Application de l'analyse des données fonctionnelles à l'identification de blé dur fusarié, moucheté et mitadiné

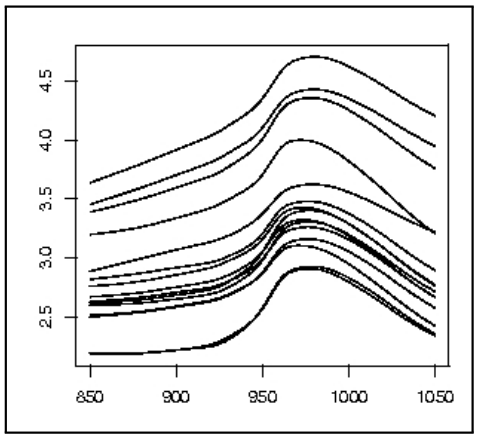
Nathalie Villa-Vialaneix
http://www.nathalievilla.org
En collaboration avec **Cécile Levasseur** (École d'Ingénieurs de Purpan) &
Fabrice Rossi (TELECOM ParisTech)

Institut de Mathématiques de Toulouse, France -
nathalie.villa@math.univ-toulouse.fr

Toulouse, Hélio-SPIR, 30 septembre 2009

Introduction

Statisticienne spécialisée dans l'analyse des données fonctionnelles :



$\longrightarrow$ Valeur d'intérêt, Y

Sommaire

Sommaire

Qu'est-ce qu'une donnée "fonctionnelle" ?

En théorie, une donnée fonctionnelle est une fonction,
$X : [0, 1] \to \mathbb{R}$.

Qu'est-ce qu'une donnée "fonctionnelle" ?

En théorie, une donnée fonctionnelle est une fonction, $X : [0, 1] \to \mathbb{R}$. À partir des observations de cette variable et d'une autre variable réelle d'intérêt, Y, on cherche à prédire Y à partir de X.

Qu'est-ce qu'une donnée "fonctionnelle" ?

En théorie, une donnée fonctionnelle est une fonction, $X : [0, 1] \to \mathbb{R}$. À partir des observations de cette variable et d'une autre variable réelle d'intérêt, Y, on cherche à prédire Y à partir de X.

En pratique, on n'observe jamais la fonction mais une discrétisation de celle-ci :

$$X(t_1), \ldots, X(t_d)$$

Problèmes posés par ce type de données

Du point de vue de la régression :

- Grande dimension : Le nombre de fonctions observées, n, est souvent plus petit que d (la dimension des données) $\longrightarrow$ Problème mal posé ;

Problèmes posés par ce type de données

Du point de vue de la régression :

- Grande dimension : Le nombre de fonctions observées, n, est souvent plus petit que d (la dimension des données) $\longrightarrow$ Problème mal posé ;
  $\Rightarrow$ Par ex, le modèle $Y = a^T X + b + \epsilon$ n'a plus une unique solution selon les moindres carrés.

Problèmes posés par ce type de données

Du point de vue de la régression :

- Grande dimension : Le nombre de fonctions observées, n, est souvent plus petit que d (la dimension des données) $\longrightarrow$ Problème mal posé ;
  $\Rightarrow$ Par ex, le modèle $Y = a^T X + b + \epsilon$ n'a plus une unique solution selon les moindres carrés.
- Discrétisations très corrélées entre elles pour une même observation du fait de la structure sous-jacente.

Problèmes posés par ce type de données

Du point de vue de la régression :

- Grande dimension : Le nombre de fonctions observées, n, est souvent plus petit que d (la dimension des données) $\longrightarrow$ Problème mal posé ;
  $\Rightarrow$ Par ex, le modèle $Y = a^T X + b + \epsilon$ n'a plus une unique solution selon les moindres carrés.
- Discrétisations très corrélées entre elles pour une même observation du fait de la structure sous-jacente.

Conséquences : Les méthodes statistiques appliquées aux données discrétisées donnent de mauvais résultats et, notamment, se généralisent mal à de nouvelles observations.

Comment traiter ce type de donnée ?

Méthodes de régularisation ou de réduction de dimension.

Comment traiter ce type de donnée ?

Méthodes de régularisation ou de réduction de dimension.
Si L^2 est l'ensemble des fonctions de carré intégrables, on sait qu'il existe une (des, en fait) base(s) de fonctions $(e_i)_i$ telles que toute fonction de L^2 s'exprime comme combinaison linéaire des $(e_i)_i$:

$$X = \sum_i \alpha_i e_i$$

Par exemple, la base trigonométrique :

$e_1(t) = \cos(t)$; $e_2(t) = \sin(t)$; $e_3(t) = \cos(2t)$; $e_4(t) = \sin(2t) \ldots$

Comment traiter ce type de donnée ?

Méthodes de régularisation ou de réduction de dimension.
Si L^2 est l'ensemble des fonctions de carré intégrables, on sait qu'il existe une (des, en fait) base(s) de fonctions $(e_i)_i$ telles que toute fonction de L^2 s'exprime comme combinaison linéaire des $(e_i)_i$:

$$X = \sum_i \alpha_i e_i \simeq \sum_{i=1}^{q} \alpha_i e_i$$

Par exemple, la base trigonométrique :

$$e_1(t) = \cos(t)\,;\ e_2(t) = \sin(t)\,;\ e_3(t) = \cos(2t)\,;\ e_4(t) = \sin(2t)\ldots$$

Pourquoi ? La dimension nécessaire pour bien représenter les données, q, est souvent très inférieure à d si la base est bien choisie.

Comment traiter ce type de donnée ?

Méthodes de régularisation ou de réduction de dimension.
Si L^2 est l'ensemble des fonctions de carré intégrables, on sait qu'il existe une (des, en fait) base(s) de fonctions $(e_i)_i$ telles que toute fonction de L^2 s'exprime comme combinaison linéaire des $(e_i)_i$:

$$X = \sum_i \alpha_i e_i \simeq \sum_{i=1}^{q} \alpha_i e_i$$

Par exemple, la base trigonométrique :

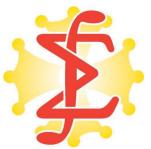
$$e_1(t) = \cos(t)\,;\ e_2(t) = \sin(t)\,;\ e_3(t) = \cos(2t)\,;\ e_4(t) = \sin(2t)\ldots$$

Pourquoi ? La dimension nécessaire pour bien représenter les données, q, est souvent très inférieure à d si la base est bien choisie. Cette représentation permet d'avoir accès à des opérations fonctionnelles comme la dérivée.

Sommaire

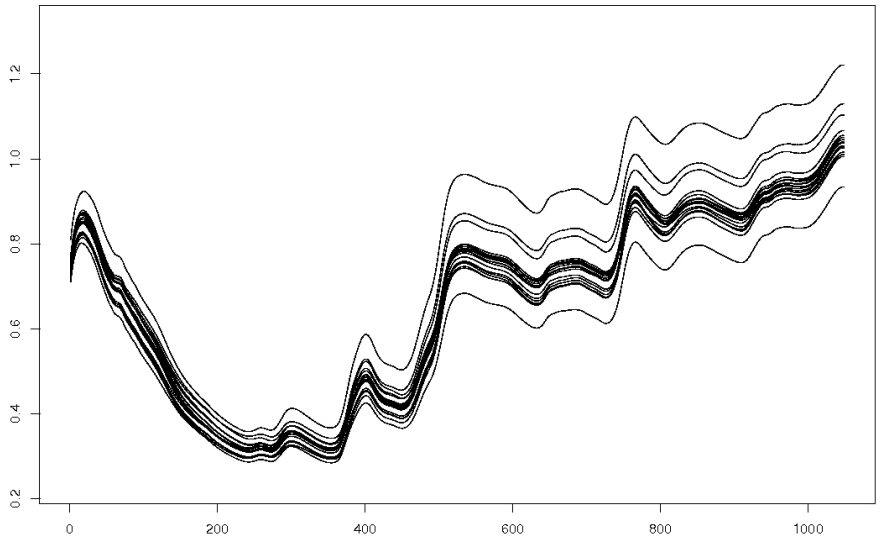
Principe général

Principe général : L'hypothèse de base est que la fonction sous-jacente est régulière.



Principe général

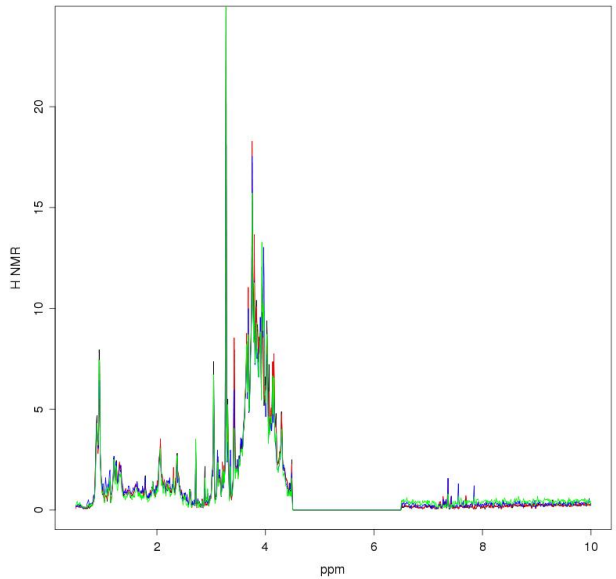
Principe général : L'hypothèse de base est que la fonction sous-jacente est régulière.



convient pour des spectres infra-rouges

Principe général

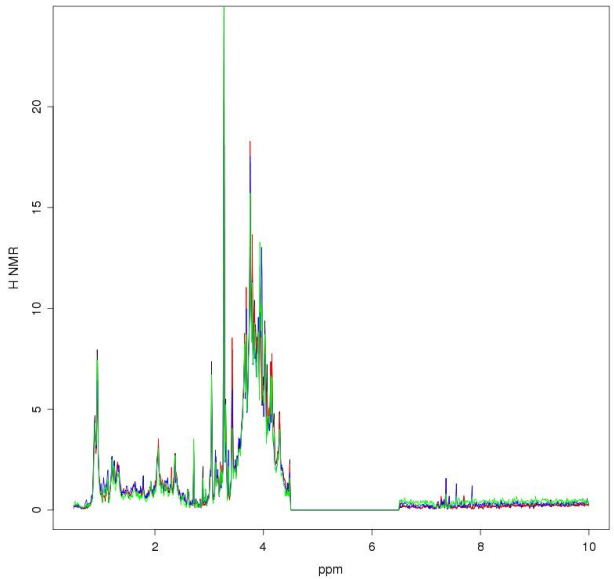
Principe général : L'hypothèse de base est que la fonction sous-jacente est régulière.

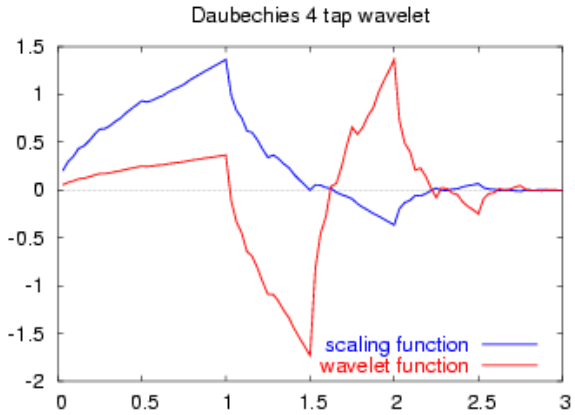


ne convient pas pour la spectrométrie de masse

Principe général

Principe général : L'hypothèse de base est que la fonction sous-jacente est régulière.





Dans ce cas, utilisation d'une projection sur des bases d'ondelettes... (Travail en cours avec A. Paris (INRA) et N. Hernandez (CENATAV, Cuba))

Principe général

Principe général : L'hypothèse de base est que la fonction sous-jacente est régulière.
Représentation des spectres dans un espace $\mathcal{H}^m$ dans lequel la norme d'un spectre est sensiblement égal à la norme de sa dérivée d'ordre m

Principe général

Principe général : L'hypothèse de base est que la fonction sous-jacente est régulière.
Représentation des spectres dans un espace $\mathcal{H}^m$ dans lequel la norme d'un spectre est sensiblement égal à la norme de sa dérivée d'ordre m $\Rightarrow$ régularité et prise en compte de la courbure !

Mise en œuvre pratique

Les données :

$$\begin{matrix} x_1(t_1) \dots x_1(t_d) \\ \dots \\ x_n(t_1) \dots x_n(t_d) \end{matrix}$$

Mise en œuvre pratique

Les données :
$$\begin{matrix} x_1(t_1) \dots x_1(t_d) \\ \dots \\ x_n(t_1) \dots x_n(t_d) \end{matrix}$$

Étape 1 : "Reconstruire", à partir de $x_i(t_1), \dots, x_i(t_d)$, la fonction sous-jacente. Elle est approchée par $\hat{x}_i$ telle que

$$\sum_{l=1}^{d} (x_i(t_l) - \hat{x}_i(t_l))^2 \text{ est petit} \quad \text{et} \quad \|D^m \hat{x}_i\|^2 \text{ est petit}$$

Mise en œuvre pratique

Les données :
$$\begin{matrix} x_1(t_1) \dots x_1(t_d) \\ \dots \\ x_n(t_1) \dots x_n(t_d) \end{matrix}$$

Étape 1 : "Reconstruire", à partir de $x_i(t_1), \dots, x_i(t_d)$, la fonction sous-jacente. Elle est approchée par $\hat{x}_i$ telle que

$$\sum_{l=1}^{d} (x_i(t_l) - \hat{x}_i(t_l))^2 \text{ est petit} \quad \text{et} \quad \|D^m \hat{x}_i\|^2 \text{ est petit}$$

Fidélité aux données Régularité

Mise en œuvre pratique

Les données :
$$\begin{matrix} x_1(t_1) \dots x_1(t_d) \\ \dots \\ x_n(t_1) \dots x_n(t_d) \end{matrix}$$

Étape 1 : "Reconstruire", à partir de $x_i(t_1), \dots, x_i(t_d)$, la fonction sous-jacente. Elle est approchée par $\hat{x}_i$ telle que

$$\sum_{l=1}^{d} (x_i(t_l) - \hat{x}_i(t_l))^2 \text{ est petit} \quad \text{et} \quad \|D^m \hat{x}_i\|^2 \text{ est petit}$$

Fidélité aux données Régularité

Étape 2 : Utiliser $D^m \hat{x}^i$ comme entrée d'une méthode de régression (ici SVM) $\Rightarrow$ nécessite de savoir calculer

$$\langle D^m \hat{x}_i, D^m \hat{x}_j \rangle$$

Mise en œuvre pratique

Les données :
$$\begin{matrix} x_1(t_1) \dots x_1(t_d) \\ \dots \\ x_n(t_1) \dots x_n(t_d) \end{matrix}$$

Étape 1 : "Reconstruire", à partir de $x_i(t_1), \dots, x_i(t_d)$, la fonction sous-jacente. Elle est approchée par $\hat{x}_i$ telle que

$$\sum_{l=1}^{d} (x_i(t_l) - \hat{x}_i(t_l))^2 \text{ est petit} \quad \text{et} \quad \|D^m \hat{x}_i\|^2 \text{ est petit}$$

Fidélité aux données Régularité

Étape 2 : Utiliser $D^m \hat{x}^i$ comme entrée d'une méthode de régression (ici SVM) $\Rightarrow$ nécessite de savoir calculer

$$\langle D^m \hat{x}_i, D^m \hat{x}_j \rangle \simeq \mathbf{Q}_{d,m} \times \begin{pmatrix} x_1(t_1) \dots x_1(t_d) \\ \dots \\ x_n(t_1) \dots x_n(t_d) \end{pmatrix}$$

Conclusion théorique

L'utilisation d'une méthode de régression de type SVM (voir Vapnik, *The Nature of Statistical Learning Theory*) permet de garantir que la méthode atteint asymptotiquement une performance optimale (lorsque n est "suffisamment grand" et d est "suffisamment grand").

Sommaire

Présentation des données

953 échantillons de blé dur ont été analysés :

- spectrométrie infra-rouge : 1049 longueurs d'onde uniformément réparties entre 400 et 2498 nm ;

Présentation des données

953 échantillons de blé dur ont été analysés :

- spectrométrie infra-rouge : 1049 longueurs d'onde uniformément réparties entre 400 et 2498 nm ;
- fusariose : déterminée en % de la masse des grains par triage préalable des grains affectés ;

Présentation des données

953 échantillons de blé dur ont été analysés :

- spectrométrie infra-rouge : 1049 longueurs d'onde uniformément réparties entre 400 et 2498 nm ;
- fusariose : déterminée en % de la masse des grains par triage préalable des grains affectés ;
- moucheture : déterminé en % de la masse des grains par triage préalable des grains affectés ;

Présentation des données

953 échantillons de blé dur ont été analysés :

- spectrométrie infra-rouge : 1049 longueurs d'onde uniformément réparties entre 400 et 2498 nm ;
- fusariose : déterminée en % de la masse des grains par triage préalable des grains affectés ;
- moucheture : déterminé en % de la masse des grains par triage préalable des grains affectés ;
- mitadinage : déterminé en % du nombre de grains affectés par comptage.

Présentation des données

953 échantillons de blé dur ont été analysés :

- spectrométrie infra-rouge : 1049 longueurs d'onde uniformément réparties entre 400 et 2498 nm ;
- fusariose : déterminée en % de la masse des grains par triage préalable des grains affectés ;
- moucheture : déterminé en % de la masse des grains par triage préalable des grains affectés ;
- mitadinage : déterminé en % du nombre de grains affectés par comptage.

Question : Comment prédire les valeurs de qualité correspondant à la fusariose, à la moucheture et au mitadinage à partir de la collecte des spectres infra-rouge ?

Présentation des données

953 échantillons de blé dur ont été analysés :

- spectrométrie infra-rouge : 1049 longueurs d'onde uniformément réparties entre 400 et 2498 nm ;
- fusariose : déterminée en % de la masse des grains par triage préalable des grains affectés ;
- moucheture : déterminé en % de la masse des grains par triage préalable des grains affectés ;
- mitadinage : déterminé en % du nombre de grains affectés par comptage.

Question : Comment prédire les valeurs de qualité correspondant à la fusariose, à la moucheture et au mitadinage à partir de la collecte des spectres infra-rouge ?

Les méthodes habituelles (PLS, réseau de neurones ...) donnent ici des résultats décevants.

Présentation des données

953 échantillons de blé dur ont été analysés :

- spectrométrie infra-rouge : 1049 longueurs d'onde uniformément réparties entre 400 et 2498 nm ;
- fusariose : déterminée en % de la masse des grains par triage préalable des grains affectés ;
- moucheture : déterminé en % de la masse des grains par triage préalable des grains affectés ;
- mitadinage : déterminé en % du nombre de grains affectés par comptage.

Question : Comment prédire les valeurs de qualité correspondant à la fusariose, à la moucheture et au mitadinage à partir de la collecte des spectres infra-rouge ?

Les méthodes habituelles (PLS, réseau de neurones ...) donnent ici des résultats décevants. ⇒ Présentation des résultats de la mise en œuvre de la méthode sur le mitadinage.

Méthodologie pour évaluation de la validité de l'approche par splines

- Séparation aléatoire du jeu de données en apprentissage (pour la définition de la régression) et test (pour évaluer les performances de la régression) : cette séparation est répétée 50 fois pour évaluer la variabilité de la qualité de l'approximation ;

Méthodologie pour évaluation de la validité de l'approche par splines

- Séparation aléatoire du jeu de données en apprentissage (pour la définition de la régression) et test (pour évaluer les performances de la régression) : cette séparation est répétée 50 fois pour évaluer la variabilité de la qualité de l'approximation ;
- Sur les 50 ensembles d'apprentissage, des erreurs de prédiction sont calculées :

$$MSE = \frac{1}{|\text{Test}_k|} \sum_{\text{Test}_k} (\text{Mitadinage} - \text{Mitadinage estimé par la régression})^2$$

Méthodologie pour évaluation de la validité de l'approche par splines

- Séparation aléatoire du jeu de données en apprentissage (pour la définition de la régression) et test (pour évaluer les performances de la régression) : cette séparation est répétée 50 fois pour évaluer la variabilité de la qualité de l'approximation ;
- Sur les 50 ensembles d'apprentissage, des erreurs de prédiction sont calculées :

$$MSE = \frac{1}{|\text{Test}_k|} \sum_{\text{Test}_k} (\text{Mitadinage} - \text{Mitadinage estimé par la régression})^2$$

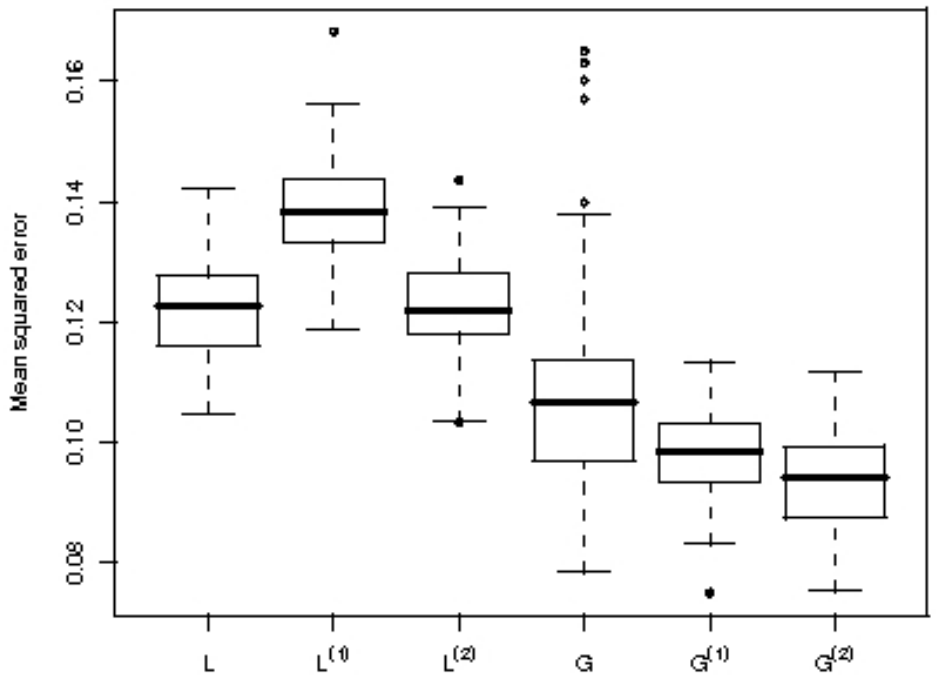
- Les divers paramètres du modèle sont évalués par validation croisée sur l'ensemble d'apprentissage.

Résultats

Méthodes comparées : SVM linéaire et non linéaire (Gaussien) sur les données initiales et les dérivées d'ordre 1 à 2 déterminées par splines.

Résultats

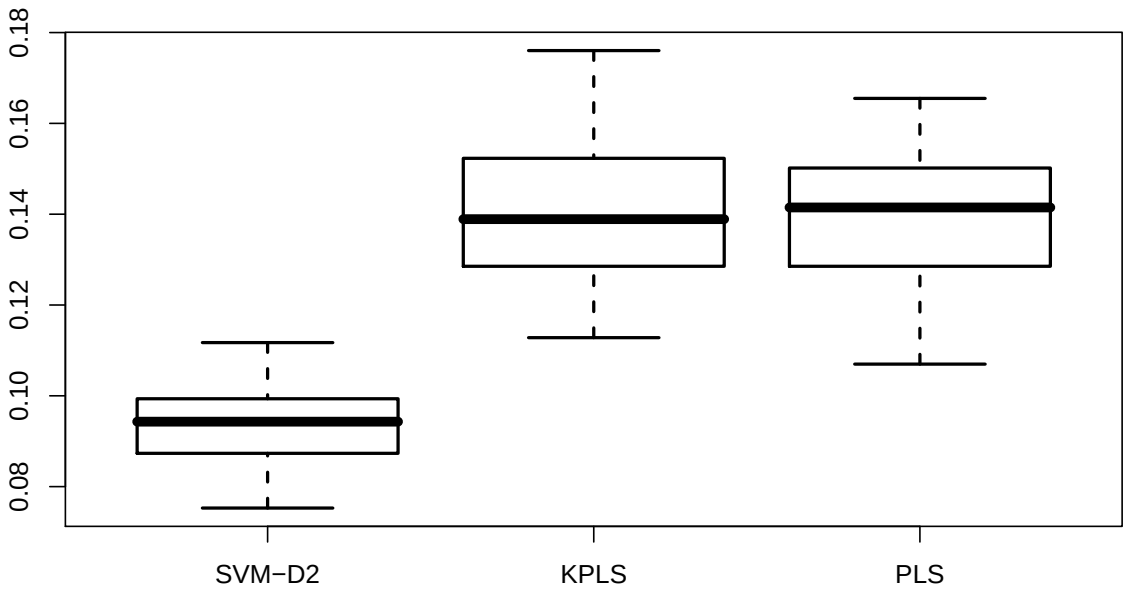
Méthodes comparées : SVM linéaire et non linéaire (Gaussien) sur les données initiales et les dérivées d'ordre 1 à 2 déterminées par splines.



Pour comparaison avec PLS...

	MSE moyenne (test)	Écart type MSE
PLS sur données initiales	0.154	0.012
Kernel PLS	0.154	0.013
SVM splines (reg. D^2)	0.094	0.008

Gain de près de 40 % sur la prédiction moyenne.



Conclusions

Résumé des résultats : Les différences sont significatives entre

- l'utilisation des dérivées d'ordre 2 et d'ordre 1 ;
- l'utilisation des dérivées et l'utilisation des données initiales ;
- l'utilisation d'une approche non linéaire et d'une approche linéaire.

Conclusions

Résumé des résultats : Les différences sont significatives entre

- l'utilisation des dérivées d'ordre 2 et d'ordre 1 ;
- l'utilisation des dérivées et l'utilisation des données initiales ;
- l'utilisation d'une approche non linéaire et d'une approche linéaire.

Perspectives :

- recherche des endroits du spectres impliqués dans la prédiction ;
- comparaison méthodologique similaire avec les splines combinées avec diverses méthodes : PLS, forêts aléatoires ...