervisée
  
- ✚ La relation entre les données et la propriété est-elle **linéaire** ?
  - → méthode (multi) linéaire / non linéaire
  
- ✚ **Prétraitement !**



## ✚ Prétraitements **LIBS** :

- Normalisation des données : par rapport à une raie, au fond, à l'intensité intégrée du spectre...
- Soustraction du fond
- Calcul de l'intensité des raies

## ✚ Prétraitements **numériques** :

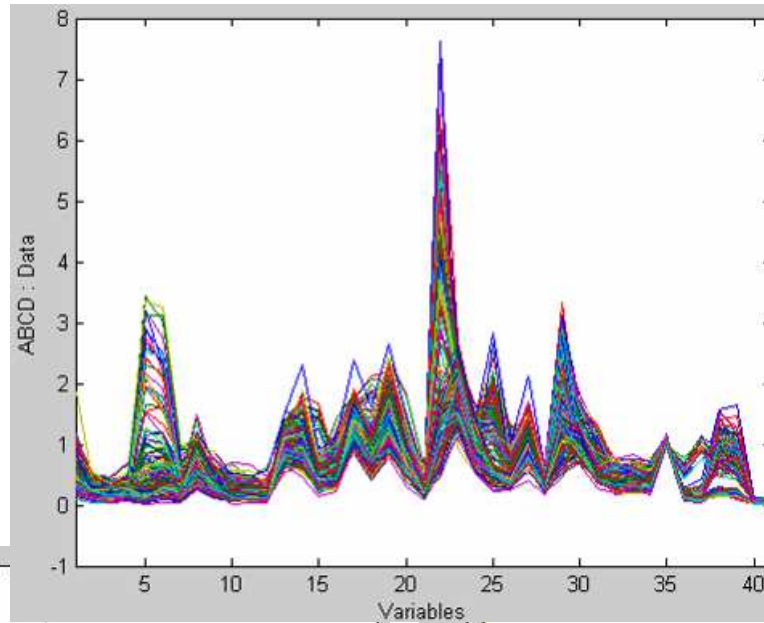
- Débruitage
- Déconvolution

## ✚ Prétraitements spécifiques au **modèle** :

- En spectroscopie on s'intéresse essentiellement aux variations autour de la moyenne → **données centrées**  $S_i(\lambda) - S(\lambda)_{MOY}$
- Si l'on souhaite donner la même importance à toutes les variables (intensité de raies par ex.) ou si les variables sont de dimension différente (spectroscopiques / instrumentales par ex.) → **données centrées-réduites**  $(S_i(\lambda) - S(\lambda)_{MOY})/S(\lambda)_{SIG}$
- Autres prétraitements : pondération des variables, transformation logarithmique, dérivation...

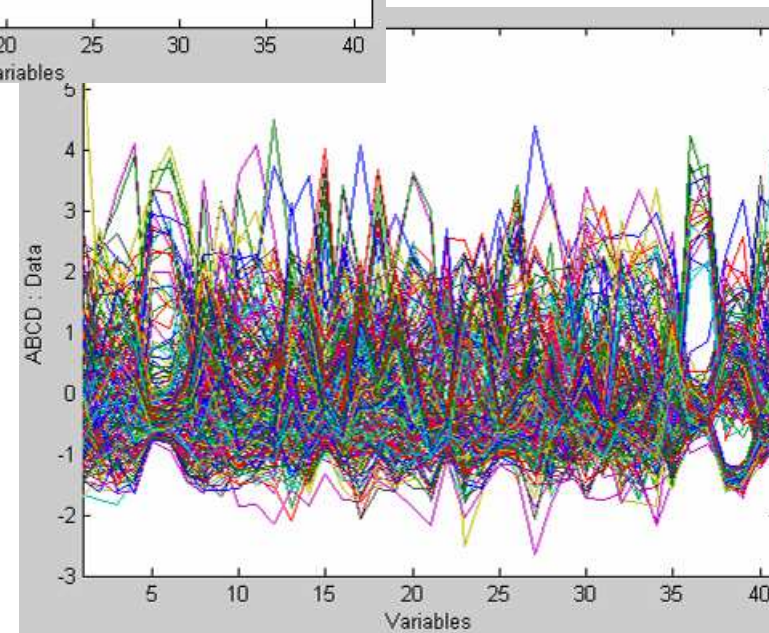
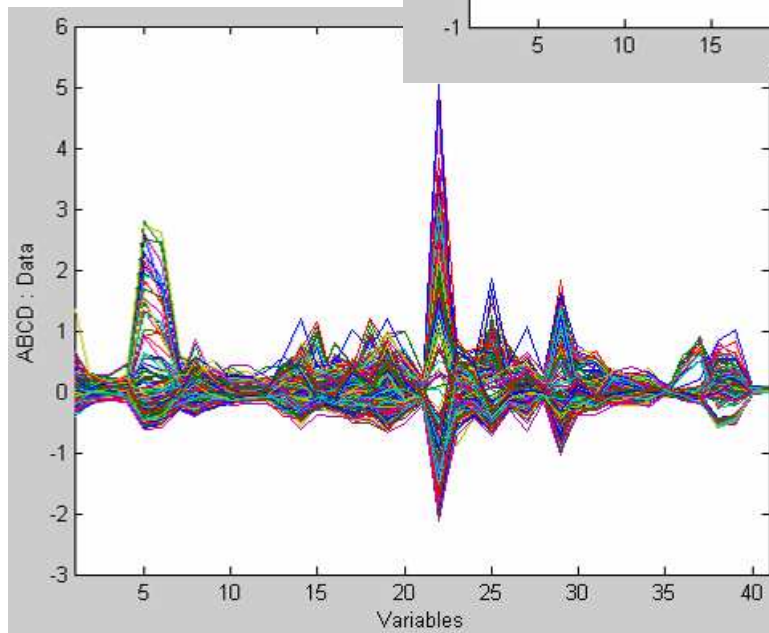


**Données  
brutes**



**Données  
centrées-  
réduites**

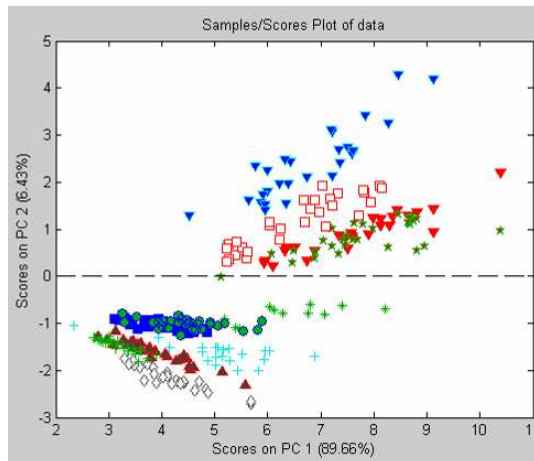
**Données  
centrées**



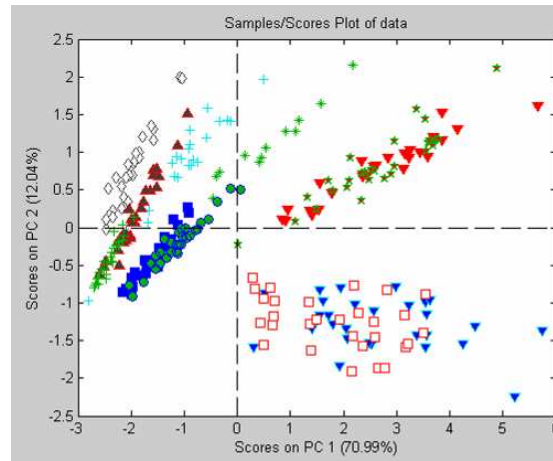


Les résultats d'un modèle dépendent fortement du prétraitement

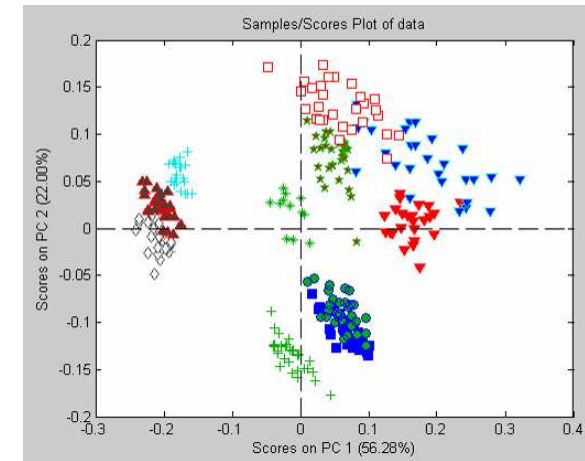
## Données brutes



## Données centrées

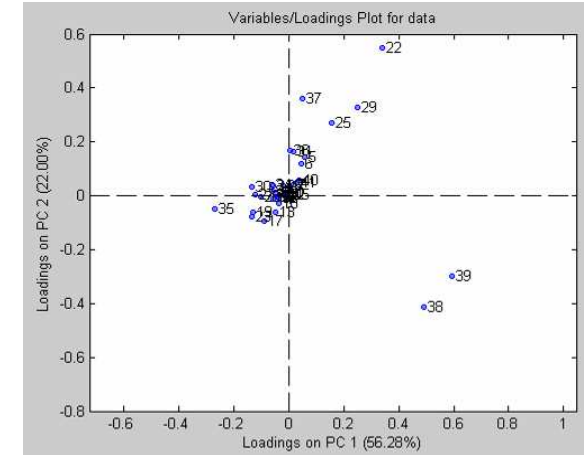
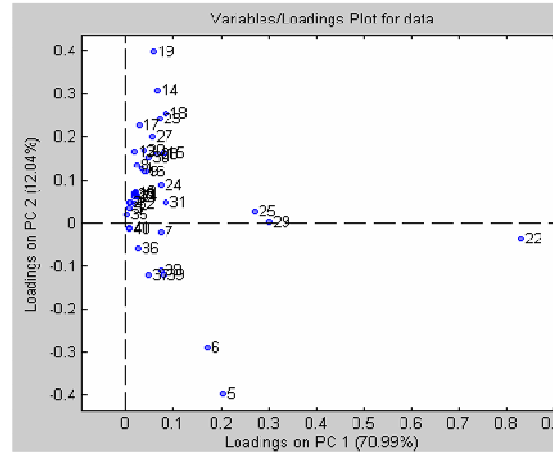
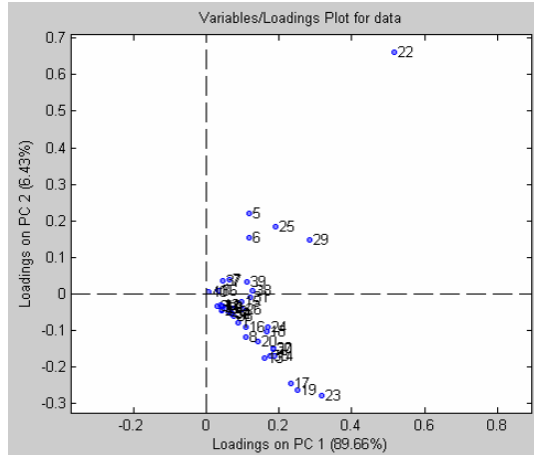


## Données normalisées/centrées



Scores

Loadings



# Calcul du modèle : précautions à prendre



## Avant :

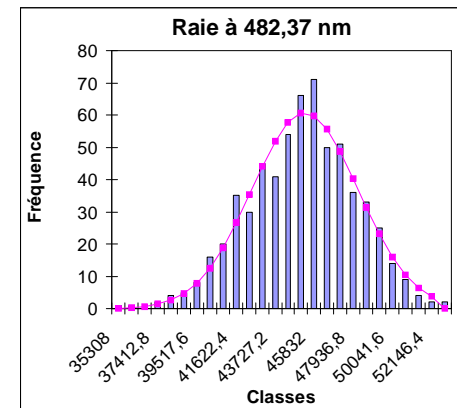
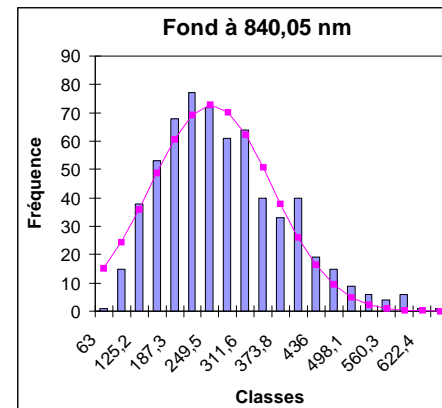
- **Normalité** des variables
- 3 lots de données **représentatifs et indépendants** : étalonnage, validation (→ optimisation du modèle), test (→ performances du modèle).  
*Typiquement 20-30 spectres par concentration pour l'étalonnage + 10-15 pour la validation + 10-15 pour le test (penser aux outliers !)*
- $N_{\text{observations}} / N_{\text{paramètres}}$  « grand » ...

## Pendant :

- Absence d'**outliers**
- **Surapprentissage**

## Après :

- Test du modèle sur **données indépendantes**
- Évaluation de la confiance dans le modèle

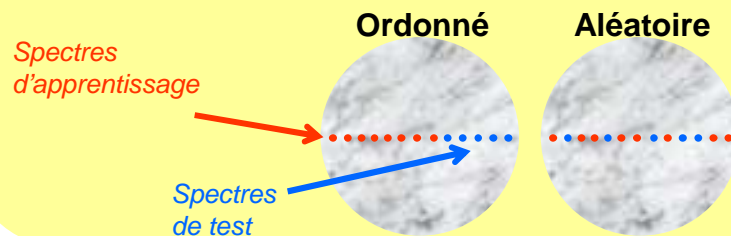


# Qu'est-ce qu'un lot de données représentatif ?



- ✚ Un ensemble de spectres représentatifs des différentes **sources de variance** attendues, par exemple :
- ✚ Variance associée à **l'échantillon** :
  - Mesures répliquées sur un même échantillon
  - Mesures sur plusieurs échantillons de la même famille
  - Mesures sur des échantillons de concentration variable

## Echantillonnage de roche



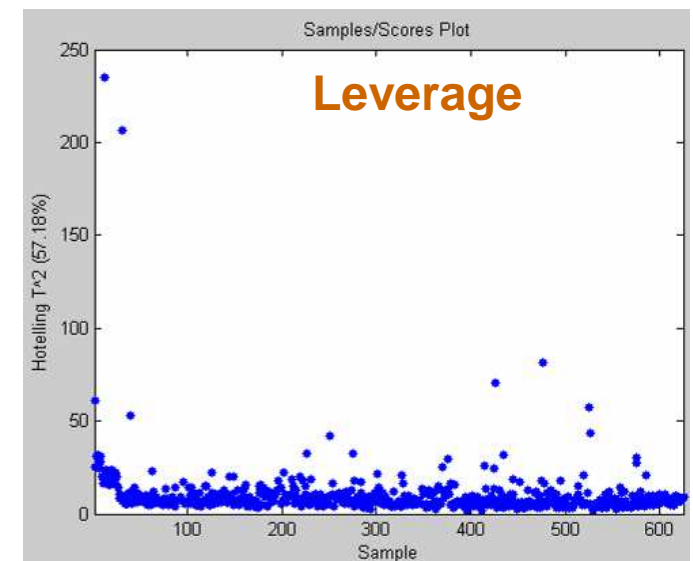
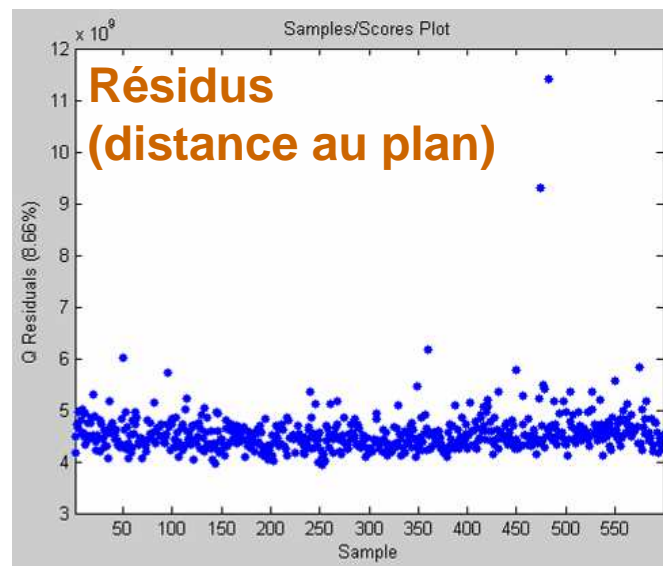
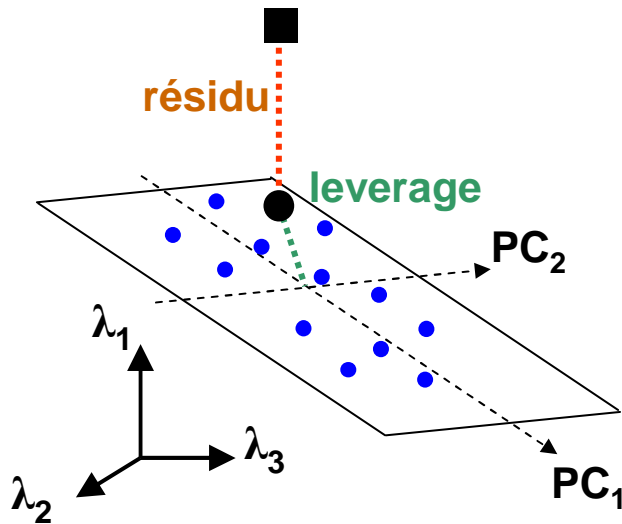
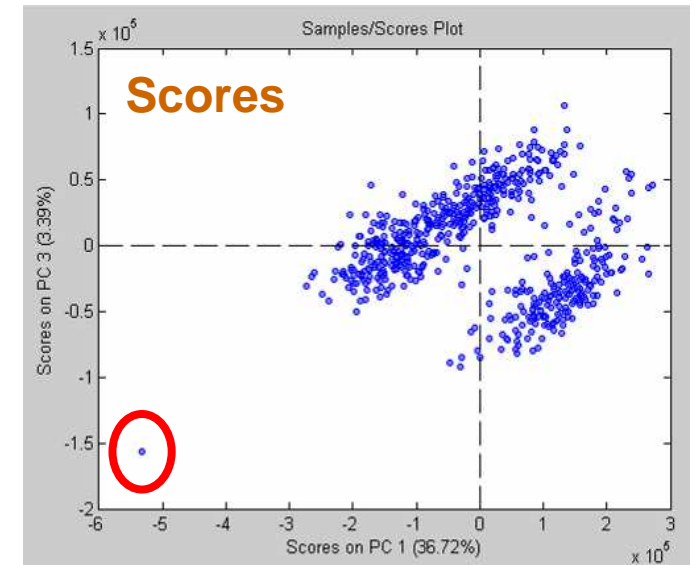
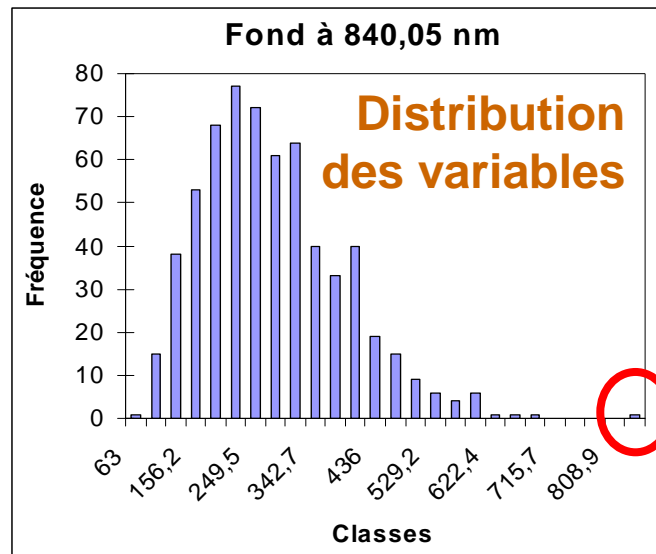
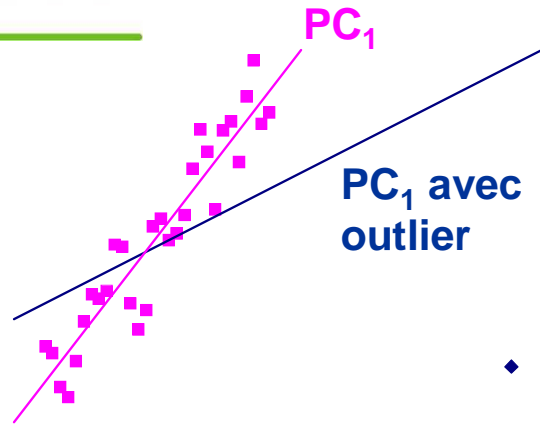
## Analyse par SIMCA

Taux moyen de...	classification correcte	"non classification"	fausse classification
Ordonné	0.517	0.458	0.025
Aléatoire	0.775	0.225	0

- ✚ Variance associée à la **mesure** :
  - Mesures par plusieurs opérateurs
  - Mesures à différentes périodes
- ✚ En pratique on a rarement (jamais...) de lot représentatif

**→ les modèles sont souvent optimistes !**

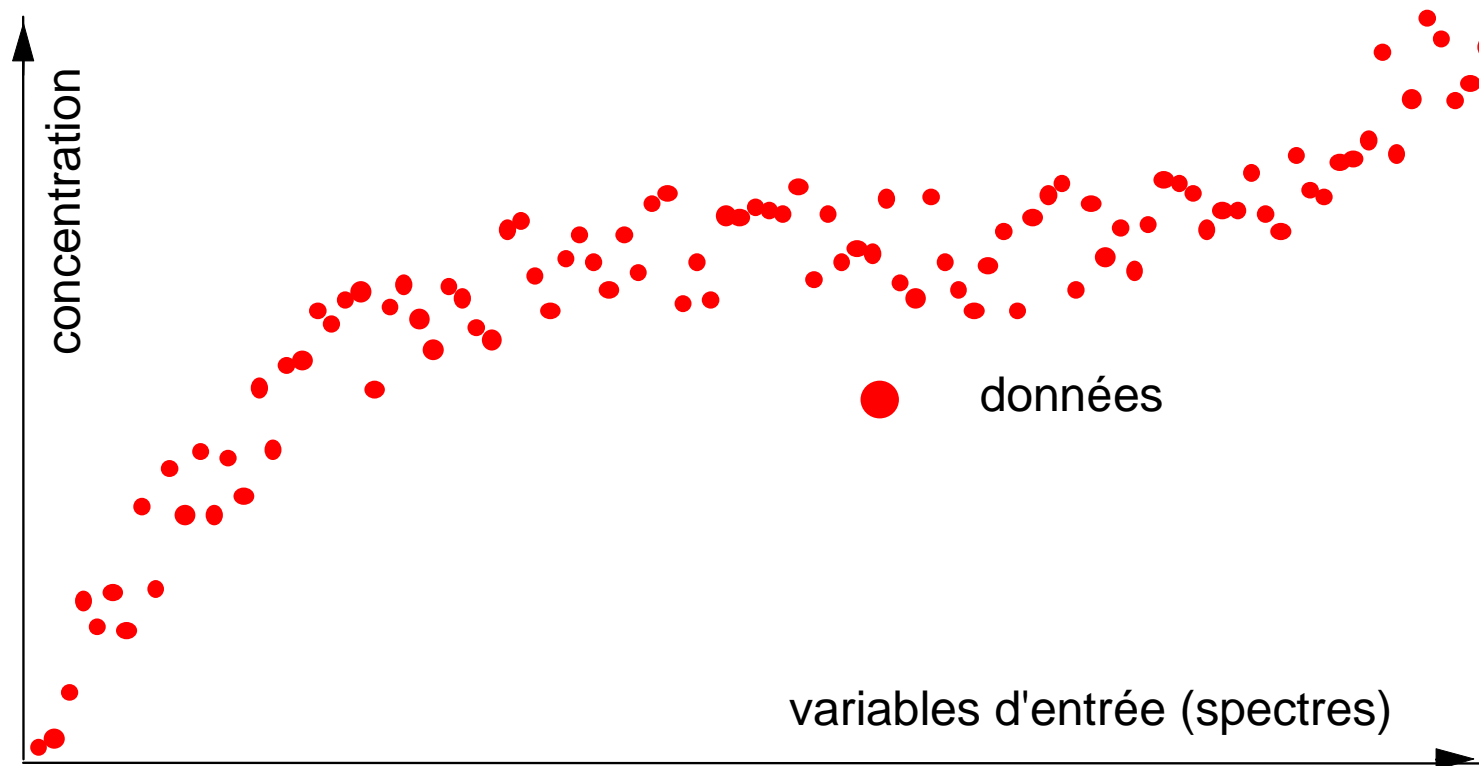
# Détection des outliers : diagnostics



# Surapprentissage (« overfitting »)



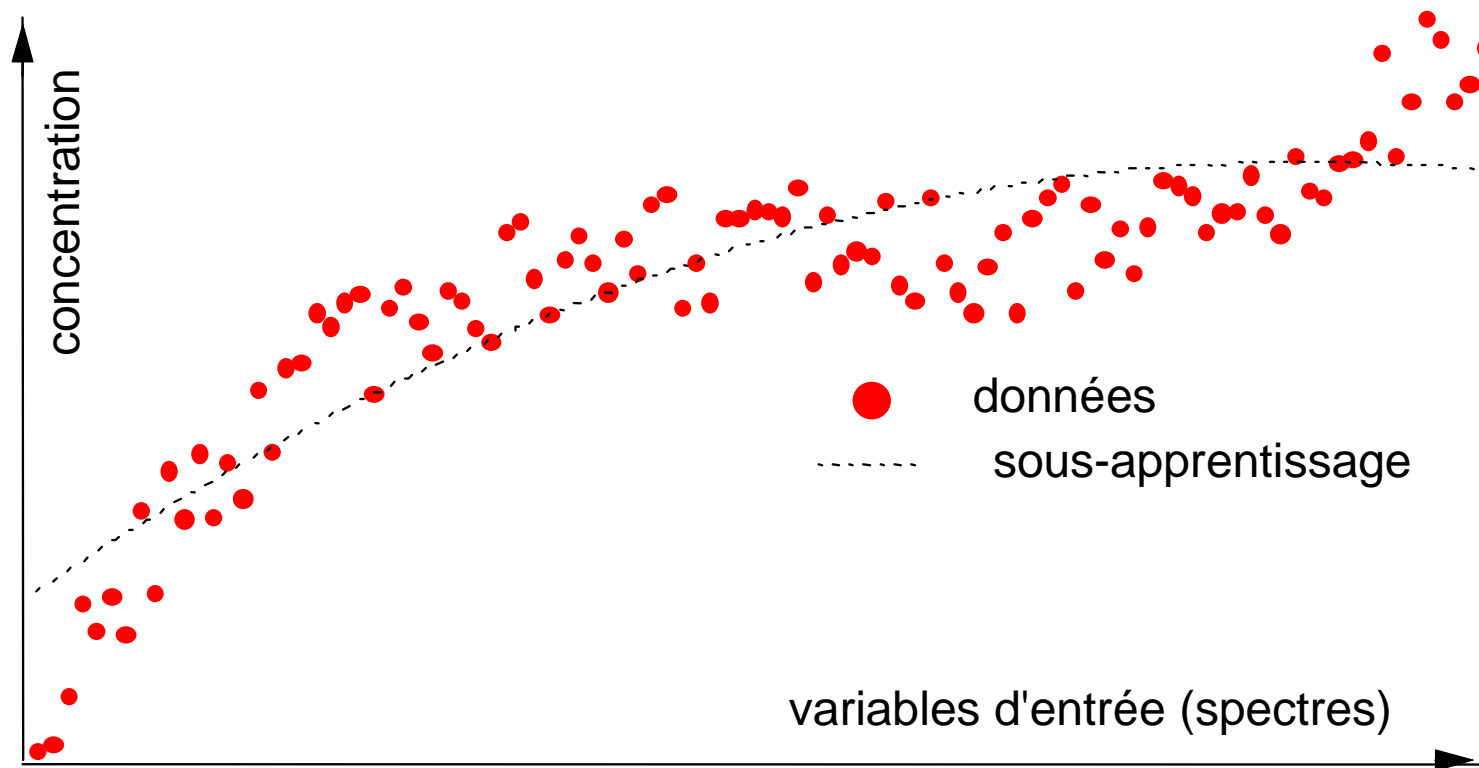
- Optimisation du nombre de paramètres pour limiter le surapprentissage (le modèle ne doit pas être trop complexe)



# Surapprentissage (« overfitting »)



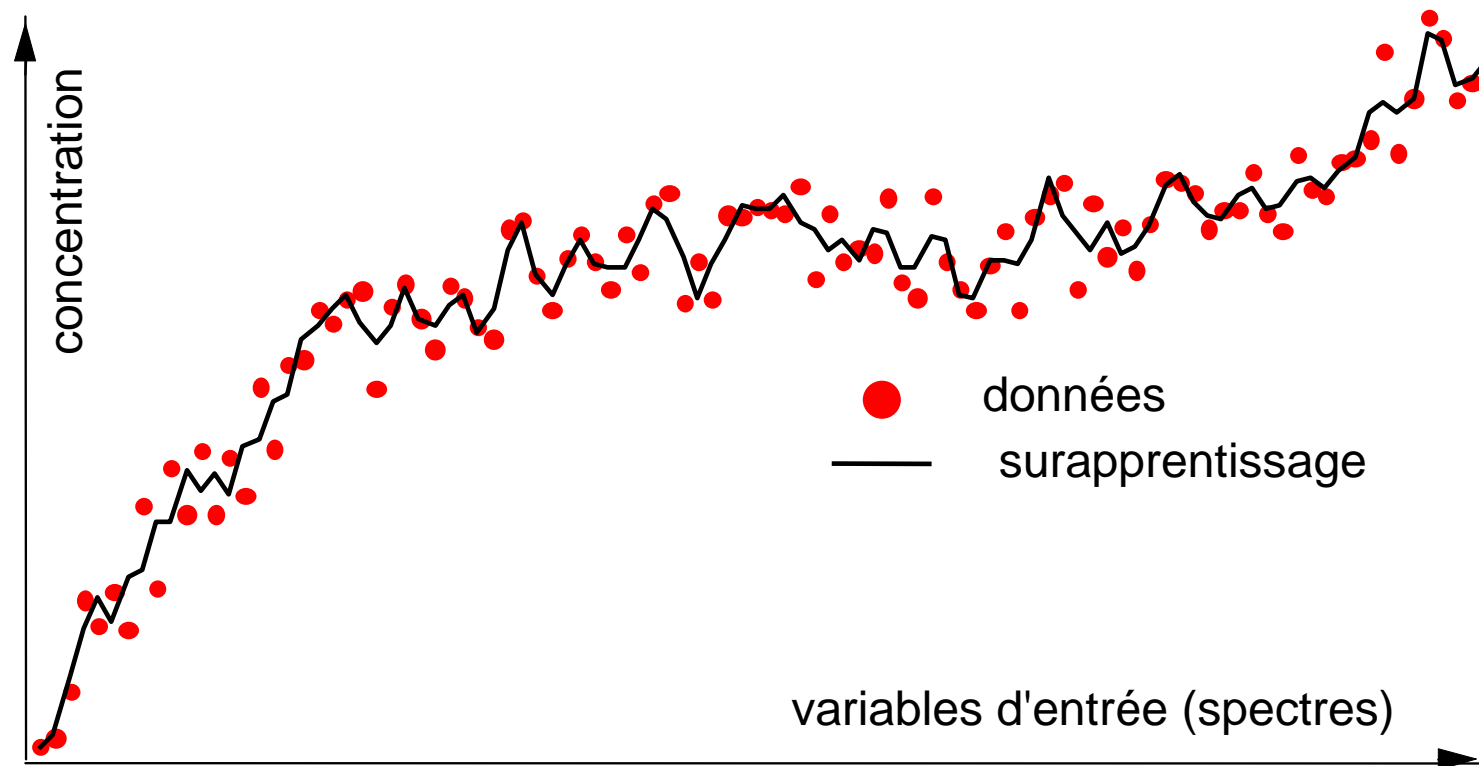
- Optimisation du nombre de paramètres pour limiter le surapprentissage (le modèle ne doit pas être trop complexe)



# Surapprentissage (« overfitting »)



- Optimisation du nombre de paramètres pour limiter le surapprentissage (le modèle ne doit pas être trop complexe)

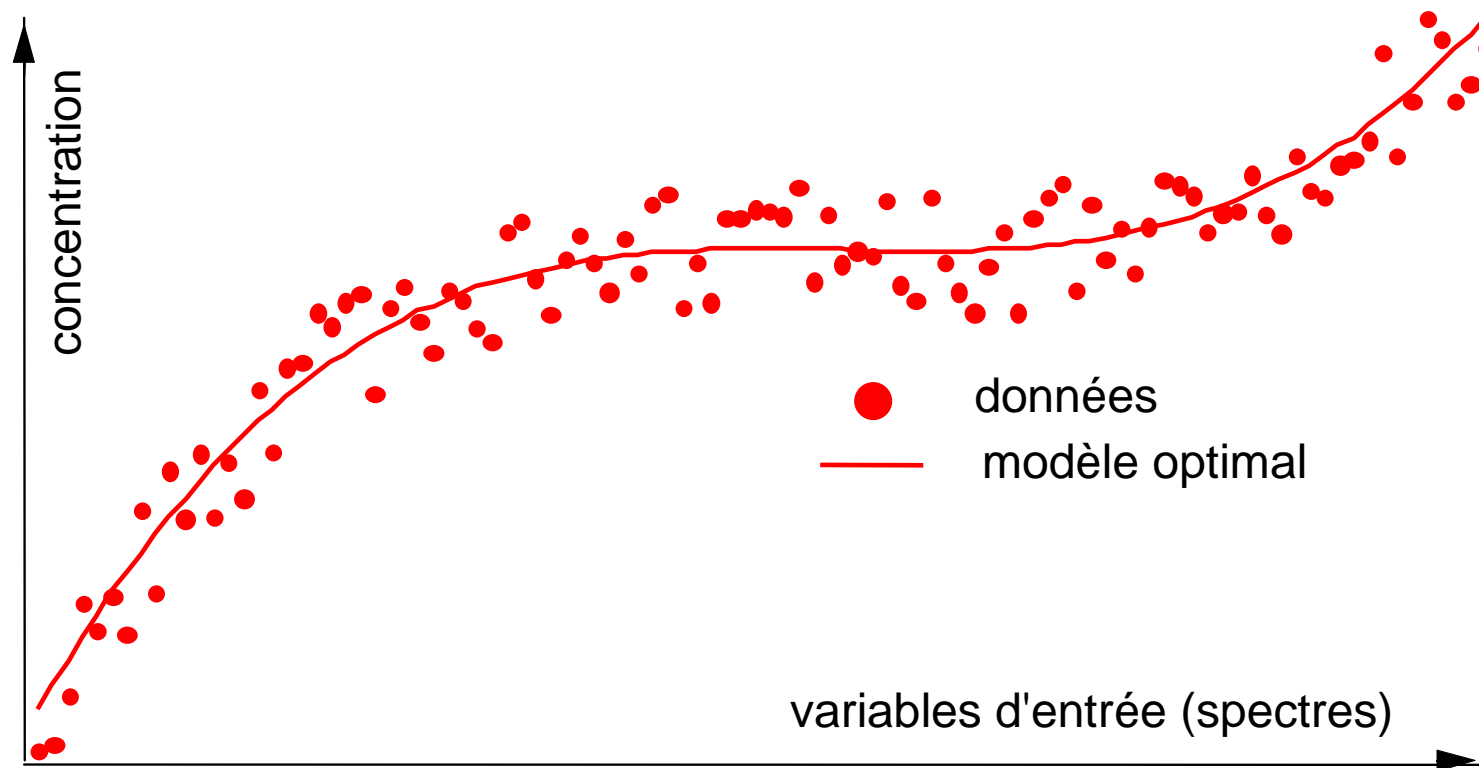


- Le modèle  $c = f(\text{spectres})$  ajuste les données d'étalonnage mais présente une forte variance en prédiction

# Surapprentissage (« overfitting »)



- Optimisation du nombre de paramètres pour limiter le surapprentissage (le modèle ne doit pas être trop complexe)



# Comment limiter le surapprentissage ?

---



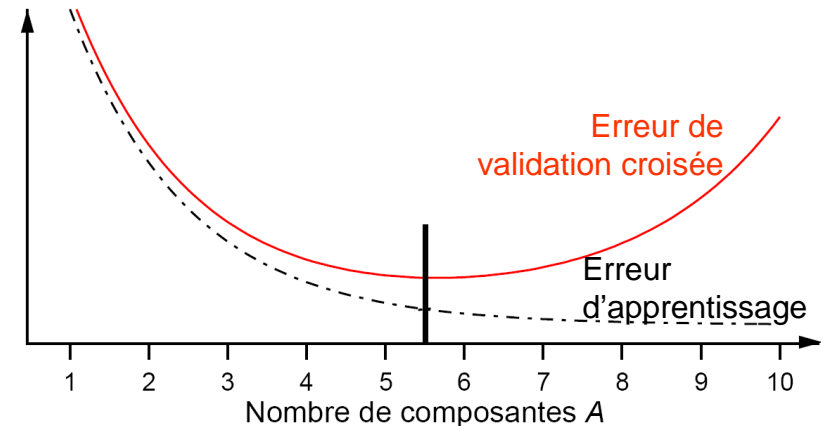
- ✚ Par une procédure de **validation** du modèle
- ✚ En maximisant le rapport  $N_{\text{observations}} / N_{\text{paramètres}}$ . 2 stratégies :
- ✚ **Sélectionner** les longueurs d'onde :
  - Réduction de la bande spectrale
  - Sélection de raies
  - Algorithme génétique
  - *Simulated annealing*
  - Information mutuelle
  - Autres critères...
- ✚ **Compresser** les données :
  - Utilisation des scores d'une ACP en entrée du réseau
  - Autres techniques de compression : régression PLS, Projection en Coordonnées Polaires...
- ✚ Autre avantage : interprétation du modèle plus simple, temps de calcul réduit

# Confiance dans le modèle



Les prédictions du lot d'étalonnage doivent être bonnes...

- Méthodes de validation :
- Optimisation du modèle sur **lot de validation** indépendant
  - **Validation croisée**



Critère : les performances du modèle pour le lot de validation doivent être proches de celles obtenues avec le lot d'étalonnage

Une fois le modèle validé, caractérisation des incertitudes de prédiction avec **lot de test** indépendant

Incertitude + intervalle de confiance si possible

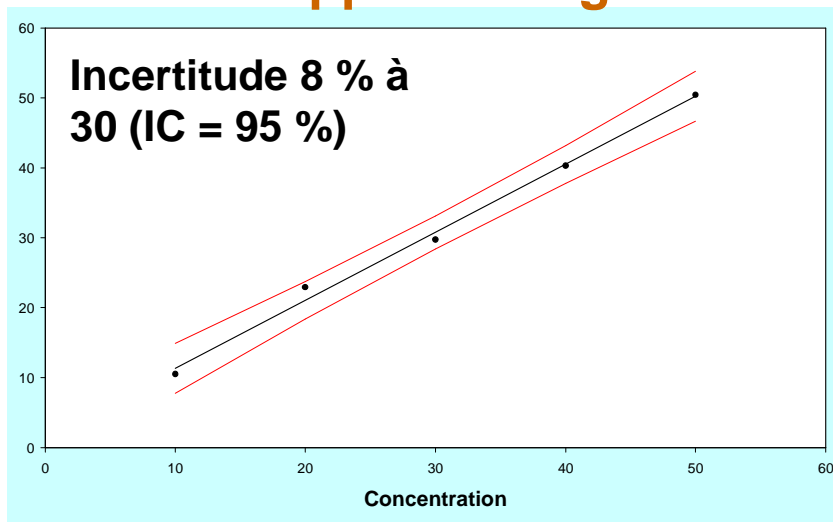
**Prédiction : attention aux outliers !**

# Concentration prédite VS concentration mesurée (PLS)

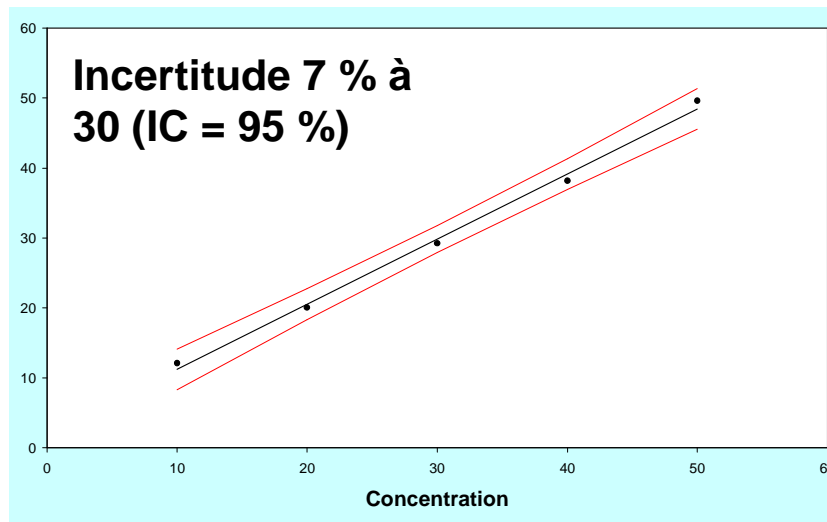


2 composantes

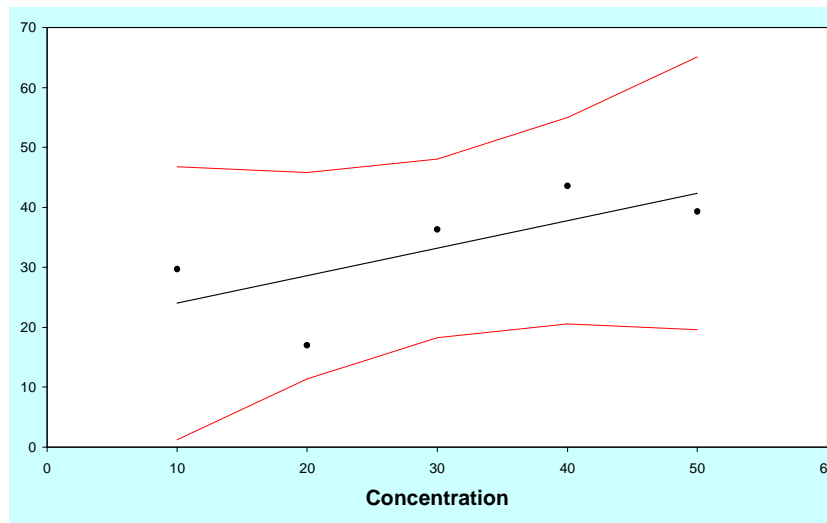
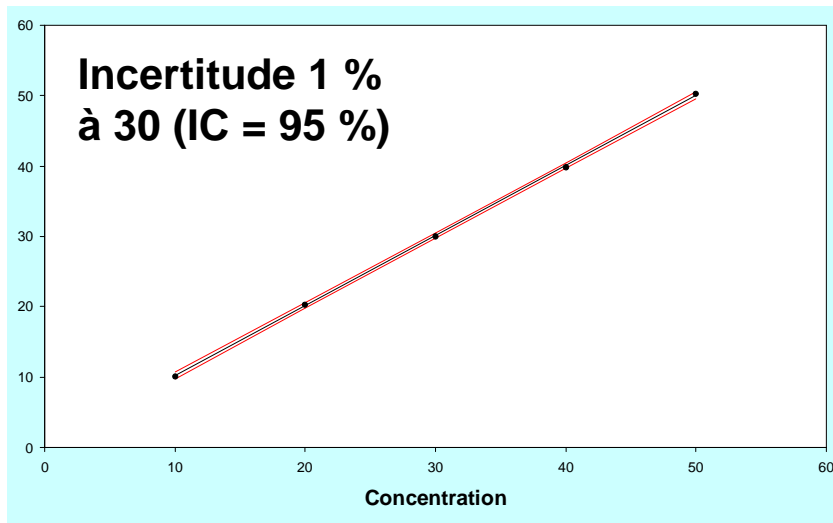
## Lot d'apprentissage



## Lot de test



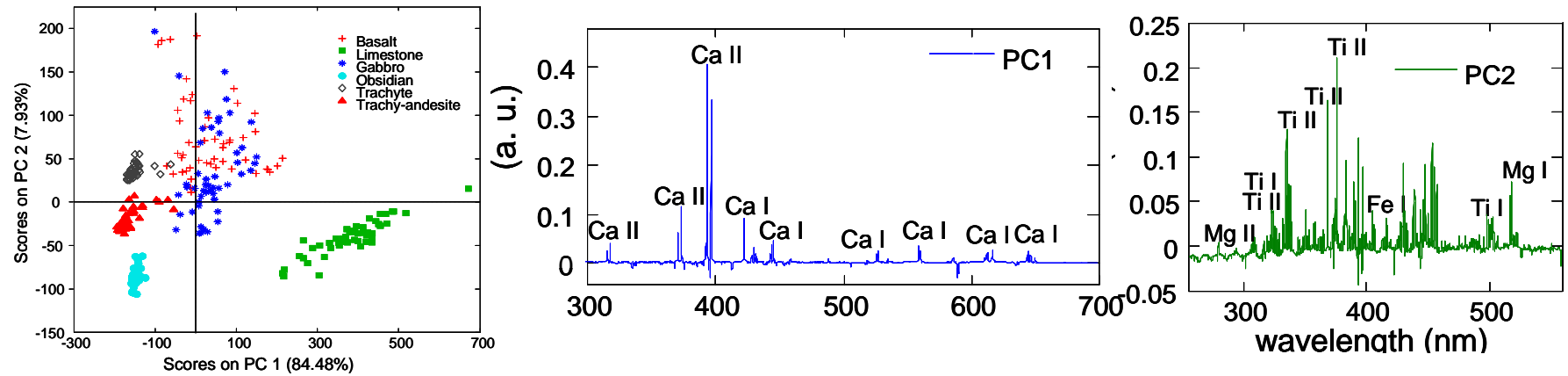
20 composantes



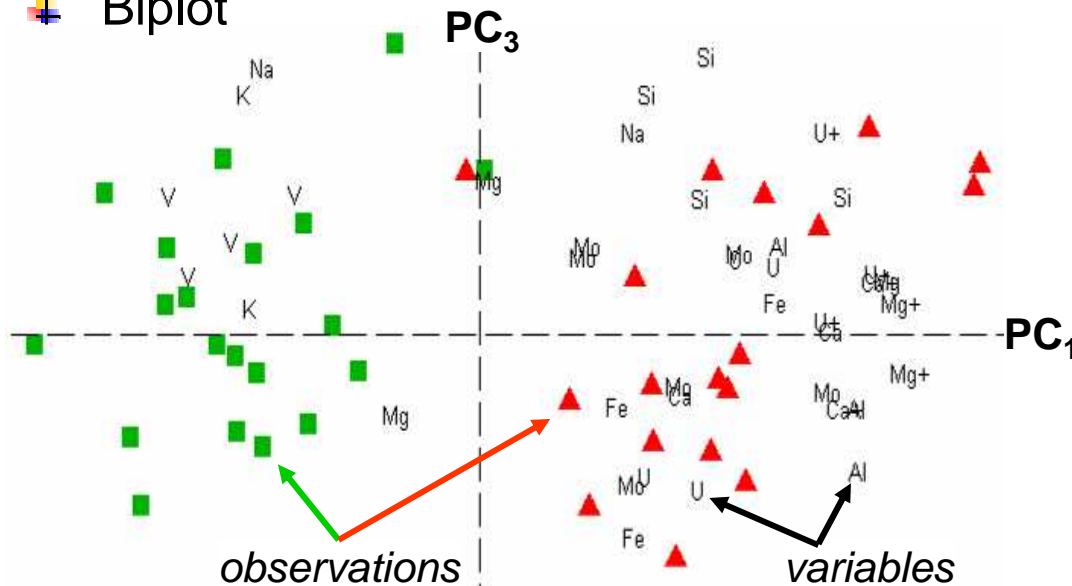
# Interprétabilité du modèle



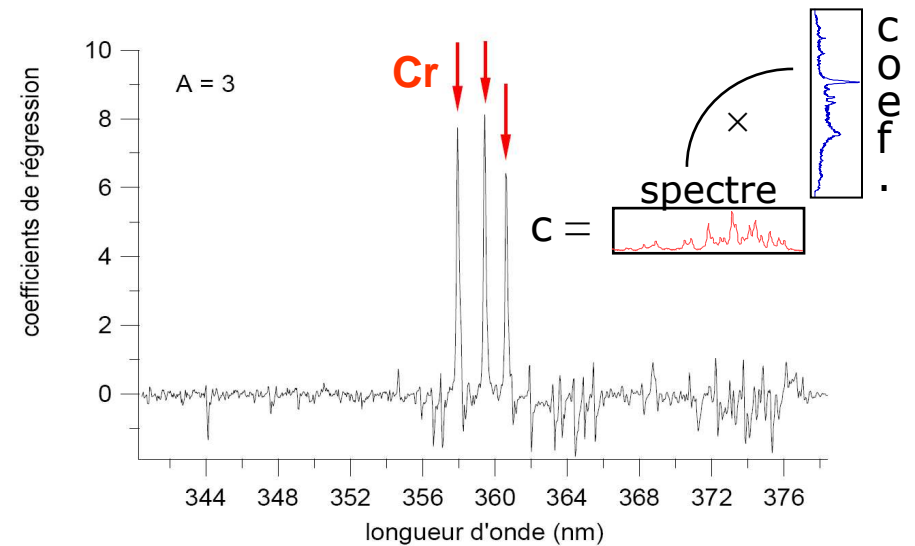
## Représentation conjointe scores / composantes



## Biplot



## Vecteur de coefficients



# Confrontation de différentes méthodes



- ✚ Classification de roches par LIBS (6 classes)
  - Cas 1 : les 6 classes sont incluses dans le lot d'apprentissage
  - Cas 2 : chaque roche est tour à tour enlevée du lot d'apprentissage → test de robustesse en cas d'analyse d'une classe inconnue
  
- ✚ Test de 2 méthodes : SIMCA et PLS-DA
  
- ✚ Prédications en accord pour 69 % des spectres
  
- ✚ Cas 2 : 88 % de ces spectres sont classés. **Bilan : 4 spectres sur 10 de « perdus » mais 100 % de réussite pour tous les autres.**

		SIMCA	PLS-DA	SIMCA $\cap$ PLS-DA
Cas 1	Taux de classification correcte (%)	77,5	85,9	97,6
	Taux de non-classification (%)	22,5	5,8	2,4
	Taux de fausse classification (%)	0,0	8,3	0,0
	<i>Parmi les spectres classés, taux de classification correcte (%)</i>	100,0	90,6	100,0
Cas 2	Taux d'allocation correcte (%)	81,4	83,1	97,3
	Taux de non-allocation (%)	18,6	4,3	2,7
	Taux de fausse classification (%)	0,0	12,6	0,0
	<i>Parmi les spectres classés, taux de classification correcte (%)</i>	100,0	86,3	100,0



- ✚ Les **méthodes multivariées** sont utiles quand :
  - Les performances des méthodes univariées ne sont pas suffisantes
  - La quantité de données à traiter est importante (atout de la LIBS !)
  
- ✚ La chimiométrie **n'est pas** une « boîte noire »
  
- ✚ **Interprétabilité** des résultats
  
- ✚ **Prétraitement** au moins aussi important que le modèle lui-même
  
- ✚ **Principales erreurs** à éviter :
  - Lot d'apprentissage non représentatif
  - Outliers → modèle biaisé
  - Surapprentissage → modèle peu prédictif
  
- ✚ Intérêt de **confronter** les résultats de différents modèles