Chimiométrie 2004, Paris 30 Nov. et 1er déc. 2004, Groupe Français de Chimiométrie, pp 108-111.

La discrimination par régression PLS sur indicatrices

Dominique **Bertrand***, Philippe **Courcoux***, Cédric **Camps** **
* Unité de Sensométrie et de Chimiométrie. INRA-ENITIAA. Rue de la Géraudière-BP 82 225 44322 Nantes cedex 03
** Institut National d'Horticulture, 2 rue LeNôtre 49000 Angers, France

MOTS CLÉS : discrimination, PLS, indicatrices

1. Introduction

La régression PLS est le plus couramment appliquée sur des variables quantitatives. Cependant, un nombre croissant d'utilisateurs utilise également cette méthode dans le cadre de la discrimination [BAR 2003]. Lorsque la matrice des variables prédictives est mal conditionnée, la discrimination PLS permet d'obtenir des résultats robustes, en général supérieurs aux méthodes alternatives telles que la discrimination sur composantes principales [BERT 1990]. La discrimination PLS repose sur un principe simple, qui consiste à remplacer le tableau de la variable qualitative à prédire par le tableau des indicatrices. Plus précisément, soit $\boldsymbol{X}$ la matrice des variables prédictives, de dimensions $n \times p$ et $\boldsymbol{y}$ ($n \times 1$) le vecteur indiquant l'appartenance d'une observation à un groupe donné. Les éléments de $\boldsymbol{y}$ sont des nombres entiers, allant de 1 à g . Un élément y_i de $\boldsymbol{y}$ indique que l'individu d'indice i appartient au groupe de numéro y_i. Pour mettre en œuvre la discrimination PLS, on remplace le vecteur $\boldsymbol{y}$ par une matrice des indicatrices, notée $\boldsymbol{Z}$, de dimensions $n \times g$. Chaque élément de z_{ij} de $\boldsymbol{Z}$ prend la valeur 1 pour indiquer que l'individu i appartient au groupe j, et 0 dans le cas contraire. La régression PLS2 est alors appliquée en prenant $\boldsymbol{Z}$ comme matrice des variables dépendantes. On peut montrer que le mode de codage de $\boldsymbol{Z}$ ne joue pas de rôle dans l'analyse.

Dans certaines situations, l'expérimentateur peut disposer de plusieurs variables qualitatives pour décrire la même observation. C'est le cas, par exemple, lorsque l'expérimentateur s'intéresse à la composition biochimique de différentes variétés d'une même espèce végétale cultivée dans différentes régions. Les variétés et les régions peuvent être considérées comme étant deux variables qualitatives.

L'analyse en variables canoniques est susceptible de traiter ce type de données. Cependant, cette méthode est en général plutôt considérée comme une extension multidimensionnelle de l'analyse de variance, c'est à dire que l'une des variables qualitatives est vue comme étant représentative d'une variable incontrôlée dont on souhaite limiter l'effet. De plus, cette méthode n'est pas directement applicable dans le cas de variables prédictives présentant une quasi-colinéarité.

Il paraît naturel d'étendre la discrimination PLS au cas où chaque observation est décrite par plusieurs variables qualitatives. Soit $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k$ ces variables. On peut comme précédemment créer pour chacune d'entre elle, les tableaux des indicatrices $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_k$.

Le nombre de colonnes de chacun de ces tableaux est égal au nombre de groupes dans chacune des variables qualitatives. On forme alors le tableau concaténé des indicatrices $\boldsymbol{R}$, soit : $\boldsymbol{R}=[\boldsymbol{Z}_1|\boldsymbol{Z}_2|\ldots|\boldsymbol{Z}_k]$. La régression PLS2 est ensuite appliquée sur $\boldsymbol{X}$ et $\boldsymbol{R}$.

A titre illustratif, cette approche a été appliquée à des populations de courbes de pénétrométrie utilisées pour caractériser les propriétés mécaniques de pommes entières.

2. Matériel et Méthodes

On tente de caractériser le degré de maturité de pommes et de suivre son évolution au cours du stockage des fruits, à température constante, pendant des temps variables. L'essai porte sur deux variétés de fruits, *Gala* et *Elstar*, et les temps de stockage sont de 8, 14, 21 et 28 jours. Les variétés et les durées de stockage sont considérées comme formant deux variables qualitatives différentes. Environ 40 fruits pour chaque couple {*variété* ; *temps de stockage*} sont soumis à un test de pénétrométrie (Figure 1).

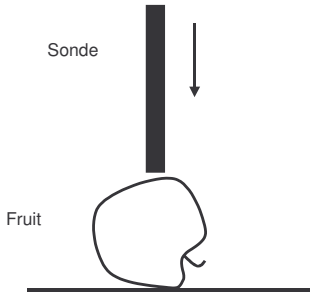

Figure 1 : Principe de l'essai de pénétrométrie

Dans ces essais, une sonde calibrée se déplace à vitesse constante et finit par pénétrer dans le fruit étudié. On enregistre la force appliquée au fruit en fonction du déplacement de la sonde . On obtient ainsi une courbe avec en abscisses le déplacement observé, et en ordonnées la force appliquée (Figure 2).

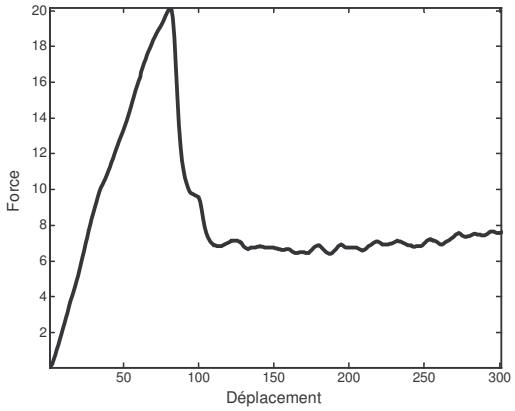

Figure 2 : Courbe de Pénétrométrie

Chaque fruit étudié donne ainsi une courbe qui est numérisée sur 300 points. On a collecté 315 courbes associées à différents fruits. Par nature, les courbes pénétrométriques sont très irrégulières et présentent une forte variabilité. La matrice des données $\boldsymbol{X}$ est ainsi formée de 315 lignes et 300 colonnes.
La matrice $\boldsymbol{R}$ est formée par les indicatrices associées aux variétés et aux temps de stockage.

3 . Résultats

Afin de permettre une comparaison, une analyse en composantes principales a été appliquée sur la matrice ***X*** formées par les courbes de pénétrométrie. Les ellipses de confiance des observations associées à chaque variété pour un temps de stockage donné (au seuil de 95%) sont présentée sur la figure 3.

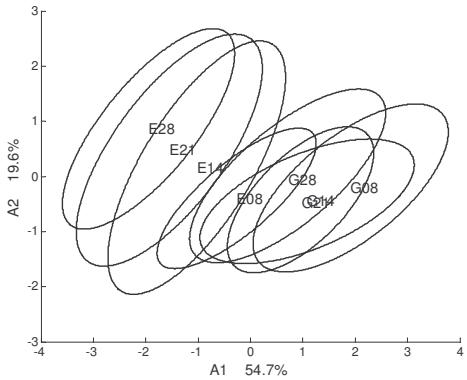

Figure 3 : Premier plan factoriel de l'ACP des courbes pénétrométriques.
Ellipses de confiance des groupes Variété x temps (seuil de 95%)
Code : E *:* Elstar *;* G *:*Gala*.* 08 ; 14 ; 21 ; 28 *: temps de stockage.*

On voit que les deux variétés codées *E* et *G* forment deux groupes assez distincts sur le premier plan de cette ACP. Les temps de stockage, notamment pour la variété *G* sont en revanche assez faiblement séparés.

Les résultats de la régression PLS2 (5 dimensions) sont illustrés par la figure 4. Nous présentons ici la carte factorielle obtenue après une ACP sur les indicatrices prédites par PLS2. La représentation directe des variables latentes de PLS2, qui est plus habituelle, donne des résultats moins facilement interprétables, qui ne sont pas présentés ici.

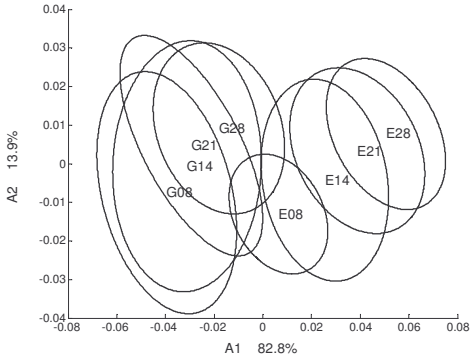

Figure 4 : Premier plan factoriel de l'ACP des indicatrices estimées par PLS2.
Ellipse de confiance des groupes Variété x temps (seuil de 95%)
Code : E *:* Elstar *;* G *:*Gala*.* 08 ; 14 ; 21 ; 28 *: temps de stockage.*

Sur cette représentation factorielle, les deux variétés sont assez bien séparées suivant la première composante. La deuxième composante est principalement représentative du temps de stockage. Les barycentres forment un arc paramétré dans lequel les temps de stockage sont ordonnés, ce qui semble logique du point de vue expérimental.
Le modèle PLS2 permet *in fine* de prédire les indicatrices par une fonction linéaire appliquée sur $\boldsymbol{X}$, soit $\hat{\boldsymbol{R}}=\boldsymbol{X}\hat{\boldsymbol{\beta}}$. Il est intéressant d'examiner les colonnes de $\hat{\boldsymbol{\beta}}$ comme des « profils pénétrométriques » associés aux indicatrices de certains groupes. A titre d'exemple, la figure 5 montre le graphe des colonnes de $\hat{\boldsymbol{\beta}}$ associées aux temps de stockage 8 et 21 jours. Malgré la nature très bruitée de ces courbes, on peut voir que la première partie de la courbe (avant le déplacement 50) évolue très fortement en fonction du temps de stockage. Cette partie représente la force appliquée avant la rupture de la cuticule (« peau ») du fruit. La partie de la courbe située après le déplacement 200 possède également un fort pouvoir discriminant. Cette région correspond à une pénétration profonde de la sonde à l'intérieur du fruit.

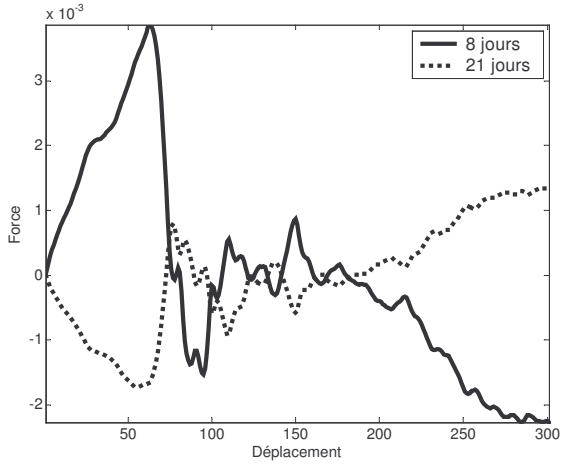

Figure 5 : Coefficients $\hat{\boldsymbol{\beta}}$ associés aux indicatrices des groupes « 08 » et « 21 » jours. (Représentation sous la forme de profils).

4 . Conclusion

La régression PLS2 appliquée sur des indicatrices peut être étendue utilement au cas où chaque observation appartient à plusieurs groupes qualitatifs. La possibilité offerte par PLS d'optimiser les dimensions du modèle, de manière à éviter les problèmes de surajustement, est particulièrement intéressante dans le cas de données quasi-colinéaires comme les courbes numérisées et les signaux physiques. Un autre intérêt, non présenté ici, est la possibilité de prédire l'appartenance d'un échantillon à un groupe inconnu en tenant compte d'autres informations qualitatives disponibles.

Bibliographie

[BAR 2003], BARKER M., RAYENS W. Partial least squares for discrimination. *Journal of Chemometrics*. 2003, Vol **17**, 166-173.
[BERT 1990], BERTRAND D., COURCOUX P., *et al.*. Stepwise discriminant analysis of continuous digitized signals. *Journal of Chemometrics*. 1990, Vol **4**, 413-427.