

INRA

Institut National de la Recherche Agronomique



Sélection d'échantillons représentatifs

Application à des modèles d'étalonnage

Delphine Jouan-Rimbaud Bouveresse

UMR INRA/INA P-G 214 "IAQA"
16, rue Claude Bernard
75231 Paris cedex 05

L' utilisation de méthodes spectroscopiques en chimie analytique est de plus en plus utilisée en routine (analyse qualitative et quantitative) :

- rapidité
- précision
- coûts moins élevés qu' une analyse classique
- pas (ou peu) de préparation d' échantillons

=>

- moins d' erreur expérimentale
- pas d' utilisation de solvants toxiques

Les échantillons analysés dans les applications industrielles sont souvent des mélanges complexes d' un grand nombre de composants (réactifs, produits intermédiaires, produits, impuretés):

- Pas de longueur d' onde sélective
- Impossibilité de préparer des solutions étalons.

(L'étalonnage est donc basé sur un jeu d'échantillons réels.

Etalonnage

Matrice spectrale \mathbf{X} ($n \times p$)

Vecteur des concentrations \mathbf{y} ($n \times 1$)

Le modèle linéaire cherche à relier \mathbf{X} et \mathbf{y} selon l' équation

$$\mathbf{y} = \mathbf{X} \mathbf{b}$$

(\mathbf{b} : vecteur des coefficients de régression)

Régression linéaire multiple (MLR)

Régression sur les composantes principales (PCR)

Régression "partial least squares" (PLS)

etc.

Pendant sa construction, le modèle doit être **validé**

- ◇ Détermination du nombre de variables latentes (PCR, PLS)
- ◇ Estimation de l' erreur de prédiction associée au modèle

Dans le cas idéal, on utilise deux jeux de données:

- $\{\mathbf{X}_e, \mathbf{y}_e\}$ pour l'étalonnage
- $\{\mathbf{X}_v, \mathbf{y}_v\}$ pour la validation

Il faut donc répartir avec soin les différents échantillons (étalonnage / validation) dans ces deux groupes :

- si les deux groupes sont trop similaires \Rightarrow La validation est optimiste, l' erreur de prédiction est sous-estimée;
- si les deux groupes sont trop différents \Rightarrow La validation n' est pas correcte, l' erreur de prédiction est surestimée.

Choix des échantillons du jeu d'étalonnage

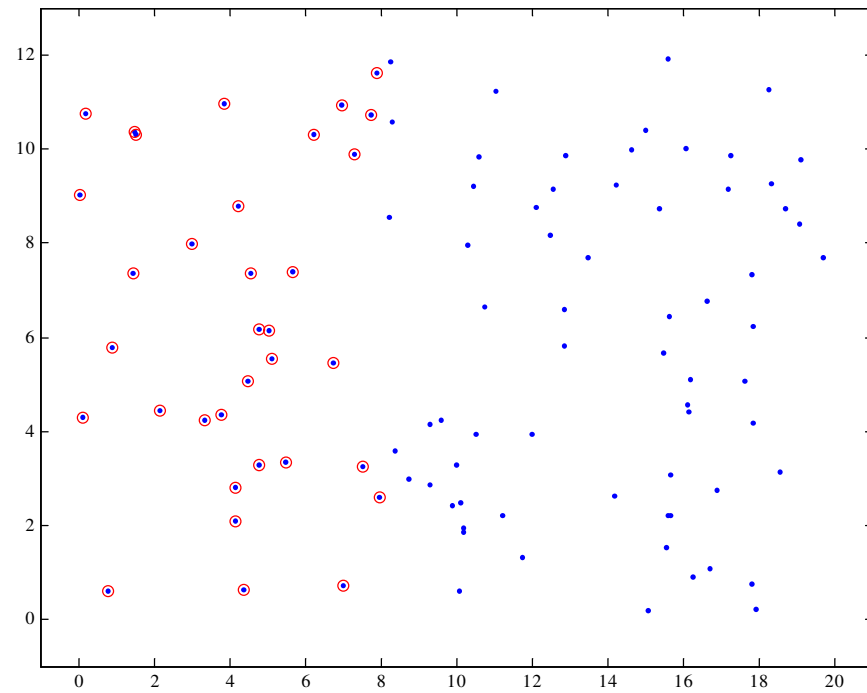
Le choix des échantillons est une étape TRES IMPORTANTE d' une procédure d' étalonnage multivarié.

Les échantillons d' étalonnage doivent donc répondre à certains critères.

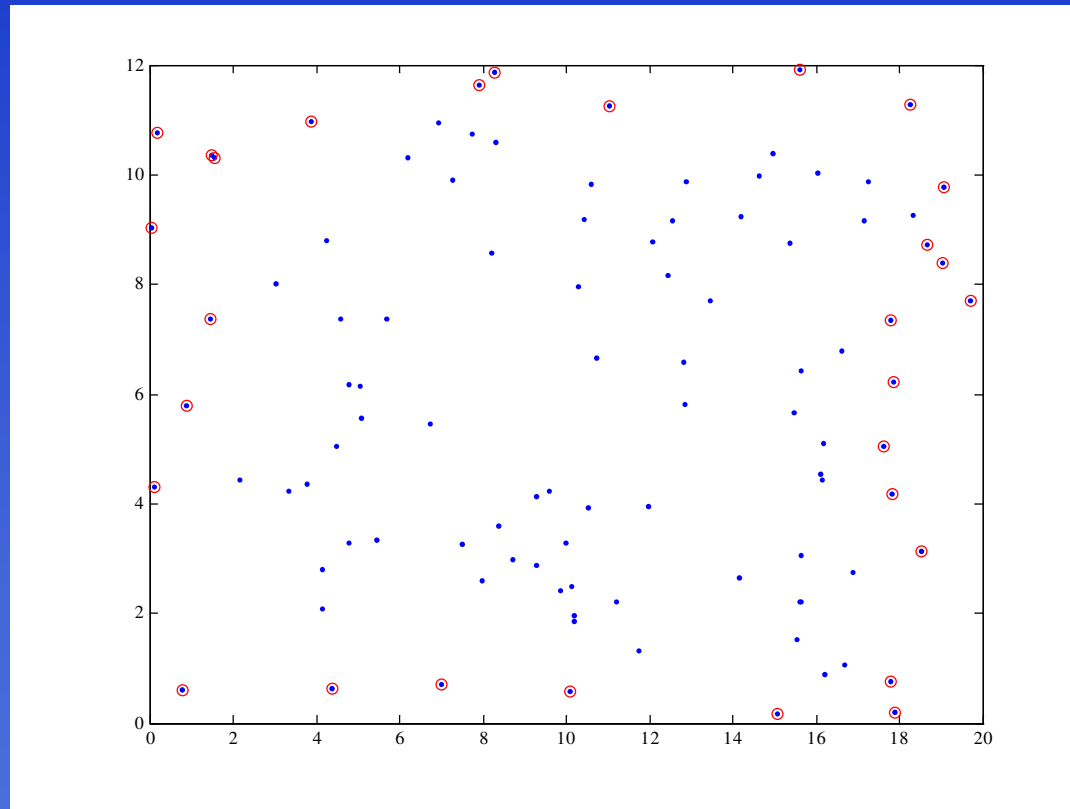
Isaksson et Næs ont identifié 3 règles d' optimalité pour les échantillons d' étalonnage:

- les échantillons retenus doivent présenter une **variabilité spectrale maximale**;
- la **plage de variation** des valeurs spectrales doit être **la plus grande possible**, mais **limitée aux valeurs rencontrées** dans la pratique;
- les échantillons doivent être **uniformément répartis**.

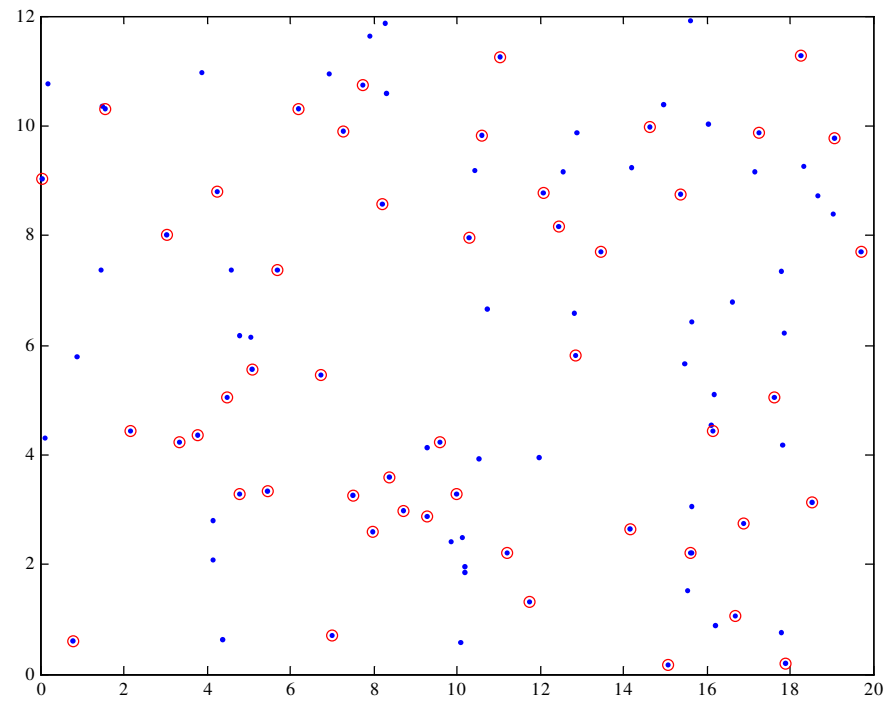
Exemple



Exemple



Exemple



Différentes méthodes de sélection d'échantillons

- Sélection aléatoire
- Répartition uniforme des échantillons sur la variable dépendante
- Algorithme de Kennard and Stone
- Algorithme DUPLEX
- OPTISIM
- etc.

Répartition uniforme des échantillons

$$y = X b$$

(

-toute variation de concentration est linéairement reliée à une variation spectrale

-chaque variation spectrale est due à une variation de concentration.

Répartition uniforme des échantillons

$y =$

2.1
2.5
3.2
3.3
4.2
4.8
5.2
5.3
5.6
5.8
6.2
6.3
6.4

- Les éléments de y sont classés par valeur croissante;

-on sélectionne alors les échantillons répartis régulièrement

Les échantillons en blanc seront utilisés pour **construire le modèle d'étalonnage.**

Les échantillons en rouge seront utilisés pour **valider le modèle d'étalonnage.**

Avantages / Inconvénients

- Méthode simple et rapide
- La répartition des échantillons est uniforme

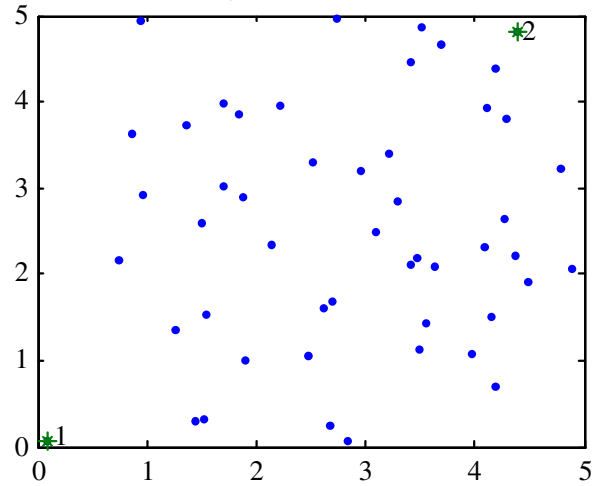
MAIS

Les variations spectrales dans X ne sont peut-être pas exclusivement liées à la variation dans y (la répartition n' est peut-être pas la plus appropriée

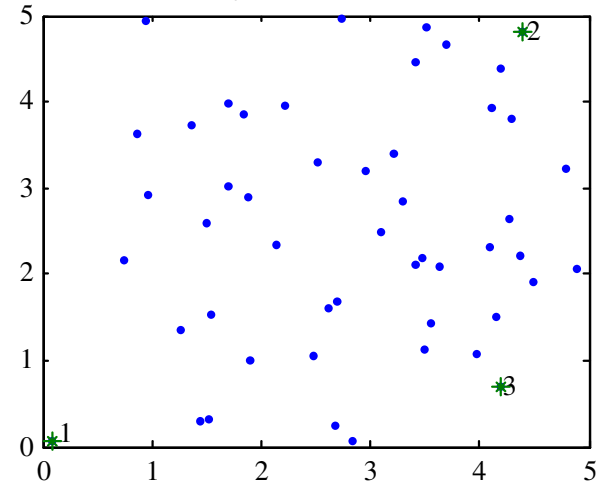
Algorithme de Kennard and Stone (K&S)

- Les **deux échantillons les plus éloignés l' un de l' autre** sont sélectionnés
- Pour chaque échantillon non sélectionné (e_i), l' algorithme
 - calcule la distance vers chaque échantillon déjà sélectionné
 - attribue à e_i la plus petite de ces distances
- ◇ L'échantillon e_i associé à la plus grande distance est donc le plus éloigné de tous les échantillons déjà sélectionnés : c' est donc lui qui est sélectionné.

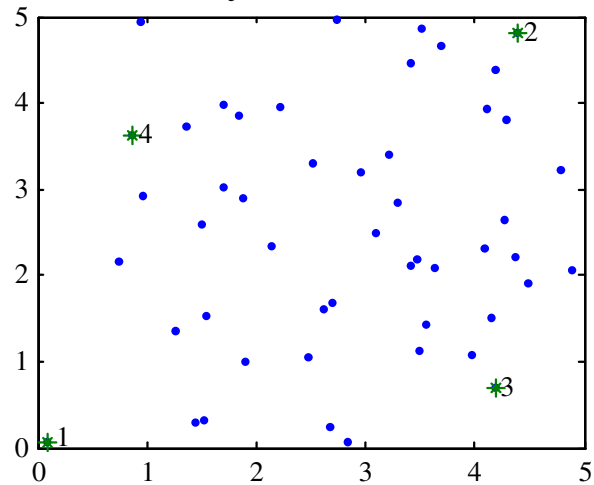
2 objets sélectionnés



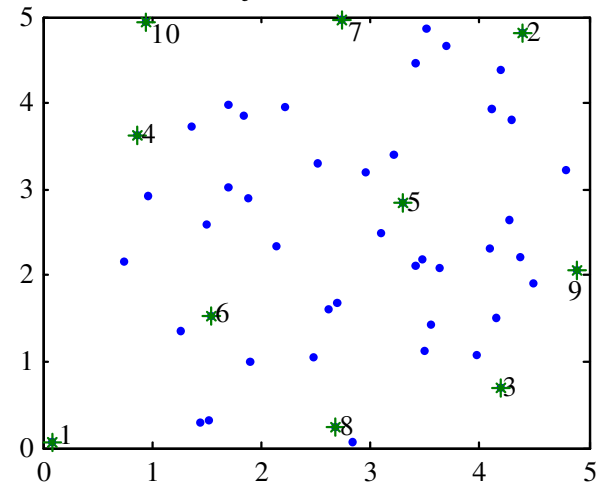
3 objets sélectionnés



4 objets sélectionnés



10 objets sélectionnés



Avantages / Inconvénients

- Le fait de sélectionner les échantillons les plus éloignés les uns des autres introduit une grande diversité dans le jeu d' étalonnage.
- Répartition uniforme des échantillons
- Toutefois, le domaine de variation spectral des échantillons de validation est moins "large" que celui des échantillons d' étalonnage, et le modèle n' est donc validé que sur une partie de ce domaine

L' **Algorithme DUPLEX** utilise alternativement deux algorithmes de Kennard and Stone

- Les deux échantillons les plus éloignés sont sélectionnés pour l' étalonnage
- Puis parmi les échantillons restants, les deux plus éloignés sont sélectionnés pour la validation

- Puis parmi les échantillons restants, le plus éloignés des échantillons d' étalonnage précédemment sélectionnés est sélectionné pour l' étalonnage
 - Puis parmi les échantillons restants, le plus éloignés des échantillons de validation précédemment sélectionnés est sélectionné pour la validation
- etc.

Ainsi, les deux domaines (étalonnage/validation) ont un domaine de variation similaire.

OPTISIM

Cette méthode nécessite la définition de 3 paramètres par l' utilisateur :

- k : Nombre d' échantillons à sélectionner
- ε : distance minimale entre deux échantillons sélectionnés
- S : Un nombre d' échantillons aléatoirement sélectionnés

OPTISIM

Le fonctionnement de cet algorithme est basé sur l' utilisation de 4 matrices :

- \mathbf{X} : Matrice initiale
- \mathbf{X}_k : Matrice des échantillons sélectionnés
- \mathbf{X}_s : Matrice d' échantillons aléatoirement sélectionnés
- \mathbf{X}_r : Matrice utilisée pour recycler les données

OPTISIM

1. L' échantillon le plus proche du spectre moyen est sélectionné
2. Sélectionner aléatoirement un échantillon e_i de \mathbf{X} , et calculer sa distance d_i par rapport à l' échantillon déjà sélectionné:

si $d_i \geq \varepsilon$, e_i est placé dans \mathbf{X}_s .

3. Répéter l'étape 2 jusqu'à l' obtention de s échantillons dans \mathbf{X}_s .
4. L'échantillon e_i associé à la plus grande distance d_i est introduit dans \mathbf{X}_k , et les autres échantillons de \mathbf{X}_s sont transférés vers \mathbf{X}_r .
5. Les étapes 2 à 4 sont répétées jusqu'à ce que k échantillons soient présents dans \mathbf{X}_k .

(Si \mathbf{X} est vide avant d' avoir obtenu k échantillons, les échantillons de \mathbf{X}_r sont transférés vers \mathbf{X} , et la procédure reprend)

Avantages / Inconvénients

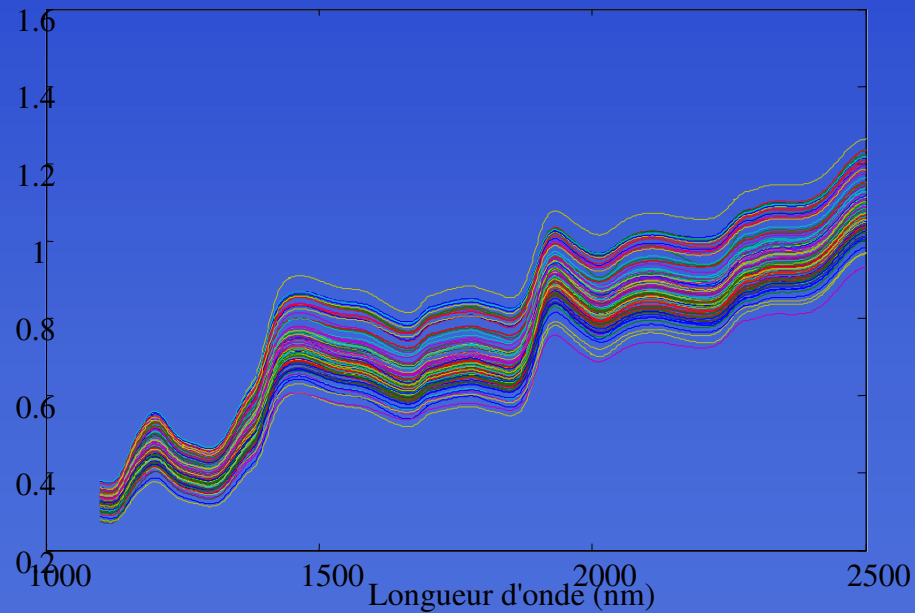
Cette méthode est une bonne alternative à KS pour les jeux de données volumineux (puisque on ne calcule pas toutes les distances).

Toutefois, elle est plus compliquée à mettre en place, puisque l'utilisateur doit lui-même choisir la valeur de trois paramètres, dont les résultats dépendront.

Application à des données réelles : Données "blé"

X : spectres proche infrarouge de 100 éch. de blé

y : teneur en eau de ces échantillons



On veut répartir les données en deux jeux de 50 échantillons.

Application de l' algorithme K&S sur les données "blé"

Application de l' algorithme Duplex sur les données "blé"

Comparaison des différents modèles

	Tous les objets	Aléatoire (1)	Aléatoire (2)	Aléatoire (3)	Kennard and Stone	DUPLEX
# LVs	3	3	3	3	3	3
% RMSEP		1.59	2.44	2.23	2.01	1.59
R ² (%)		97.38	68.88	95.50	96.09	97.39

Application à des données réelles : Données "essence"

X : spectres proche infrarouge de 60 éch. d' essence (30/30)

y : indice d' octane de ces échantillons

	Tous les objets	Aléatoire (1)	Aléatoire (2)	Aléatoire (3)	Kennard and Stone	DUPLEX
# LVs	3	3	5	3	3	4
% RMSEP		0.34	0.42	0.72	0.36	0.32
R ² (%)		96.67	49.91	73.04	96.7	96.60

Application à des données réelles : Données "betterave"

X : spectres proche infrarouge de 297 betteraves sucrières

y : teneur en sucre

	Tous les objets	Aléatoire (1)	Aléatoire (2)	Aléatoire (3)	Kennard and Stone	DUPLEX
# LVs	9	8	9	9	9	7
% RMSEP		1.48	1.58	2.39	1.44	1.65
R ² (%)		96.79	62.69	96.02	95.83	96.66

Conclusion

- Une sélection raisonnée des échantillons pour la construction d' un modèle d' étalonnage multivarié est essentielle.
- La sélection aléatoire est à déconseiller, les deux jeux de données ne recouvrant pas toujours le même domaine de variation.
- Les méthodes K&S et Duplex conduisent à des résultats satisfaisants, et assez similaires.

Conclusion

- La méthode Optimisim a été présentée comme une bonne alternative à K&S; toutefois, elle est plus compliquée à mettre en place
- La sélection d' échantillons revient, par nature, à sélectionner les spectres présentant les plus grandes différences (attention aux spectres aberrants.
- Un autre aspect important est le nombre d'échantillons à utiliser dans le jeu d'étalonnage.

Bibliographie

D. Jouan-Rimbaud Bouveresse, J. Maalouly, B. Jaillais, *Spectra Analyse* (2004), **33**, 23

M. Daszykowski, B. Walczak, D.L. Massart, *Anal. Chim. Acta* (2002), **468**, 91

D. Bertrand, "La spectroscopie infrarouge et ses applications analytiques", 2000, Tech. et Doc., Paris