Une approche orientée données pour la projection de variables spectrales en spectrométrie

Catherine **Krier**[1], Damien **François**[2], Michel **Verleysen**[1 *]
[1] *Université catholique de Louvain – Machine Learning Group, DICE, 3 place du Levant, 1348 Louvain-la-Neuve, Belgique,* {krier,verleysen}@dice.ucl.ac.be
[2] *Université catholique de Louvain – Machine Learning Group, CESAME, 4 av. G. Lemaître, 1348 Louvain-la-Neuve, Belgique,* francois@csam.ucl.ac.be

MOTS CLÉS : Spectroscopie, modèles non-linéaires multivariés, sélection de variables, analyse en composantes indépendantes, information mutuelle

1. Introduction

Beaucoup de problèmes en chimiométrie nécessitent la prédiction d'une variable quantitative à partir des variables spectrales mesurées. Il peut s'agir de prédire la concentration d'un composant dans un produit, de vérifier une conformité à une norme alimentaire, etc. La variable quantitative étant souvent difficile et/ou onéreuse à quantifier, une solution consiste à développer un modèle permettant de prédire cette variable à partir des spectres infrarouges (IR) mesurés en transmission ou réflexion sur le produit.

De très nombreux travaux ont été et sont encore actuellement réalisés afin de développer des modèles performants en termes de qualité de prédiction. La performance est bien évidemment un critère essentiel ; l'absence de hautes performances rend en effet tout modèle simplement inexploitable. Mais il ne s'agit pas de l'unique objectif. L'utilisateur des modèles, le chimiométricien ou le praticien, doit être capable d'interpréter le modèle afin d'en extraire l'information adéquate quant au problème étudié. Interpréter signifie, par exemple, identifier les longueurs d'onde ou plages de celles-ci qui contribuent à expliquer la variable quantitative, afin de vérifier, ou de découvrir, les composants qui en sont responsables.

Ce papier présente une approche permettant de construire un modèle prédictif tout en extrayant une information interprétable. La recherche de performances en termes de qualité de prédiction est couplée à une sélection drastique du nombre de composantes entrant en ligne de compte dans le modèle, rendant son interprétation aisée. La sélection de plages spectrales suit une projection préalable sur une base afin de réduire préalablement la dimension du problème ; l'originalité de l'approche présentée ici réside dans la construction d'une base de projection adaptée aux données, et non fixée a priori.

La suite de ce papier est organisée de la façon suivante. La Section 2 introduit la nécessité de combiner la projection et la sélection de variables dans des problèmes de spectrométrie, et montre certaines limitations des approches existantes. La Section 3 définit la construction du modèle prédictif, combinant projection, sélection et prédiction. La Section 4 montre les résultats de cette approche sur un problème classique de prédiction

[*] M. Verleysen est Directeur de Recherches du Fonds National de la Recherche Scientifique. Le travail de C. Krier et D. François est financé par une bourse FRIA. Une partie de cet article présente des résultats de recherche financée par le programme belge des Pôles d'Attraction Interuniversitaires, mis en place par les Services fédéraux des affaires Scientifiques, Techniques et Culturelles de l'Etat belge. La responsabilité scientifique appartient à ses auteurs.

à partir de spectres infrarouges.

2. Réduction du nombre de variables spectrales

De façon générale, les spectres sont considérés comme des objets de « haute dimension », ce qui signifie qu'ils sont chacun décrits par un grand nombre de variables (appelées variables spectrales dans la suite de ce texte). Il n'est pas rare d'avoir des centaines ou des milliers de variables spectrales, quel que soit le contexte. Ceci entraîne des difficultés importantes dans la construction de modèles prédictifs. Comme souvent le nombre de spectres disponibles pour construire le modèle est inférieur au nombre de variables spectrales, le problème tel quel est mal posé, quelque soit le modèle. Il faut alors recourir à des techniques de réduction du nombre de variables spectrales, ce que font tous les modèles linéaires comme la PCR et la PLSR, ainsi que les modèles non-linéaires.

La réduction du nombre de variables spectrales peut se faire par sélection (on en garde certaines, parmi les variables initiales), ou par projection (on en construit d'autres, à partir des variables initiales). D'un point de vue interprétabilité, la sélection est le moyen le plus adéquat. En effet, les variables résultant de la procédure sont des variables de départ, identifiées avec leur longueur d'onde, et donc que le chimiométricien peut relier à sa connaissance des constituants recherchés. Les méthodes de projection, y compris linéaires comme la PLSR, n'apportent en général pas cette facilité d'interprétation ; les axes de projection « utilisent » en effet l'ensemble de la plage spectrale des spectres, rendant impossible leur identification avec certains constituants. La situation s'empire encore lorsque des modèles a priori plus généraux et plus performants, comme des modèles non-linéaires de prédiction, sont utilisés.

La sélection de variables a néanmoins deux limites. D'une part, elle est moins générale que la projection (elle en est un cas particulier) ; on peut donc imaginer que des performances accrues pourraient être obtenues en utilisant des projections. D'autre part, elle est tout simplement impossible à mettre en pratique lorsque le nombre de variables est trop important, comme dans le cas de données spectrales ; la sélection passe en effet par l'évaluation d'un critère (mesurant l'apport d'information d'une variable ou d'un groupe de variables), et les critères adéquats (détaillés dans la Section 3) sont très difficiles à estimer lorsque le nombre de variables augmente.

La solution passe alors par une approche consistant en une projection préalable, suivie d'une sélection. Si la projection est construite de façon à conserver l'interprétabilité des variables, les avantages respectifs de la sélection et de la projection seront additionnés. Une projection préalable est en tout cas possible dans le cas où les spectres sont « lisses », ce qui est principalement le cas des spectres IR. Dans le cas de spectres formés de nombreux pics étroits, comme les spectres de masse par exemple, la sélection doit être abordée directement, avec les limitations précitées [Krier 06].

Dans le cas de spectres lisses, la procédure consiste donc à d'abord projeter les spectres sur une base préalablement choisie et à considérer que chaque spectre est défini par les coefficients de sa projection sur la base, puis à sélectionner le plus petit nombre possible de coefficients. Si la base elle-même est interprétable, c'est-à-dire si chacun de ses éléments couvre une plage limitée et identifiable de longueurs d'onde, chaque coefficient sera interprétable de la même façon, ce qui rencontre les objectifs précités.

3. Modèle de prédiction

La section précédente a montré l'intérêt de décomposer le problème de prédiction en trois opérations consécutives : une projection (conservant l'interprétabilité des variables), une sélection, et la prédiction proprement dite. Ces trois opérations sont

détaillées dans cette section.

3.1 Projection

Exploiter le caractère lisse des spectres revient à considérer le problème sous l'angle d'une branche relativement récente de l'analyse de données, à savoir l' « analyse de données fonctionnelles » [Ramsay-Silverman 97]. Les spectres peuvent en effet être vus comme des fonctions (échantillonnées aux longueurs d'onde correspondant aux variables spectrales), plutôt que comme des données (vecteurs) de très haute dimension. L'avantage de la première vue est que le caractère lisse des spectres est pris en compte dans la continuité des fonctions spectrales (et de leur dérivées), alors qu'il n'est tout simplement pas pris en compte dans la seconde vue, ce qui constitue indéniablement une perte d'information. En pratique, l'analyse de données fonctionnelles consiste à projeter les fonctions sur une base, et ensuite à travailler avec les coefficients de la projection plutôt qu'avec les fonctions elles-mêmes. La base étant connue (et commune à toutes les données), il y a en effet une équivalence entre une fonction et ses coefficients, aux erreurs de projection près. Si ces dernières sont suffisamment faibles (par exemple en choisissant un nombre suffisamment élevé d'éléments dans la base), l'équivalence est assurée.

Une manière traditionnelle de procéder en analyse de données fonctionnelles est de considérer une approche par splines. Les splines sont des fonctions polynomiales par morceaux, définies en coupant l'intervalle de définition des fonctions en sous-intervalles. Des polynômes sont alors définis sur un ensemble restreint d'intervalles consécutifs. Ces polynômes constituent la base et sont appelés B-splines. L'ordre des polynômes définit leur degré. Les fonctions sont lors vues comme des combinaisons linéaires des B-splines. La continuité de fonctions et de ses dérivées successives est garantie par un choix adéquat des bases de B-splines. Les splines sont utilisés en analyse de données spectrales, par exemple dans [Rossi 06] et [Durand 01].

Une limitation importante de l'approche par splines est que la base est constituée de fonctions choisies a priori, sans aucune information extraite des données spectrales elles-mêmes. On choisit les splines parce qu'ils sont lisses et ont les propriétés désirées de continuité, on choisit leur ordre en fonction de la connaissance que l'on a, dans certains cas, du problème (par exemple, on augmentera l'ordre des splines si on sait que l'information adéquate se retrouve dans la dérivée des spectres plutôt que dans leur valeur [Rossi 06]), mais les valeurs spectrales elles-mêmes ne sont pas utilisées (si ce n'est éventuellement dans le choix du nombre de splines, à travers une minimisation de l'erreur de projection). La situation est donc que des problèmes très différents, dans des contextes différents et avec des données différentes, sont résolus en utilisant la même base, ce qui n'est certainement pas optimal, ni au point de vue des performances, ni au point de vue de l'interprétabilité. Par exemple, le nombre de B-splines ayant été fixé, chacun occupera une plage de valeurs spectrales de longueur égale, aussi bien dans les parties plates du spectre que dans celles où on retrouve des pics plus pointus, correspondant à des constituants particuliers. On aimerait pouvoir adapter la base en fonction des données, en définissant les éléments de la base comme étant les constituants que l'on recherche.

Cette question apparemment complexe peut être abordée par les techniques d'analyse en composantes indépendantes (ICA). L'ICA consiste, à partir de la connaissance de mélanges de fonctions (on parle de sources), à retrouver celles-ci, sans aucune connaissance si sur les sources ni sur le mélange. Afin de résoudre ce problème, l'ICA fait bien entendu un certain nombre d'hypothèses. Les plus fréquentes sont la linéarité du mélange, et l'indépendance des sources. Dans le cas des spectres, les sources considérées sont les spectres des constituants (inconnus). On peut considérer que les constituants sont indépendants entre eux, ce qui satisfait la seconde hypothèse.

L'hypothèse de linéarité est plus difficile à valider, mais n'est pas plus contraignante que la linéarité supposée dans beaucoup d'autres modèles, comme la PLSR. Les sources ainsi identifiées par l'ICA sont des spectres, interprétables, et peuvent servir de base à la place des B-splines. Elles n'ont plus les contraintes de ces derniers, et surtout sont déduits directement des données, donc sont mieux adaptées. De plus amples détails sur les nombreuses techniques d'ICA peuvent par exemple être consultés dans [Oja 01]. Une fois la base identifiée par ICA, les spectres son projetés sur les éléments de la base et remplacés par leurs coefficients, exactement comme dans le cas des splines.

3.2 Sélection

Tout comme le nombre de splines, le nombre de composantes indépendantes recherchées par l'ICA est un paramètre à ajuster. Bien qu'il ne s'agisse pas d'un ajustement très critique, il ne fait néanmoins pas descendre en-dessous d'un nombre trop faible, au risque d'une mauvaise projection des spectres sur la base. Un nombre élevé signifie néanmoins l'impossibilité de construire de manière efficace des modèles de prédiction directement sur les coefficients. Une étape de sélection (des coefficients résultant de l'ICA) est alors nécessaire. L'objectif étant de pouvoir construire des modèles non-linéaires de prédiction, la sélection se doit d'être non-linéaire également. La corrélation (de chaque coefficient avec la quantité à prédire), critère purement linéaire, n'est donc pas un critère de sélection adéquat. Elle doit être remplacée par l'information mutuelle, critère non-linéaire, qui a comme second avantage d'être extensible aisément à des groupes de variables. L'information mutuelle entre deux variables X et Y mesure la quantité d'information sur Y apportée par la connaissance de la variable X, par rapport à une situation où X ne serait pas connue, et vice-versa. Il s'agit d'un concept dérivé de la théorie de l'information et appliqué depuis quelques années à la sélection de variables en général [Battiti 94]. Son utilisation en spectrométrie est illustrée par exemple dans [Rossi 06b].

Le critère (l'information mutuelle) est utilisé pour sélectionner les variables (dans notre cas les coefficients de l'ICA) suivant une procédure de sélection. La procédure utilisée ici est une procédure « forward ». La première étape consiste à sélectionner la variable ayant l'information maximale avec la valeur à prédire. La seconde étape consiste à sélectionner une autre variable spectrale qui, combinée avec la première, apporte l'information mutuelle maximale avec la valeur à prédire, et ainsi de suite. En tout, N variables sont ainsi sélectionnées, en s'arrêtant lorsque l'information mutuelle n'augmente plus ; ensuite, comme la procédure forward ne remet jamais en cause le choix d'une variable et peut donc conduire à un ensemble de taille surestimée, chacune des 2^N combinaisons possibles des N variables est testée. Le critère utilisé pour comparer les 2^N cas possibles peut à nouveau être l'information mutuelle, ou mieux, un critère de type wrapper, c'est-à-dire l'erreur de prédiction du modèle elle-même. Si le nombre 2^N reste raisonnable, il est envisageable de construire 2^N modèles et de comparer leurs erreurs de prédictions, ce qui aurait bien entendu été irréalisable sur toutes les combinaisons possibles des variables spectrales initiales.

3.3 Prédiction

Les variables étant maintenant sélectionnées, la prédiction consiste à construire un modèle non-linéaire sur celles-ci. Un réseau RNFN (Radial-Basis Function Networks) est choisi, mais tout autre modèle non-linéaire pourrait être utilisé également. Les réseaux RBFN offrent de manière générale un bon compromis entre performances de prédiction et complexité-calcul. Ils comportent deux paramètres, le nombre d'unités et un facteur d'échelle (WSF) ; une procédure d'apprentissage des réseaux RBFN peut être trouvée dans

[Benoudjit 03].

4. Résultats

La méthodologie détaillée dans la section précédente est expérimentée sur un problème traditionnel d'analyse spectrale, la base de données Tecator [Tec]. Le problème consiste à prédire le taux de graisse dans des échantillons de viande, à partir de leurs spectres mesures en NIR, entre 850 et 1050 nm. Au total, 215 spectres sont disponibles ; 160 sont choisis aléatoirement comme base d'apprentissage, le reste étant réservé pour le test. L'ICA est réalisée sur les 160 spectres d'apprentissage (préalablement centrés et réduits), et 12 composantes indépendantes sont extraites. Elles sont représentées à la Figure 1 ; comme on peut le voir, les composantes indépendantes sont fort localisées, ce qui semble correspondre à une identification de constituants.

Ensuite les 12 coefficients de la projection de chaque spectre sur la base de sources résultant de l'ICA, constituent les entrées de la procédure de sélection par information mutuelle. Six coefficients ($N = 6$) sont sélectionnés avant la recherche exhaustive.

Trois choix doivent encore être définis : d'une part les deux méta-paramètres des modèles RBFN (le nombre d'unités et le facteur d'échelle, par recherche exhaustive dans une plage de valeurs), et d'autre part les variables qui seront effectivement utilisées parmi les 6 sélectionnées (recherche exhaustive parmi les 2^6 possibilités). Pour ce faire, une procédure de validation croisée est utilisée. L'ensemble de 160 spectres d'apprentissage est divisé aléatoirement en quatre parties égales. Chaque apprentissage est répété alors quatre fois, en utilisant trois des quatre parties comme ensemble d'apprentissage et la quatrième comme ensemble de validation ; la moyenne des quatre expériences est prise en compte. L'ensemble de test de 55 spectres n'est évidemment pas utilisé à ce niveau. Cette procédure résulte en la sélection de deux coefficients (correspondant aux 7 et 8ème composantes indépendantes), et des nombres d'unités et de facteur d'échelle respectivement égaux à 28 et 8.

La validité et le bien-fondé de la procédure doivent être vérifiés à deux niveaux : d'une part les performances en termes d'erreur de prédiction, et d'autre part l'interprétabilité des composantes indépendantes sélectionnées. En ce qui concerne le premier point, les erreurs de prédiction (NMSE, Normalized Mean Square Error) sur les ensembles de validation et de test sont respectivement de 0.0061 et 0.0255. Ceci est tout à fait comparable avec les performances d'autres modèles publiés récemment (voir par exemple [Rossi 06b]) sur la même base de données. La reconstruction des spectres des données Tecator (représentés à la Figure 2) sur base des deux composantes indépendantes sélectionnées est illustrée par la Figure 3. Comme le montre cette figure, les longueurs d'onde apportant l'information utile à la construction du modèle sont comprises dans un intervalle autour de 940 nm, où la reconstruction des spectres est non nulle, ceci correspond, par exemple, aux variables spectrales sélectionnées dans [Rossi 06b].

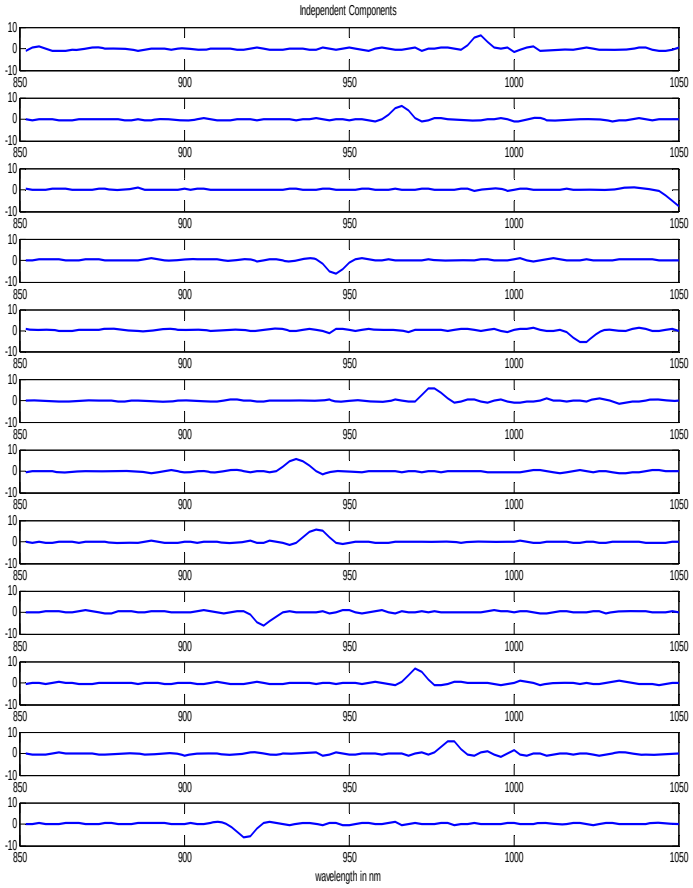

Figure 1: Composantes indépendantes extraites par ICA.

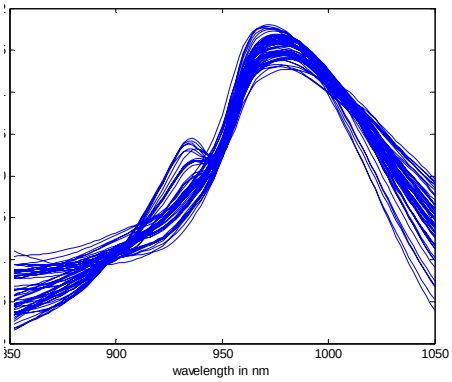

Figure 2: Spectres de la base de données.

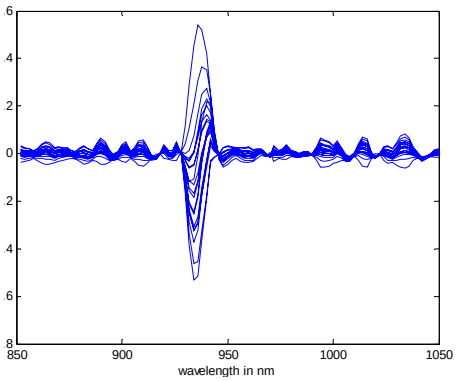

Figure 3: Spectres reconstruits sur base des composantes indépendantes sélectionnée.

5. Conclusion

L'utilisation de l'information mutuelle, couplée à une projection préalable des spectres sur une base de composantes indépendantes déterminées par ICA, conduit à la construction de modèles de prédiction dont les performances sont similaires à celles référencées dans la littérature. Par ailleurs, le caractère local des composantes indépendantes sélectionnées permet une interprétation aisée du modèle en termes de plages de longueurs d'onde.

Cette approche a pour avantage d'exploiter le caractère lisse des spectres par le biais d'une projection sur une base de fonctions lisses. De plus, la base utilisée ici est adaptée aux données, car construite au départ de celles-ci, et non choisie a priori comme dans le cas de splines, par exemple.

Références

[Krier 06] Krier C., François D., Wertz V.,Verleysen M., "Feature Scoring by Mutual Information for Classification of Mass Spectra", FLINS 2006, 7th International FLINS Conference on Applied Artificial Intelligence, pp. 557-564, Genova (2006).

[Ramsay-Silverman 97] Ramsay J.O., Silverman B.W., "Functional Data Analysis", Springer-Verlag, New York, NY, 1997.

[Rossi 06] Rossi F., François D., Wertz V., Verleysen M., "Fast Selection of Spectral Variables with B-Spline Compression", Chemometrics and Intelligent Laboratory Systems, Elsevier, in press.

[Durand 01] Durand J.-F., "Local polynomial additive regression through PLS and splines: PLSS", Chemometrics and Intelligent Laboratory Systems, Vol. 58, Issue 2, 2001, pp. 235-246.

[Oja 01] A. Hyvarinen, J. Karhunen, and E. Oja., "Independent Component Analysis". J. Wiley, 2001. 49

[Battiti 94] Battiti R., "Using the mutual information for selecting features in supervised neural net learning", IEEE Transactions on Neural Networks, Vol. 5, pp. 537-550 (1994).

[Rossi 06b] Rossi F., Lendasse A., François D., Wertz V., Verleysen M., "Mutual information for the selection of relevant variables in spectrometric nonlinear modeling", Chemometrics and Intelligent Laboratory Systems, Elsevier, Vol. 80, No. 2 (February 2006), pp. 215-226.

[Benoudjit 03] Benoudjit N., Verleysen M., "On the Kernel Widths in Radial-Basis Function Networks", Neural Processing Letters, Vol. 18, pp. 139-154 (2003).

[Tec] Tecator meat sample dataset. Available on statlib :
http:// lib.stat.cmu.edu/datasets/tecator.