IDC: amélioration de l'étalonnage direct, application à la quantification de l'éthanol dans les moûts en fermentation

Jean Claude J. C. Boulet, Jean-Michel Roger

IDC: amélioration de l'étalonnage direct, application à la quantification de l'éthanol dans des moûts en fermentation

M. BOULET Jean-Claude[1], M. ROGER Jean-Michel[2]

[1] INRA-UMR-SPO, F-34060 Montpellier (bouletjc@supagro.inra.fr)

[2] CEMAGREF-UMR-ITAP, F-34196 Montpellier (jean-michel.roger@montpellier.cemagref.fr)

Mots-clef: direct, calibration, DC, SBC, IDC.

1 Introduction

En spectroscopie, deux types de connaissances peuvent être utilisées pour construire un étalonnage : les connaissances expertes, telles que les spectres purs des composés, et les connaissances expérimentales, comme les bases d'étalonnage. Les méthodes de régression utilisent des connaissances expérimentales, mais se privent des connaissances expertes. A l'opposé, les méthodes d'étalonnage direct utilisent la connaissance experte mais se passent des connaissances expérimentales. Quelques méthodes, comme la Multi Curve Resolution (MCR, [1]) ou la Science Based Calibration (SBC, [2]) utilisent avec succès les deux types de connaissances. Ce texte propose une nouvelle méthode d'étalonnage, inspirée de la SBC, appelée IDC (Improved Direct Calibration), qui utilise à la fois tous les spectres purs disponibles et les connaissances expérimentales exprimant les influences.

2 Théorie

Soient $\mathbf{X}$ une matrice de N spectres et P variables, et $\mathbf{y}$ les valeurs correspondantes de la grandeur d'intérêt. La prédiction $\hat{\mathbf{y}}$ de $\mathbf{y}$ obtenue à partir de $\mathbf{X}$ peut être calculée avec un modèle linéaire d'étalonnage utilisant un vecteur $\mathbf{b}$ de b-coefficients:

$$\hat{\mathbf{y}}=\mathbf{Xb}$$

Les méthodes d'étalonnage diffèrent selon la manière dont $\mathbf{b}$ est estimé. Nous ne considérerons ici que les méthodes linéaires d'étalonnage direct.

La Calibration Directe, ou DC (Martens et Naes, [3]) est basée sur la loi des mélanges issue de la loi de Beer-Lambert. Soient $\mathbf{K}$ la matrice contenant les spectres purs de tous les composés présents, à l'exception de celui, $\mathbf{k}$, de la grandeur d'intérêt, et $\mathbf{\Sigma}_{DC}$ le projecteur orthogonal à $\mathbf{K}$. On a :

$$\hat{\mathbf{y}}_{DC} = \mathbf{X}\,\mathbf{\Sigma}_{DC}\,\mathbf{k}\,(\mathbf{k}'\,\mathbf{\Sigma}_{DC}\,\mathbf{k})^{-1}$$

La DC est rarement utilisée car elle ne tient pas compte des bruits causés par les diverses grandeurs d'influence physique et elle suppose de connaître les spectres purs de tous les composés présents.

La Science Based Calibration (SBC), proposée par Marbach [2], modifie la DC en pondérant les directions de l'espace, d'une manière inversement proportionnelle au bruit. Ce bruit est recueilli par un plan d'expériences fournissant une matrice de spectres $\mathbf{X}_G$ où les variations sont dues uniquement aux grandeurs d'influence. Notons $\mathbf{\Sigma}_{SBC}= (\mathbf{X}'_G\mathbf{X}_G)^{-1}$. De manière analogue à la DC:

$$\hat{\mathbf{y}}_{SBC}=\mathbf{X}\,\mathbf{\Sigma}_{SBC}\mathbf{k}(\mathbf{k}'\,\mathbf{\Sigma}_{SBC}\mathbf{k})^{-1}$$

La SBC permet de prendre en compte des influences diverses, telles qu'elles s'expriment dans $\mathbf{X}_G$. Par contre, elle ne tient pas compte explicitement des spectres purs des autres composés.

De manière complémentaire à la DC et à la SBC, l'IDC propose de réaliser un étalonnage direct, en tenant compte à la fois des spectres purs disponibles et de l'information expérimentale sur le bruit. Soit $\mathbf{K}$ la matrice des spectres purs disponibles, autres que celui de la grandeur d'intérêt. Soit $\mathbf{P}$ de dimension (A,P) la matrice regroupant les A premiers vecteurs propres de l'ACP sur $\mathbf{X}_G$. $\mathbf{K}$ et $\mathbf{P}$ sont

responsables des influences qui s'ajoutent aux effets du produit d'intérêt, comme schématisé par l'équation suivante :

$$\mathbf{X}=\mathbf{y}\mathbf{k}' + \mathbf{T}_\chi \mathbf{K} + \mathbf{T}_\psi \mathbf{P} \quad (1)$$

Soit **R** la matrice construite en concaténant **K** et **P**, et $\mathbf{\Sigma}_{IDC}$ le projecteur orthogonal à **R**. Par construction, $\mathbf{K}\mathbf{\Sigma}_{IDC} = \mathbf{P}\mathbf{\Sigma}_{IDC}=\mathbf{0}$. En multipliant l'équation (1) à droite par $\mathbf{\Sigma}_{IDC}\,\mathbf{k}(\mathbf{k}'\,\mathbf{\Sigma}_{IDC}\mathbf{k})^{-1}$, on obtient :

$$\hat{\mathbf{y}}_{IDC}=\mathbf{X}\,\mathbf{\Sigma}_{IDC}\mathbf{k}(\mathbf{k}'\,\mathbf{\Sigma}_{IDC}\mathbf{k})^{-1}$$

donnant le modèle :

$$\mathbf{b}_{IDC}=\mathbf{\Sigma}_{IDC}\mathbf{k}(\mathbf{k}'\,\mathbf{\Sigma}_{IDC}\mathbf{k})^{-1}$$

En procédant par projection orthogonale, et en incluant les spectres purs et l'espace engendré par le bruit expérimental, l'IDC combine les avantages de la DC et de la SBC.

3 Matériels et méthodes

Deux applications de l'IDC seront présentées: la quantification de l'éthanol dans des moûts et vins en cours de fermentation alcoolique, à partir des spectres acquis entre 500 et 2500nm, et la quantification des protéines totales dans des échantillons de blé broyé à partir des spectres entre 1300 et 2400nm (Challenge-Chimiométrie 2007). Seule la première application est détaillée ici.

2.1. Données

Les données expertes étaient constituées des spectres purs $\mathbf{k}_{ethanol}$, $\mathbf{k}_{glycerol}$, $\mathbf{k}_{lactate}$, et $\mathbf{k}_{eau}$ acquis en référence à l'air. Les données expérimentales étaient constituées des spectres de 1480 échantillons de vins et de moûts en cours de fermentation, acquis en transmission entre 500 et 2500 nm, en référence à l'eau. Ces données ont été réparties en trois jeux :

- les 165 spectres des moûts non fermentés (y=0) ont été rassemblés dans $\mathbf{X}_G$, pour représenter la variabilité indépendante de y ;
- les 315 premiers individus de la base ont été rassemblés dans ($\mathbf{X}_{ETAL}$, $\mathbf{y}_{ETAL}$), pour étalonner une PLSR
- le reste de la base (1000 individus) a constitué ($\mathbf{X}_{TEST}$, $\mathbf{y}_{TEST}$)

2.2. Calcul du modèle IDC

- Détermination de A : Plusieurs modèles IDC ont été calculés avec des valeurs de A allant de 1 à 20. La valeur de A retenue est celle qui minimise l'erreur de prédiction sur $\mathbf{X}_G$ (prédiction qui devrait être nulle).
- Détermination de **k** et **K**: **k** est $\mathbf{k}_{ethanol}$, le spectre de la grandeur d'intérêt. La matrice **K** est composée des spectres $\mathbf{k}_{glycerol}$, $\mathbf{k}_{lactate}$, et $\mathbf{k}_{eau}$. Les deux premiers ont été choisis car le glycérol et l'acide lactique sont des sous-produits des fermentations alcooliques et malolactiques, ils ne sont pas représentés dans $\mathbf{X}_G$. Le troisième permettra de compenser le fait que les spectres de moûts sont acquis en référence à l'eau. En effet, comme $\mathbf{k}_{eau}$ est introduit dans **R**, tous les spectres du jeu de test sont projetés orthogonalement au spectre de l'eau, ce qui conduit à supprimer de la prédiction tout effet lié au choix de la référence, air ou eau.

2.3. Calcul du modèle PLSR

Le modèle PLSR a été construit à partir du jeu ($\mathbf{X}_{ETAL}$, $\mathbf{y}_{ETAL}$). Le nombre de variables latentes est celui qui minimise l'erreur standard de validation croisée sur ($\mathbf{X}_{ETAL}$, $\mathbf{y}_{ETAL}$).

2.4. Test des modèles IDC et PLSR

Les modèles IDC et PLSR ont été appliqués sur le jeu ($\mathbf{X}_{TEST}$, $\mathbf{y}_{TEST}$), et leurs erreurs standard de prédiction comparées.

4 Résultats.

L'IDC a été calculée avec A=4 (Fig.1), ce qui correspond à 99,91% de la variance de $\mathbf{X}_G$. La PLSR a été calculée avec 5 variables latentes. Les erreurs standard de prédiction (RMSEP) sont

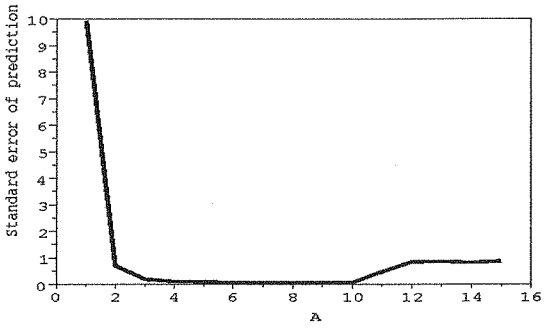

Fig.1: Modèles IDC: erreurs de prédiction sur $\mathbf{X}_{2a}$ en fonction du nombre A de vecteurs propres de l'ACP sur $\mathbf{X}_G$

respectivement pour l'IDC et la PLSR: de 0,87° et 0,90° pour les teneurs en éthanol inférieures à 10°; de 1,01° et 0,92° pour les teneurs en éthanol supérieures à 10°, et de 0,96° et 0,92° sur l'ensemble du jeu de test. Ainsi, l'IDC est très légèrement plus performante que la PLSR pour les teneurs en éthanol inférieures à 10°, ce rapport s'inverse toutefois pour les teneurs en éthanol supérieures à 10°, ainsi que sur l'ensemble du jeu de test. Ces deux modèles ont donc des performances comparables.

5 Discussion et Conclusion

L'IDC est une amélioration de la DC consistant à prendre en compte les grandeurs d'influence non représentées par un spectre pur. Correctement paramétrée, ses performances peuvent être comparables à celles de la PLSR. Cela est rendu possible car l'IDC utilise simultanément des connaissances expertes et des connaissances expérimentales se complétant mutuellement pour caractériser complètement les grandeurs d'influence. L'IDC a ainsi une grande souplesse d'utilisation dans la caractérisation des grandeurs d'influence, quelle que soit leur origine. De manière générale, toute grandeur d'influence non prise en compte dans $\mathbf{R}$ et ayant un effet identique sur chaque spectre pourra se traduire par une modification de pente et/ou biais. Par exemple, la non prise en compte du spectre de l'eau dans le cas présent donne des prédictions beaucoup plus bruitées, le long d'une droite de pente différente de 1 (résultat non présenté).

L'IDC est un modèle basé sur le concept de NAS-Net Analyte Signal (Lorber,[5]). En effet, le terme $\mathbf{\Sigma}_{IDC}\mathbf{k}$ inclus dans $\mathbf{b}_{IDC}$ correspond à la définition du NAS selon Lorber: « *the net analyte signal may be computed as the part of its spectrum orthogonal to the contribution of other coexisting constituants* », à la différence près que l'IDC étend cette définition aux grandeurs d'influence physiques. L'obtention d'un modèle IDC performant indique que toutes les grandeurs d'influence ont bien été prises en compte. En conclusion, à la suite de la SBC, l'IDC montre que les méthodes d'étalonnage direct ont un réel potentiel de prédiction, particulièrement intéressant lorsque l'obtention des jeux d'étalonnage est difficile ou impossible.

6 Références

[1] A.DeJuan, R.Tauler. Comparison of three-way resolution methods for non-linear chemical datasets. *J. of Chemometrics*, 15:759-772, 2001.

[2] R.Marbach. A new method for multivariate calibration. *J. Near Infrared Spectroscopy*, 13:241-254, 2005.

[3] H.Martens et T.Naes. *Multivariate Calibration*. Wiley, 1989.

[4] J.M.Roger, F.Chauchard et V.Bellon-Maurel. EPO-PLS. External parameter orthogonalisation of PLS. Application to temperature-independant measurement of sugar contents in fruits. *Chem.Intell. Lab. Systems*, 66:191-204, 2003.

[5] A.Lorber, K.Faber et B.R.Kowalski. Net analyte signal calculation in multivariate calibration. *Analytical Chemistry*, 69:1620-1626, 1997