Chimiométrie 2005
Villeneuve d'Ascq (France), 30 November - 1 December 2005, pp. 52-55

Sélection de groupes de variables spectrales par information mutuelle grâce à une représentation spline

Fabrice **Rossi**[1], Damien **François**[2], Vincent **Wertz**[2], Michel **Verleysen**[3]
[1] *Projet AxIS, INRIA-Rocquencourt, Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex, France,* Fabrice.Rossi@inria.fr.
[2] *Université catholique de Louvain – Machine Learning Group, CESAME, 4 av. G. Lemaître, 1348 Louvain-la-Neuve, Belgique,* francois@csam.ucl.ac.be, wertz@inma.ucl.ac.be
[3] *Université catholique de Louvain – Machine Learning Group, DICE, 3 place du Levant, 1348 Louvain-la-Neuve, Belgique,* verleysen@dice.ucl.ac.be

MOTS CLÉS : Spectroscopie, modèles non-linéaires multivariés, sélection de variables, information mutuelle, B-splines

1. Introduction

De nombreuses applications de la spectrométrie nécessitent la prédiction d'une variable quantitative à partir des variables spectrales mesurées ; l'apprentissage d'un modèle sur base de données connues est alors réalisé. Souvent, le modèle utilisé est linéaire ; on peut citer comme exemples la régression sur composantes principales (PCR), la régression en moindres carrés partiels (PLSR) ou la régression linéaire multiple pas-à-pas (SMLR) [BER 00], [MAS 97]. Dans certains cas cependant, la relation physique à modéliser se distancie fortement d'une relation linéaire ; des modèles non-linéaires doivent alors être utilisés. Par rapport aux modèles linéaires, les modèles non-linéaires ont un potentiel (en termes de précision) plus important car ils peuvent a priori modéliser n'importe quelle relation entre les variables spectrales et la variable à prédire. Ce potentiel se paie néanmoins par un plus grand nombre de paramètres à ajuster, des procédures d'apprentissage plus complexes, et une interprétabilité plus difficile.

Les problèmes de modélisation en spectrométrie comportent un grand nombre de variables d'entrée (les variables spectrales), et ne sont en général connus qu'à travers un nombre réduit d'exemples ; il est alors est nécessaire de construire les modèles sur un nombre réduit de variables, déterminées à partir des variables spectrales initiales. Il s'agit des variables latentes, que l'on trouve dans les modèles linéaires cités ci-dessus, et qui sont en général construites grâce à une mesure de corrélation. Dans le cas de modèles non-linéaires cependant, la corrélation, qui ne mesure que des phénomènes linéaires, n'est pas appropriée ; il faut utiliser la mesure non-linéaire correspondante, à savoir l'information mutuelle [ROS05]. Par rapport à la corrélation, l'information mutuelle nécessite néanmoins le recours à des estimateurs plus complexes et plus bruités, ce qui entraîne respectivement des temps-calcul plus élevés et des résultats moins précis.

Pour construire un modèle non-linéaire sur un nombre réduit de variables issues de N variables spectrales, on pourrait songer à tester les 2^N sous-ensembles possibles et à sélectionner celui ayant la plus grande information mutuelle avec la variable à prédire. Le temps-calcul d'une telle procédure serait néanmoins prohibitif ; de plus, les estimations d'information mutuelle étant peu précises, cette procédure mènerait à la sélection de variables spectrales particulières, alors que d'autres, proches dans le spectre et donc proches (et corrélées) aux premières auraient pu donner un résultat similaire, voir meilleur.

Dans cet article, nous proposons d'utiliser un critère d'information mutuelle non pas pour sélectionner des variables spectrales individuelles, mais bien des groupes de variables consécutives. Pour ce faire, les spectres sont projetés sur des bases de B-splines [BOR 78], ces derniers ayant l'avantage d'être bien localisés en longueur d'onde. Ce sont alors les B-splines eux-mêmes qui sont sélectionnés à travers leurs coefficients. Nous montrons que la combinaison de la projection des spectres sur des B-splines, la sélection de ceux-ci par information mutuelle et la construction d'un modèle non-linéaire sur base des coefficients de B-splines sélectionnés, permet d'obtenir des performances de prédiction meilleures que dans le cas de modèles linéaires, en conservant la possibilité d'interpréter

les plages spectrales obtenues.

La section 2 décrit le concept et l'estimation de l'information mutuelle entre variables ; la section 3 rappelle brièvement le principe de la représentation de spectres sur une base de B-splines, avant la présentation de résultats expérimentaux dans la section 4.

2. Sélection de variables par information mutuelle

L'information mutuelle (IM) est une mesure non paramétrique de dépendance entre deux variables aléatoires, éventuellement vectorielles. Elle peut être définie comme la réduction d'entropie de la variable *X* apportée par la connaissance de la variable *Y*, c'est-à-dire :

$$\mathrm{IM}(X,Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y),$$

où *H(X)* est l'entropie de *X* et *H(X|Y)* l'entropie conditionnelle de *X* sachant *Y*. Les variables X et Y peuvent être des vecteurs, ce qui permet de calculer l'IM sur des groupes de variables réelles. Ceci est cependant délicat car la plupart des méthodes de calcul de l'entropie s'appuient sur une estimation de la densité de probabilité de la variable considérée, ce qui est difficile en dimension élevée. Une méthode récente [KRA 04] permet cependant ce calcul, même quand le nombre de variables considérées est important. Dans le cas de spectres, il est donc possible de calculer $IM(X_{[s]},Y)$ où $X_{[s]}$ est un sous-ensemble des longueurs d'onde présentes dans le spectre d'origine et *Y* la grandeur numérique à prédire. On peut ainsi chercher le sous-ensemble qui explique le mieux la cible. Cependant, le calcul de $IM(X_{[s]},Y)$ est assez coûteux car la méthode de [KRA 04] est basée sur un calcul de plus proches voisins. Le temps de calcul est ainsi proportionnel au nombre de variables considérées, i.e., $|s|+1$. Il est dès lors totalement impossible en pratique de tester les 2^N sous-ensembles de variables spectrales. On applique donc une procédure heuristique de type *forward backward* [ROS05].

Cette procédure consiste dans la phase *forward* à ordonner les variables spectrales en utilisant l'IM. La première variable X_1 est celle qui maximise $IM(X_i,Y)$ quand X_i parcourt toutes les variables spectrales. La deuxième variable X_2 maximise $IM((X_1, X_i), Y)$ quand X_i parcourt toutes les variables spectrales sauf X_1. Plus généralement, X_k maximise $IM((X_1,X_2,...,X_{k-1},X_k),Y)$ pour X_k choisi dans les variables spectrales restantes après *k*-1 itérations. En théorie, on doit avoir $IM((X_1,...,X_{k-1}),Y) <= IM((X_1,...,X_k),Y)$, mais l'estimation de l'IM n'est pas parfaite et les mesures sont bruitées. Il est donc possible que l'IM de la suite de variables décroisse à partir d'un certain rang. On se contente donc de calculer l'ordre des *M* meilleures variables, puis de déterminer *k* tel que $IM((X_1,X_2,...,X_k),Y)$ soit maximal. On applique alors une phase *backward* d'élagage qui consiste à supprimer les variables qui ne font pas décroître l'IM : on cherche s tel que $IM((X_1,...,X_k),Y) <= IM((X_1,...,X_{s-1},...,X_{s+1},...,X_k),Y)$ et on supprime X_s de la liste courante. On répète cette opération de suppression jusqu'à ne plus pouvoir faire augmenter l'IM.

Le temps de calcul de cette procédure est proportionnel à L^2NM^2, où *L* désigne le nombre de spectres, ce qui peut être prohibitif quand *M* est grand, c'est-à-dire quand on cherche à obtenir des résultats précis en classant le plus de variables possibles dans la phase *forward*.

3. Représentation des spectres sur une base de B-splines

La sélection de variables spectrales brutes par information mutuelle présente deux inconvénients : elle est coûteuse en temps-calcul, et son estimation est imprécise, ce qui résulte en la sélection de variables spectrales particulières, alors que d'autres, proches dans le spectre et donc proches (et corrélées) aux premières auraient pu donner un résultat similaire, voir meilleur.

Pour réduire le nombre de variables à sélectionner, nous proposons d'utiliser la régularité des spectres. Ceux-ci peuvent en effet être représentés sans perte sur une base de B-splines [BOR 78] comportant beaucoup moins de fonctions que de variables spectrales d'origine. Considérons en effet un intervalle de longueurs d'onde, [a,b] et un entier p>1. On découpe l'intervalle [a,b] en p sous-intervalles réguliers et on considère les splines d'ordre d associés à ce découpage, c'est-à-dire les fonctions C^{d-2} dont la restriction à chaque

intervalle est un polynôme de degré *d-1*. Cet ensemble de fonctions possède une base de *p-1+d* B-splines. Chaque B-spline est elle-même une spline localisée, nulle en dehors de d intervalles consécutifs.

Nous proposons donc de représenter chaque spectre par son approximation (au sens des moindres carrés) sur une base de B-splines. L'ordre d est fixé à 4 ou 5 selon la régularité attendue pour les spectres. Le paramètre *p*, commun à tous les spectres, est déterminé par une procédure de leave-one-out. On représente alors chaque spectre par ses *p-1+d* coordonnées sur la base de B-splines retenue et on applique la procédure de sélection de variables par IM à cette représentation. Comme les spectres sont en général très réguliers, *p-1+d* est petit devant *N*, ce qui réduit considérablement le temps de calcul nécessaire à la sélection des variables optimales. De plus, chaque B-spline est localisée ; une B-spline résume donc un intervalle de longueurs d'onde, ce qui facilite l'interprétation. Enfin, des variables spectrales correspondant à des longueurs d'onde très proches, et donc fortement corrélées, seront représentées par un seul coefficient B-spline.

4. Résultats expérimentaux

Nous illustrons la méthode proposée sur les données de la compétition logicielle de la conférence IDRC 98[1]. Il s'agit de 141 spectres en proche infrarouge (de 400nm à 2498nm, 1050 longueurs d'onde) correspondant à l'analyse d'échantillons d'herbes humides, soumises à divers niveaux de fertilisation. La variable à prédire est le niveau d'azote contenu dans chaque échantillon. Pour l'analyse, les 141 spectres ont été répartis en 36 spectres de test (ensemble indépendant grâce auquel on évalue les performances de la calibration) et 105 spectres d'apprentissage (trois groupes de 35 spectres pour effectuer une validation croisée). On mesure les performances des méthodes utilisées en termes de RMSE (*Root Mean Square Error*) sur l'ensemble de test. Pour toutes les méthodes, les spectres ont été préalablement centrés et réduits. Les informations produites par cette réduction (la moyenne et l'écart-type de chaque spectre) ont été conservées sous forme de deux variables additionnelles. Les performances de références sont données par la PCR et la PLSR appliquées aux spectres centrés et réduits (et au deux nouvelles variables). Le nombre de variables latentes dans ces méthodes est déterminé par validation croisée.

Grâce à la représentation par B-splines, chaque spectre décrit par 1052 variables spectrales est réduit à 155 coefficients de B-splines d'ordre 5, ce qui correspond à p=151 intervalles (p étant déterminé par leave-one-out). Chaque B-spline correspond donc grossièrement à 35 variables spectrales d'origine, puisqu'elle recouvre 5 intervalles. La procédure décrite dans la section 3 appliquée aux coordonnées B-splines conduit à la sélection de 10 variables (les variables moyenne et écart-type ne sont pas sélectionnées). On construit sur ces 10 variables un modèle linéaire (à titre de vérification) et un réseau de neurones de type RBFN (*Radial Basis Function Network*) dont les méta-paramètres sont déterminés par validation croisée. La Table 1 donne l'ensemble des résultats obtenus.

Méthode	***NMSE (Test)***
PCR (Régression en composantes principales, 10 composantes)	0.646
PLSR (Régression en moindres carrés partiels, 9 composantes)	0.633
B-splines + IM (10 variables) + modèle linéaire	0.829
B-splines + IM (10 variables) + RBFN	0.567

Table 1 : résultats expérimentaux

Les 10 variables obtenues correspondent à des plages de longueurs d'onde interprétables (la Figure 1 représente les 10 B-splines sélectionnées). En fait, en raison des chevauchements entre les B-splines, les 10 variables correspondent aux 5 intervalles de longueurs d'onde suivants : [402,440], [582,732], [956,1024], [2096,2164] et [2236,2386].

[1]URL : http://kerouac.pharm.uky.edu/asrg/cnirs/shoot_out_1998/shoot_out_1998.html

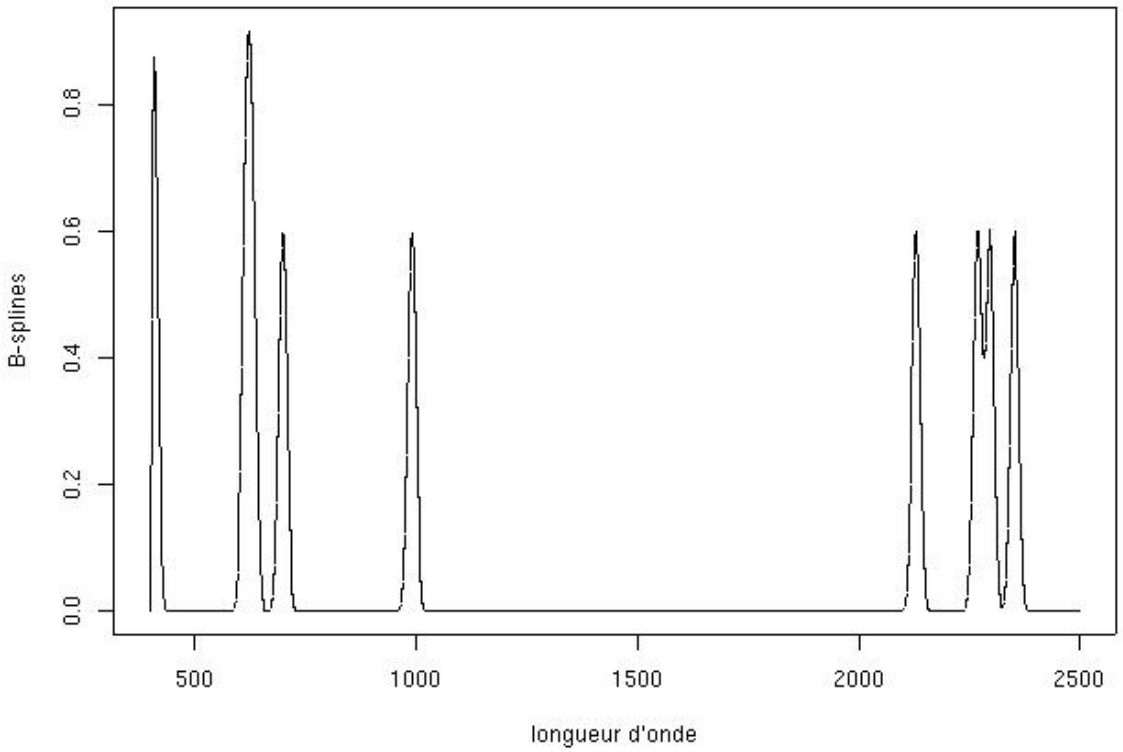

Figure 1 : B-splines sélectionnées ; certains se recouvrent, donnant lieu à 5 plages spectrales utiles au problème de prédiction.

5. Conclusion

La combinaison de la projection de spectres sur des B-splines, la sélection de ceux-ci par information mutuelle et la construction d'un modèle non-linéaire sur base des coefficients de B-splines sélectionnés, permet d'obtenir des performances de prédiction meilleures que dans le cas de modèles linéaires, tout en apportant la possibilité d'interpréter les plages spectrales obtenues. Les temps-calcul restent raisonnables sur l'exemple présenté, de l'ordre de l'heure y compris les optimisations de leave-one-out et l'optimisation des modèles non-linéaires. Cet article montre l'interprétabilité rendue possible grâce à la sélection de plages de longueur d'ondes ; la méthodologie utilisée est illustrée sur un problème typique de prédiction.

Remerciements

D. François bénéficie d'une bourse du Fond belge pour la Formation à la Recherche dans l'Industrie et l'Agriculture (FRIA). M. Verleysen est Maître de Recherches FNRS. Ce travail est en partie financé par le Pôle d'Attraction Interuniversitaire initié par le Gouvernement Fédéral Belge, Ministère des Sciences, Cultures et Technologies. La responsabilité scientifique incombe aux auteurs.

Bibliographie

[BER 00] Bertrand D., Dufour E., "La spectroscopie infrarouge et ses applications analytiques", Eds Tec& Doc, collection sciences et techniques agroalimentaires, (2000).

[BOR 78] De Boor C., « A practical guide to splines », Springer-Verlag, New York (1978).

[KRA 04] A. Kraskov, H. Stögbauer, P. Grassberger, "Estimating mutual information", Phys. Rev. E **69**, 066138 (2004).

[MAS 97] Massart D. L., Vandeginste B. G. M., Buydens L. M. C., De Jong S., Lewi P. J., Smeyers-Verbeke J., "Handbook of Chemometrics and Qualimetrics : Part A", Elsevier Science, Amsterdam, 1997.

[ROS05] Rossi F., Lendasse A., François D., Wertz V., Verleysen M., "Mutual information for the selection of relevant variables in spectrometric nonlinear modeling", Accepted for publication in Chemometrics and Intelligent Laboratory Systems, Elsevier.