La robustesse des étalonnages multidimensionnels, application aux données spectrales

J.M. Roger

Author-produced version of the article published in Techniques de l'Ingénieur, 2010, SL 265, 11 p.

La robustesse des étalonnages multidimensionnels

Application aux données spectrales

par **Jean-Michel ROGER**

Ingénieur en Chef du Génie Rural des Eaux et des Forêts

Chercheur

Résumé L'étalonnage des appareils délivrant des signaux continus, tels que des spectres, pose des problèmes particulier de robustesse, car la grande dimension de l'espace de mesure offre beaucoup de possibilités aux grandeurs de s'exprimer. Une stratégie d'appréhension de ce problème est donc nécessaire. Elle repose sur un questionnement hiérarchisé sur la significativité de la grandeur d'influence, sa contrôlabilité et sa mesurabilité. Dans le cas d'une grandeur influente, non contrôlable mais mesurable, il est possible de corriger de trois manières : soit en modifiant le signal mesuré, soit en modifiant le modèle, soit en corriger sa sortie. Dans le cas le plus défavorable, où la grandeur est influente, mais ni contrôlable, ni mesurable, des méthodes d'amélioration de la robustesse des étalonnages sont nécessaires. Des méthodes récentes permettent de corriger les influences en corrigeant l'espace de mesure par projection orthogonale au sous espace touché par la grandeur d'influence.

Abstract. The calibration of the devices which produce continuous signals, like spectra, presents a specific problem of robustness. This issue is related to the huge dimension of the measurement space, which increases the possibility for the influence factors to affect the signal. Then, this problem must be managed with a dedicated strategy, based on a hierarchical approach, answering the successive questions: Is the factor influent? Can the factor be controlled? Can the factor be measured? In the case where the influent factor can not be controlled but can be measured, three main ways of correction are possible, by modifying: the signal, the model or the estimation. In the most unfavourable case, where the influent factor can neither be controlled nor be measured, it is necessary to use some methods dedicated to the robustness improvement. For example, some recent methods based on orthogonal projection, allows the user to correct for the influences, by subtracting from the measurement space the subspace which is impacted by the influence factor.

Mots-clés robustesse, étalonnage multi-varié, espaces de grande dimension, spectrométrie.

Keywords robustness, multivariate calibration, high dimension space, spectrometry

Table analytique

Notations

Les lettres majuscules grasses seront employées pour désigner des matrices, p.e. $\mathbf{X}$; les lettres minuscules grasses désignent des vecteurs colonnes, p.e. $\mathbf{x}_j$ désigne la $j^{ème}$ colonne de $\mathbf{X}$ (les vecteurs sont toujours supposés disposés en colonne) ; les lettres minuscules non grasses désignent des scalaires, p.ex. des éléments de matrice x_{ij} ou des indices i. En cas de besoin, la dimension des matrices peut être indiquée par un double indiçage entre parenthèses, p.ex. $\mathbf{X}_{(np)}$ indique que la matrice $\mathbf{X}$ a n lignes et p colonnes. Le produit scalaire de deux vecteurs $\mathbf{a}$ et $\mathbf{b}$ est noté $\mathbf{a}^T\mathbf{b}$. Les grandeurs sont indiquées par une lettre capitale italique, par exemple la température T. Dans la suite de cet article, la grandeur d'intérêt (visée par l'étalonnage) sera notée Y, sa valeur y. Une grandeur d'influence sera notée G, sa valeur g. Les valeurs estimées par un modèle seront notées avec un chapeau, par

exemple $\hat{y}$.

Contexte et problématique

1.1 L'étalonnage des signaux multi-variés

De plus en plus d'appareils analytiques délivrent des signaux continus (des courbes) qui sont digitalisés sous forme d'un vecteur de nombres. C'est le cas notamment des spectromètres, qui produisent des spectres d'absorption, de réflectance, de masse, etc. Cependant, ces appareils sont généralement utilisés pour accéder à des grandeurs (dites d'intérêts) qui sont en relation avec la forme ou le niveau des courbes mesurées. Ainsi, le chimiste analyste n'utilise pas un spectromètre pour mesurer un spectre, mais pour estimer la concentration d'un composé du produit mesuré, c'est-à-dire la grandeur d'intérêt. Il est donc nécessaire d'étalonner une relation entre le signal acquis et la grandeur d'intérêt, dont la valeur est recherchée.

L'étalonnage d'un appareil de mesure consiste à établir une relation entre les grandeurs primaires, effectivement mesurées, et la grandeur d'intérêt, dont la valeur est recherchée. C'est une opération assez bien maîtrisée dans le cas où le nombre de grandeurs primaires est petit. Par exemple, l'étalonnage d'un thermomètre à thermocouple nécessite l'établissement d'une relation entre une seule grandeur primaire (la tension de jonction) et la grandeur d'intérêt (la température). Dans ce cas, la dimension de l'espace de mesure est unitaire. Dans le cas de la spectrométrie, cette dimension est beaucoup plus importante. Ainsi, les spectromètres infrarouge à transformée de Fourier (IRTF) mesurent des absorptions sur plus d'un millier de nombres d'onde ; dans le domaine de la Résonance Magnétique Nucléaire (RMN), ce sont des vecteurs de plusieurs centaines de milliers de points qui sont produits. L'espace de mesure devient alors impossible à appréhender dans sa globalité. Ce problème, dit « fléau de la dimensionnalité », peut être illustré par l'exemple suivant :

Soit un capteur de température utilisant un thermocouple, dont le signal de la grandeur primaire est digitalisé sur 8 bits. La relation d'étalonnage aura pour fonction de mettre en relation les $2^8 = 256$ valeurs possibles de la tension de jonction avec le même nombre de valeurs de température.

Soit un capteur de concentration en sucre utilisant un spectromètre délivrant des spectres d'absorption sur 256 longueurs d'ondes, chacune d'elle étant digitalisée sur 8 bits. La relation d'étalonnage aura pour fonction de mettre en relation les $2^{256x256} = 2^{65536} \approx 2\ 10^{19728}$ valeurs possibles de spectres mesurés avec les différentes valeurs de concentration en sucre.

Il est donc clair que les techniques statistiques classiques d'étalonnage ne peuvent pas être appliquées sans précaution au cas des signaux de grande dimension, comme les spectres. C'est pour cela que des méthodes dédiées ont été développées, au sein de la chimiométrie, comme détaillé dans [DB1]. La plupart de ces méthodes s'appuient sur l'identification du sous espace porteur de l'information utile, que nous appellerons *espace latent*, en référence aux variables latentes de la méthode d'étalonnage la plus populaire, la régression Partial Least-Squares ou PLSR (cf [DB1], [HM1]). Il s'agit du sous espace engendré par les variations des spectres reliées à celles de la grandeur d'intérêt. Cet espace est d'une dimension très inférieure à celle de l'espace de mesure.

Dans la suite de cet article, nous supposerons que le signal mesuré est un spectre $\mathbf{x}$ de dimension p (par exemple un spectre d'absorption infrarouge), que l'étalonnage concerne une grandeur d'intérêt unique Y (par exemple une concentration) et qu'il est réalisé par une méthode linéaire. Ainsi, le modèle d'étalonnage est constitué d'un vecteur $\mathbf{b}$ de p coefficients et d'une ordonnée à l'origine b_0, tels que :

$$\hat{y} = \mathbf{x}^T\mathbf{b} + b_0 \qquad (1)$$

1.2 Le problème de la robustesse

En métrologie, la qualité d'une méthode ou d'un instrument est habituellement caractérisée en utilisant différentes notions comme : la reproductibilité, la répétabilité, l'exactitude, la fidélité, la justesse ou l'incertitude. La notion de robustesse n'y apparaît pas. Elle revêt des significations différentes selon le domaine d'application, mais est toujours vue comme une qualité essentielle.

Pour une méthode d'analyse, on peut trouver une définition dans fascicule de documentation FD V01-000, et rappelée dans [MD1] :

> la robustesse est l'aptitude d'une méthode d'analyse à fournir de faibles variations du résultat lorsqu'elle est soumise à des modifications contrôlées des conditions d'application (exemple : température ambiante, lumière, pression atmosphérique, humidité, réactifs, appareillage, etc.).

Dans le domaine des étalonnages multi-variés, une définition récente proposée dans [MZ1], résume le sens communément donné à cette qualité :

> la robustesse d'un modèle d'étalonnage multivarié est la stabilité de sa capacité prédictive vis-à-vis des perturbations appliquées au voisinage des conditions d'étalonnage.

D'un point de vue formel, le problème de la robustesse peut s'exprimer de la manière suivante :

Soit G une grandeur d'influence, dont la variation de niveau δg autour des conditions d'étalonnage, entraîne une variation de spectre $\delta\mathbf{x}$. L'effet sur le modèle de cette variation est directement donné par $\delta\omega = \delta\mathbf{x}^{\mathbf{T}}\mathbf{b}$. La robustesse du modèle d'étalonnage est qualifiée par son insensibilité aux variations de *G*, c'est à dire la taille de l'erreur qui doit être petite :

$$|\delta\omega| = |\delta\mathbf{x}^{\mathbf{T}}\mathbf{b}| \qquad (2)$$

1.3 Exemples

Deux exemples issus de la spectrométrie visible et proche infrarouge seront utilisés dans cet article. Le premier concerne une grandeur d'influence continue, alors que le deuxième illustre le cas d'une grandeur discrète.

1.3.1 Exemple 1 : Effet de la température sur la mesure du taux de sucres des pommes par spectrométrie visible / très proche infrarouge

Cet exemple est décrit en détail dans [JR1]. Les spectres visible / très proche infrarouge de pommes Golden entières ont été acquis à l'aide d'un spectromètre Zeiss MMS1, en rétrodiffusion, sur la gamme spectrale 310 nm – 1050 nm. Les spectres ont été digitalisés sur p =256 longueurs d'onde, régulièrement espacées d'environ 3,3 nm. Le taux de sucres (essentiellement fructose et glucose) des pommes étaient mesurés par réfractométrie sur quelques gouttes de jus extraites par pression au voisinage de la zone de mesure du spectre.

Les spectres et les taux de sucres d'un premier ensemble E_0 de n_0=80 pommes ont été mesurés à température ambiante et stockés respectivement dans $\mathbf{X}_0$ et $\mathbf{y}_0$.

Un deuxième jeu E_1 de n_1=10 pommes de la même variété ont été immergées dans un bain marie dont la température a été portée à 8 niveaux $\{t^1, t^2, \ldots, t^8\}$ régulièrement espacés de t^1=5°C à t^8=40°C. Pour chacune des températures, les spectres des dix pommes ont été mesurés, puis stockés dans les matrices $\{\mathbf{X_1}^1, \mathbf{X_1}^2, \ldots, \mathbf{X_1}^8\}$. Le taux de sucres de ces 10 pommes a été mesuré et stocké dans $\mathbf{y}_1$.

Un troisième jeu E_2 de n_2=10 pommes a été soumis au même dispositif expérimental que le jeu E_1, fournissant les matrices $\{\mathbf{X_2}^1, \mathbf{X_2}^2, \ldots, \mathbf{X_2}^8\}$ et le vecteur $\mathbf{y_2}$.

Le jeu E_0 a été utilisé pour construire un modèle par régression PLS, sans prétraitement ni sélection de variables. La figure 1 gauche montre l'évolution de l'erreur d'étalonnage (SEC pour standard error of calibration) et de validation croisée (SECV pour standard error of cross validation) en fonction de la dimension de l'espace latent. Cette figure indique qu'une dimension de k =10 est tout à fait acceptable. La figure 1 droite montre les prédictions en validation croisée pour un espace latent de dimension k =10.

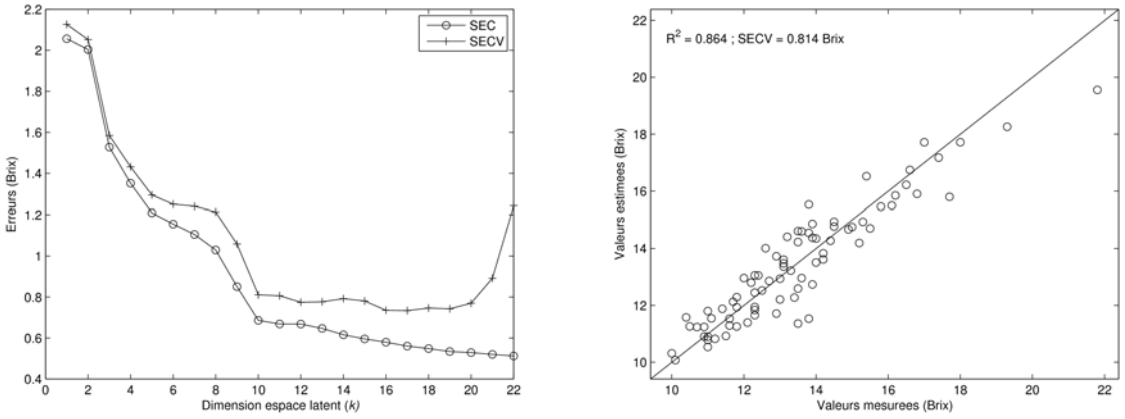


Figure 1 - Étalonnage du modèle de mesure du taux de sucres des pommes. À gauche, évolution des erreurs en fonction du nombre de variables latentes de la PLS ; À droite, prédictions de la validation croisée pour k =10 variables latentes.

Le jeu E_1 sera utilisé pour tester des méthodes d'amélioration de la robustesse. Le jeu E_2 servira d'ensemble de test des modèles. La figure 2 montre le résultat du test sur E_2 du modèle étalonné sur E_0. À gauche, on voit clairement que l'estimation $\hat{\mathbf{y}}_2$ faite par le modèle subit un décalage directement dépendant de la température, comme confirmé par la figure de droite.

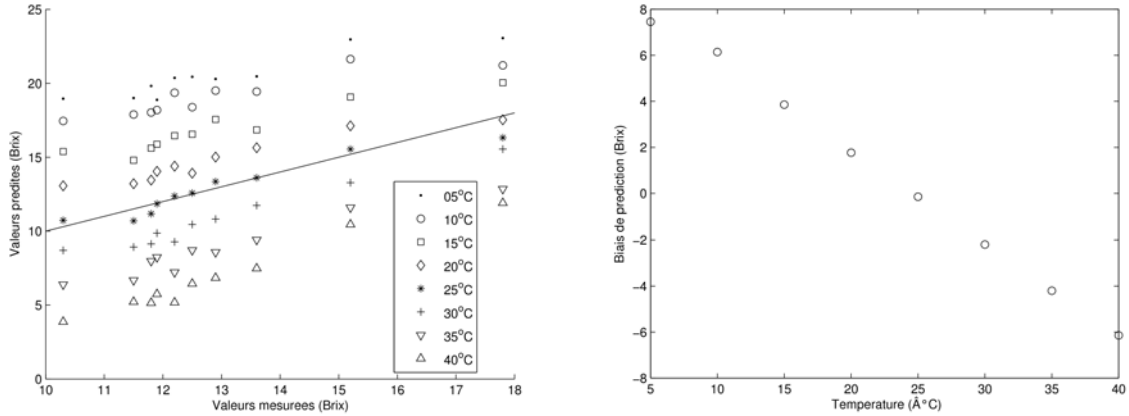


Figure 2 - Test du modèle d'estimation du taux de sucres des pommes sur différentes températures de fruits. À gauche, valeurs prédites par le modèle ; À droite, évolution du biais en fonction de la température.

1.3.2 Exemple 2 : Effet du millésime sur la mesure du taux de protéines du blé par spectrométrie visible / proche infrarouge

Cet exemple est décrit en détail dans [JR2]. Les spectres visible / proche infrarouge d'échantillons de grains de blé ont été acquis à l'aide d'un spectromètre FOSS NIR System 6500, en rétrodiffusion, sur la gamme spectrale 400 nm – 2500 nm. Les spectres ont été digitalisés sur p =1050 longueurs d'onde, régulièrement espacées de 2 nm. Le taux de protéines des échantillons était mesuré par une méthode de référence de laboratoire et exprimée en % de masse. Les échantillons de blé ont été récoltés et mesurés sur 8 années successives, de 1998 à 2005. La grandeur d'influence étudiée dans cet exemple est le millésime de la récolte.

Un premier jeu de données F_0 a été constitué en rassemblant les échantillons des 4 premières années, c'est à dire de 1998 à 2001. Ce jeu contenait une matrice $\mathbf{X}_0$ de n_0=456 spectres et un vecteur $\mathbf{y}_0$ du même nombre de taux de protéines.

Quatre jeux de test F_{2002}, F_{2003}, F_{2004} et F_{2005} ont été constitués avec les 4 années restantes, avec des effectifs de 84, 117, 107 et 121 individus pour les années 2002, 2003, 2004 et 2005.

Le jeu F_0 a été utilisé pour étalonner un modèle par régression PLS, sans prétraitement ni sélection de variables. La figure 3 gauche montre l'évolution de l'erreur d'étalonnage (SEC pour standard error of calibration) et de validation croisée (SECV pour standard error of cross validation) en fonction de la dimension de l'espace latent. Une dimension de k =17 a été choisie. La figure 3 droite montre les prédictions en validation croisée pour un espace latent de dimension k =17.

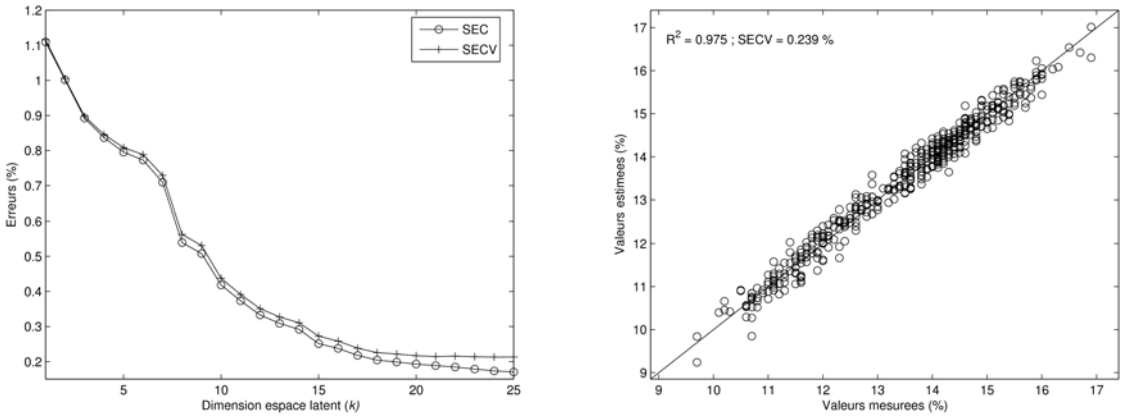
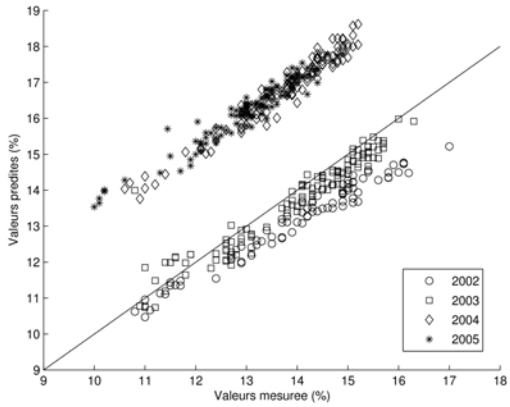


Figure 3 - Étalonnage du modèle de mesure du taux de protéines du blé. À gauche, évolution des erreurs en fonction du nombre de variables latentes de la PLS ; À droite, prédictions de la validation croisée pour k =17 variables latentes.

La figure 4 montre le résultat du test sur F_{2002}, F_{2003}, F_{2004} et F_{2005} du modèle étalonné sur F_0. À gauche, on voit clairement que le millésime de la récolte agit comme un effet de bloc, provoquant des biais et des pentes reportées sur la partie droite de la figure.



Année	2002	2003	2004	2005
Biais (%)	2,21	1,56	3,17	4,93
Pente	0,78	0,87	0,99	0,87

Figure 4 - Test du modèle d'estimation du taux de protéines du blé sur les 4 années 2002 à 2005. À gauche, graphe des prédictions ; À droite, biais et pentes des prédictions en fonction des années.

Stratégie générale de construction d'un étalonnage robuste

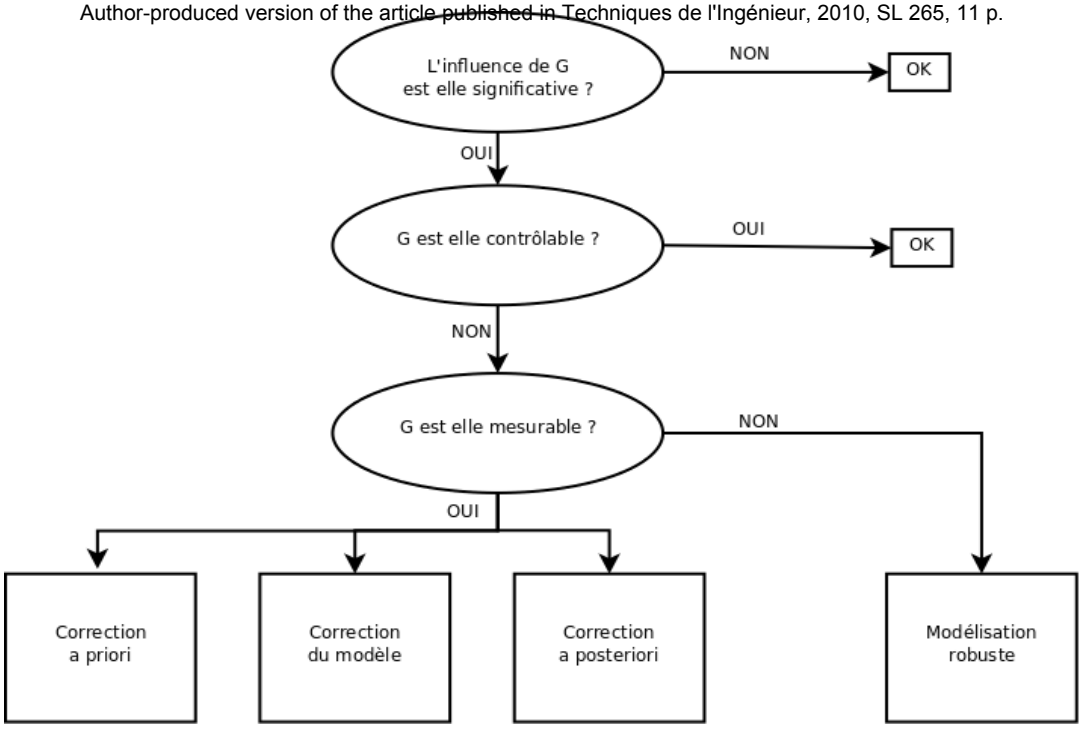


Figure 5 - Stratégie d'amélioration de la robustesse d'un modèle d'étalonnage, vis à vis d'une grandeur d'influence *G*.

Lorsque la robustesse d'un étalonnage est particulièrement recherchée, et que donc une grandeur *G* (que l'on supposera unique pour simplifier le propos) est impliquée, la construction du modèle doit suivre une démarche hiérarchique, comme indiquée par la figure 5. Cet organigramme utilise 3 niveaux de questionnement :

- En premier lieu, il conviendra de statuer sur l'importance de **l'influence** de *G* sur l'étalonnage. Cette première étape est rendue délicate par l'aspect multi-varié des signaux mesurés. En effet, dans le cas mono-varié, une influence de *G* sur le signal implique obligatoirement une influence sur $\hat{y}$. Il suffit donc d'examiner la sensibilité du signal mesuré vis à vis de *G*. Dans le cas multi-varié tel que la spectroscopie, il n'y a pas de lien univoque entre le signal mesuré et $\hat{y}$ car l'espace latent, utilisé par le modèle, est beaucoup plus petit que l'espace de mesure. Il faut donc tester l'influence de *G* sur le modèle d'étalonnage. Le problème qui survient alors est que cette sensibilité dépend de la manière dont est construit le modèle, c'est à dire : le type de régression, sa dimension, les prétraitements des signaux, les variables sélectionnées, etc. Ainsi, par exemple, certains prétraitements géométriques tels que la normalisation et la dérivation permettent de réduire considérablement l'effet de la diffusion dans les spectres acquis en rétro-diffusion (Cf [TD1]). Les méthodes les plus courantes pour caractériser l'effet de *G* utilisent donc un plan d'expériences, dont la variable de sortie est l'erreur de prédiction du modèle d'étalonnage. Les différents facteurs de ce plan, outre la grandeur *G* elle même, incluent les paramètres du modèle d'étalonnage. À noter que cette méthodologie de plans d'expériences s'adapte bien au cas de plusieurs grandeurs d'influence.
- En deuxième étape, si *G* est considérée comme influente, on se posera la question de sa **contrôlabilité**, c'est à dire de la possibilité de réduire les variations de *g*. La réponse à cette question peut être complexe. Elle doit tout d'abord tenir compte des aspects techniques. Prenons l'exemple où *G* est la température du spectromètre. S'il est techniquement facile de contrôler cette température dans une application de mesure en ligne où le spectromètre est à poste fixe, c'est beaucoup plus difficile à réaliser pour un spectromètre

portable. En complément de cet aspect technique, la possibilité de contrôler *G* peut aussi être empêchée par beaucoup d'autres raisons, par exemple économiques : si *G* est la température du fruit dont on veut mesurer le taux de sucres par spectrométrie sur une ligne de tri, le contrôle de sa température passe par un dispositif (par exemple un tunnel frigorifique) certainement très onéreux. D'autre part, même si *G* peut être contrôlée, une variation résiduelle due à l'erreur de l'asservissement subsistera. Il faudra alors repasser au point 1, pour vérifier que ces variations sont acceptables.

- Enfin, si *G* est influente et que ses variations ne peuvent pas être évitées, on se posera la question de sa **mesurabilité**, c'est à dire de connaître *g* au moment de l'utilisation du modèle d'étalonnage. Les deux cas, où *G* est ou n'est pas mesurable sont détaillés dans la suite de cet article.

1.4 Cas où la grandeur d'influence est mesurable

Dans cette partie, nous nous intéresserons aux cas où une grandeur *G* exerce une influence non négligeable, que ses variations ne peuvent être contrôlées, mais que sa valeur *g* est connue au moment de l'utilisation du modèle d'étalonnage. Deux cas peuvent se distinguer :

- Si *G* prend ses valeurs dans un ensemble continu, comme c'est le cas d'une grandeur physico-chimique classique, cela signifiera que l'on peut la mesurer en même temps que le signal multivarié. Par exemple, la température du fruit dont on veut estimer le taux de sucres par spectrométrie est mesurée par un thermomètre infrarouge.
- Si *G* prend ses valeurs dans un ensemble discret, comme c'est le cas pour une grandeur qualitative, cela signifiera que l'on est en mesure de fournir sa valeur au modèle. Par exemple, la variété du fruit dont on veut estimer le taux de sucres par spectrométrie est fournie par le système de gestion de production de l'usine.

Nous supposerons donc dans la suite de cette partie, que la valeur *g* (continue ou discrète) de *G* est disponible en même temps que le signal multivarié **x**. Comme indiqué sur la figure 5, trois classes de méthodes de correction peuvent alors être envisagées.

1.4.1 Correction *a priori*

Dans cette option, la valeur *g* de *G* est utilisée pour modifier **x**, de manière à retrouver le signal **x*** que l'on aurait mesuré en l'absence de perturbation, c'est à dire pour la valeur nominale *g** de *G* pour laquelle le modèle a été étalonné. Formellement, cela revient à appliquer une fonction *f* à **x** et *g* pour calculer **x*** et l'introduire dans le modèle qui reste inchangé :

$$\mathbf{x}^* = f(\mathbf{x}, g)$$
$$\hat{y} = \mathbf{x}^{*T}\mathbf{b} + b_0$$

Cette méthode nécessite de définir une fonction de R^{p+1} dans R^p. Elle est donc assez délicate à mettre en œuvre, et doit être réservée à des cas particuliers permettant de restreindre l'espace des possibles concernant la fonction *f*.

Dans le domaine de la spectrométrie optique, et plus particulièrement infrarouge, ce cas se présente pour le problème de la « standardisation optique » :

La standardisation optique vise à compenser les défauts d'alignement et de distorsion qui apparaissent entre l'appareil ayant servi à l'étalonnage (qualifié de maître) et les appareils utilisés dans un réseau de mesure (qualifiés d'esclave). Elle cherche donc à corriger les spectres mesurés par les appareils esclaves pour qu'ils ressemblent à ceux qui auraient été mesurés par l'appareil maître. Ici, la grandeur d'influence *G* est de nature discrète ; elle contient l'identification de l'appareil esclave.

La méthode la plus employée pour cette standardisation est sans conteste la *standardisation directe*, ou sa variante la *standardisation directe par morceaux* [YW1]. Une première hypothèse est que la fonction *f* est linéaire, c'est à dire

représentée par une matrice **F** mettant **x** et **x*** en correspondance. Une deuxième hypothèse, dans le cas de la standardisation par morceaux, est que la déformation des spectres esclaves est relativement faible, ce qui se traduit par le fait que la matrice **F** est quasi diagonale. L'identification de la matrice **F** passe par la mesure d'objets de référence appelés *« optical standards »*, qui sont des objets très stables, comme des céramiques ou des produits purs (comme le toluène) dont le spectre est connu et constant. Dans le cas de la mesure de produits stables, comme les céréales, des cellules scellées, contenant des échantillons de produit sont aussi utilisées comme standards. Elles offrent l'avantage de présenter des spectres proches de ceux qui seront mesurés par la suite. Cette option de correction *a priori* est présentée, d'une autre manière, dans le paragraphe 2.6.2 de [DB1].

1.4.2 Correction du modèle

Dans cette option, la valeur *g* de *G* est utilisée pour modifier le modèle (**b,** b_0). Formellement, cela peut s'écrire :

$$(\mathbf{b}^*, b_0^*) = f(\mathbf{b}, g)$$
$$\hat{y} = \mathbf{x}^T\mathbf{b}^* + b_0^*$$

Comme précédemment, la fonction *f* est définie dans de grands espaces ; de R^{p+1}dans R^{p+1}. Il est donc très difficile d'en estimer une expression analytique. Dans la pratique, cette correction de modèle n'est appliquée que dans le cas où *G* est discrète et ne peut prendre qu'un nombre relativement petit de valeurs. Par exemple, si *G* représente la variété d'une pomme, pouvant prendre ses valeurs dans l'ensemble {*golden*, *grany*, *fuji*}, la solution couramment employée consiste à développer un modèle spécifique à chaque variété {$\mathbf{b}_{golden}$,$\mathbf{b}_{grany}$,$\mathbf{b}_{fuji}$}, puis simplement à attribuer à **b** le modèle correspondant à la valeur de *G*. Cette solution peut aussi être appliquée au cas d'une grandeur G continue, en la discrétisant préalablement. Par exemple, on peut disposer de 3 modèles {$\mathbf{b}_{froid}$,$\mathbf{b}_{tiède}$,$\mathbf{b}_{chaud}$}, puis adopter chacun de ces trois modèles si la température des objets mesurés est dans une des trois classes {*froid*, *tiède*, *chaud*}, définies préalablement. L'inconvénient majeur de cette approche est qu'elle requiert autant d'étalonnages que de modalités de *G*.

1.4.3 Correction *a posteriori*

Cette option est très certainement la plus utilisée. Elle consiste à estimer l'erreur causée par *G*, puis de l'utiliser pour corriger l'étalonnage initial. Cette estimation est généralement fondée sur *g*, bien sûr et sur $\hat{y}$, l'estimation brute du modèle d'étalonnage. Formellement, cela revient à écrire :

$$\omega = \mathbf{x}^T\mathbf{b} + b_0$$
$$\hat{e} = f(\hat{y}, g)$$
$$\hat{y}^* = \mathbf{x}^T\boldsymbol{b} + b_0 - \hat{e}$$

Si l'influence de G se traduit par un biais pur (erreur systématique indépendante de *y*), l'estimation de $\hat{e}$ se fondera uniquement sur *g*. Dans le cas d'une erreur de type biais / pente, la prise en compte de $\hat{y}$ sera nécessaire. Cette méthode de correction est évoquée dans le paragraphe 2.6.1 de [DB1].

La correction *a posteriori* est illustrée sur l'exemple 1 (température des pommes) :

L'application sur le jeu E_1 du modèle étalonné sur le jeu E_0, montre que l'erreur de prédiction dépend essentiellement de la température (figure 6, gauche), et peu du taux de sucres (non représenté). La fonction *f* de correction, estimée par $f(\hat{y}, g) = -0.41\,g + 9.3$ est ensuite appliquée pour corriger les prédictions sur E_1 et produire les prédictions corrigées (figure 6, droite à comparer avec figure 2, gauche). Même si la plus grande partie des influences a été corrigée, il reste un biais général de prédiction. Il semblerait que le modèle d'erreur, appris sur le jeu E_1 ne soit pas parfaitement transposable à E_2.

La méthode de correction *a posteriori* est certainement la plus populaire, car la plus simple mais, comme le montre l'exemple précédent, elle nécessite que le modèle de

correction soit fiable, pour que la part d'incertitude qu'elle introduit ne soit pas trop élevée.

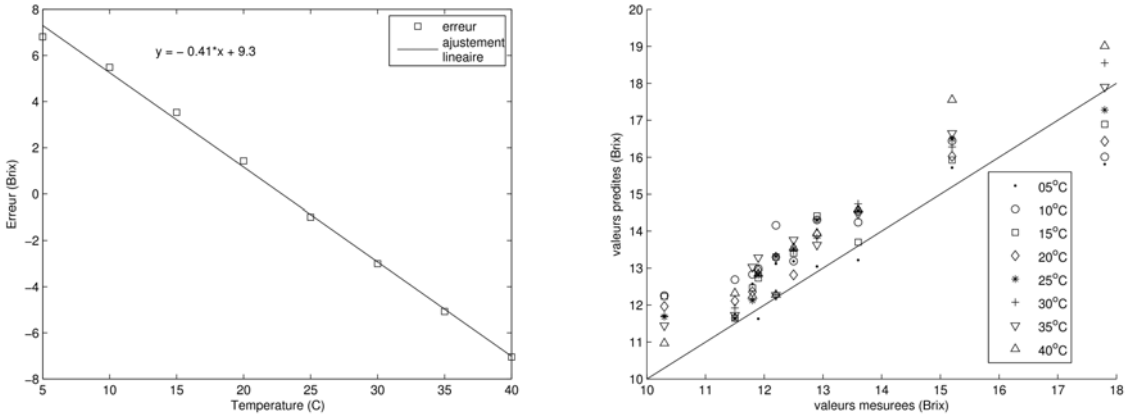


Figure 6 - À gauche ; courbe d'apprentissage de l'erreur en fonction de la température. À droite ; prédictions corrigées.

1.5 Cas des grandeurs d'influence non mesurables

Dans cette partie, nous nous intéresserons aux cas le plus défavorable, lorsqu'une grandeur *G* exerce une influence non négligeable, que ses variations ne peuvent être contrôlées, et que sa valeur *g* n'est pas connue au moment de l'utilisation du modèle d'étalonnage.

Dans ce cas, la robustesse du modèle doit être bien établie au moment de l'étalonnage. Comme nous allons le voir dans la suite, cette exigence peut simplement relever de bonnes pratiques ou bien être explicitement apportée.

L'équation (2) donnant la valeur absolue de l'erreur en fonction de l'influence, peut aussi s'écrire :

$$|\delta\omega| = ||\delta\mathbf{x}||.||\mathbf{b}||.|\cos(\delta\mathbf{x}, \mathbf{b})| \qquad (3)$$

L'erreur absolue due à l'influence de G est donc le produit de 3 termes positifs. Pour minimiser un tel produit, il suffit de minimiser un des 3 termes, c'est à dire, soit la norme de la perturbation, soit la norme du modèle, soit la dépendance entre le modèle et la perturbation.

1.5.1 Minimisation de la norme de perturbation $||\delta\mathbf{x}||$

Ce terme est la résultante de la variation de *G* sur le signal mesuré. Il est subi et ne peut normalement pas être changé. Cependant, certaines grandeurs d'influence particulières ont un effet systématique qui peut être considérablement réduit par l'application de prétraitements spécifiques.

C'est le cas en spectrométrie des influences de la diffusion de la lumière, induites par la granulométrie lorsque l'on mesure des solides ou de la turbidité, lorsque l'on mesure des liquides. Cette influence se traduit par des effets additifs et multiplicatifs, comme expliqué ci-après par l'altération de la loi de Beer Lambert :

Lorsque un rayonnement lumineux d'intensité $I_0(\lambda)$ traverse un produit d'épaisseur *e*, constitué d'un mélange de *k* produits en concentrations c_1, … , c_k, l'intensité résultante est donnée par :

$I(\lambda) = I_0(\lambda) \times \exp(-ec_1\varepsilon_1(\lambda)) \times ... \times \exp(-ec_k\varepsilon_k(\lambda))$

résultant en une absorbance

$A(\lambda) = -\log(I(\lambda) / I_0(\lambda)) = e\,[\,c_1\varepsilon_1(\lambda) + \ldots + c_k\varepsilon_k(\lambda)\,]$

Si une diffusion de la lumière intervient, une partie des photons échappent au détecteur, et cette perte d'intensité est comptée comme une absorption. Cela se traduit par un terme multiplicatif $K_1(\lambda)$ dans

l'intensité, et additif dans l'absorbance :

$I(\lambda) = I_0(\lambda) \times K_1(\lambda) \times \exp(-ec_1\varepsilon_1(\lambda)) \times ... \times \exp(-ec_k\varepsilon_k(\lambda))$, donnant :

$A(\lambda) = e\,[\, c_1\varepsilon_1(\lambda) + \ldots + c_k\varepsilon_k(\lambda)\,] + \log(\, K_1(\lambda)\,)$

D'autre part, la diffusion augmente le trajet optique, d'un facteur $K_2(\lambda)$, se traduisant par une modification exponentielle de l'intensité et multiplicative de l'absorbance :

$I(\lambda) = I_0(\lambda) \times K_1(\lambda) \times \exp(-K_2(\lambda)ec_1\varepsilon_1(\lambda)) \times ... \times \exp(-ec_k\varepsilon_k(\lambda))$, donnant :

$A(\lambda) = K_2(\lambda) \times e\,[\, c_1\varepsilon_1(\lambda) + \ldots + c_k\varepsilon_k(\lambda)\,] + \log(\, K_1(\lambda)\,)$

Dans la pratique, en spectrométrie infrarouge, le terme $K_2(\lambda)$ est considéré comme constant et le terme $\log(\, K_1(\lambda)\,)$ est modélisé par un polynôme de λ, généralement de degré 1 ou 2. Donc, comme cela est très bien montré dans [TD1] l'effet $\delta\mathbf{x}$ de la granulométrie ou de la turbidité sur les spectres d'absorbance proche infrarouge peut être considérablement réduit par des prétraitements spécifiques de normalisation (pour l'effet multiplicatif) et de dérivation (pour l'effet additif). L'exemple suivant illustre ces corrections :

Dans cet exemple, les spectres de 36 échantillons de vins sont mesurés avec un spectromètre Jasco V-570, en référence à l'eau dans une cuve de 1 mm de trajet optique. Il s'agit de 6 vins différents, à 6 niveaux de turbidité (Cf [SP1] pour une description détaillée). La figure 7, gauche, montre les spectres bruts. On y voit clairement une ligne de base, caractérisée par une tendance linéaire décroissante et un effet multiplicatif qui s'applique globalement sur tous les spectres. La figure 7, droite montre les mêmes spectres après dérivation seconde (algorithme de Savitsky et Golay, [PG1], fenêtre de lissage 140 nm, ordre du polynôme 3) et normalisation par la moyenne quadratique du spectre (traitement SNV, [RB1]). On voit clairement que les spectres ne sont plus affectés d'aucune ligne de base, et que les effets multiplicatifs ont été considérablement réduits. En effet, des classes de spectres apparaissent, correspondant aux 6 vins utilisés. On peut constater également que la correction n'est pas parfaite. Cette imperfection est probablement due au fait que la modélisation de la ligne de base est trop pauvre.

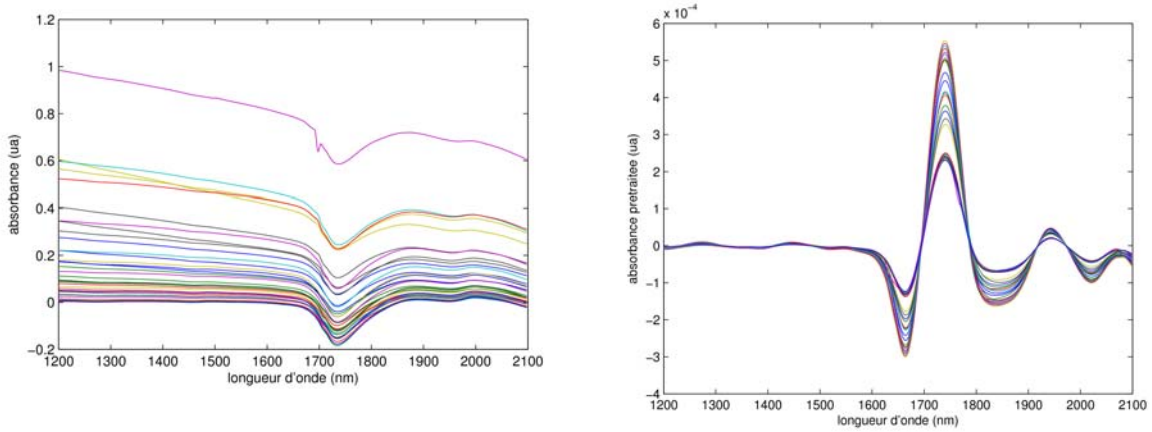


Figure 7 - Spectres infrarouges de vins turbides, avant et après normalisation et réduction de ligne de base.

1.5.2 Minimisation de la norme du modèle $||\,\mathbf{b}\,||$

Ce terme, bien connu des statisticiens, traduit la complexité du modèle. Le lecteur se rapportera utilement au paragraphe 1.8.2 de [DB1], où cette notion de complexité est largement décrite. Des explications plus théoriques peuvent aussi être trouvées dans [MS1]. Il convient de retenir que, plus le modèle est complexe, plus sa norme est élevée, et par conséquent plus sa sensibilité aux grandeurs d'influence augmente également. Dans le cadre de la spectrométrie, où les méthodes de régression factorielle sont utilisées, la complexité du modèle est directement liée au nombre de dimensions de l'espace latent utilisé, comme le montre l'exemple suivant :

Cet exemple utilise les données décrites dans le paragraphe 1.3.1. Il montre, figure 8, l'évolution du terme $||\mathbf{b}||$ en fonction du nombre de variables latentes introduites dans la régression PLS, en superposition avec les erreurs d'étalonnage (SEC) et de validation croisée (SECV). On voit clairement une augmentation brutale de $||\mathbf{b}||$ aux alentours de 10 à 13 variables latentes, lorsque l'erreur de validation croisée cesse de décroître. Si la dimension de l'espace latent est trop grande, le modèle est étalonné sur des informations qui contiennent plus de bruit que de tendances significatives du phénomène modélisé, et sa robustesse en est altérée.

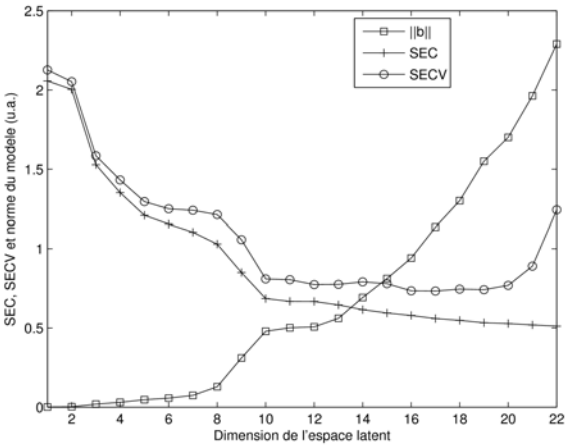


Figure 8 - Évolutions comparées de la norme du modèle et de l'erreur de validation croisée, traduisant sa capacité à généraliser.

1.5.3 Minimisation de la dépendance entre le modèle et la perturbation $|\cos(\delta\mathbf{x}, \mathbf{b})|$

C'est en agissant sur ce terme que l'on peut effectivement améliorer la robustesse du modèle d'étalonnage, en regard d'une ou plusieurs grandeurs d'influence données. Comme cela a déjà été dit plus haut, il est assez difficile d'agir sur le terme $\delta\mathbf{x}$. C'est donc sur la construction du modèle **b** que l'on doit se focaliser, pour minimiser $|\cos(\delta\mathbf{x}, \mathbf{b})|$, c'est à dire orthogonaliser le modèle vis à vis de la perturbation. Deux grandes options sont disponibles : l'orthogonalisation implicite et explicite.

L'orthogonalisation implicite est la méthode la plus naturelle. Elle consiste simplement à inclure dans la collection des échantillons d'étalonnage des exemples de variation de *G*, donc des exemples de $\delta\mathbf{x}$, ce qui permettra à la régression de trouver par elle même un espace latent indépendant de *G*. Par exemple, pour réaliser un modèle indépendant de la température, on réalise l'étalonnage sur des spectres acquis à différents niveaux de température.

Cette méthode a été appliquée aux données de l'exemple 1, décrit en 1.3.1. La base d'étalonnage isotherme E_0 a été concaténée avec E_1, dont les individus ont été mesurés à des températures variant de 5 à 40°C. Un nouvel étalonnage a été réalisé sur cette nouvelle base. La figure 9 gauche montre les résultats de la validation croisée. On voit que la dimension optimale du modèle est maintenant de 18, contre 10 dans le cas isotherme (figure 1, gauche). La figure 9 droite montre le résultat de l'application de ce modèle sur le jeu E_2. La majeure partie des influences a été corrigée, mais un biais général subsiste.

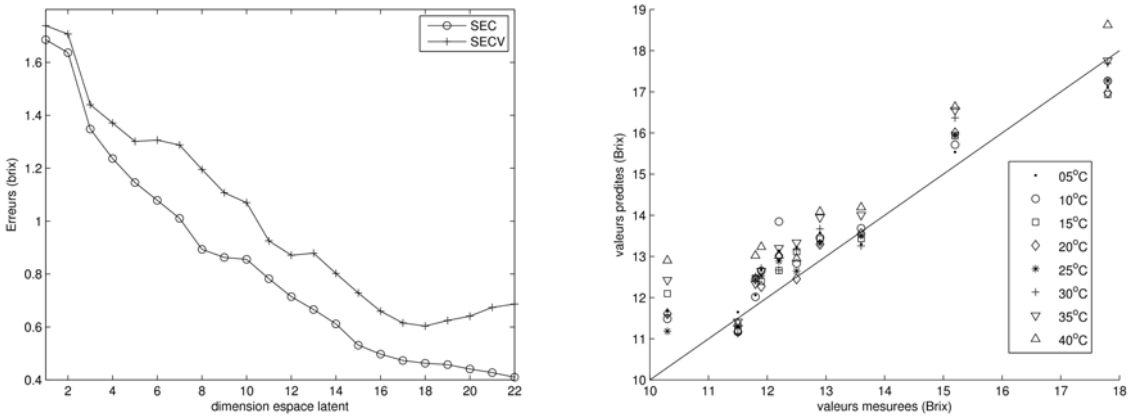


Figure 9 - Orthogonalisation implicite du modèle d'étalonnage, par incorporation dans la base d'étalonnage de spectres altérés, contenant des exemples de $\delta\mathbf{x}$. À gauche, validation croisée sur $\{E_0 \cup E_1\}$. À droite, résultat du test sur E_2.

L'orthogonalisation explicite consiste à identifier les dimensions de l'espace qui sont majoritairement touchées par les perturbations $\delta\mathbf{x}$, puis de les retirer par projection orthogonale avant de réaliser l'étalonnage. Les fondements théoriques de cette méthode ainsi que différentes manières de l'appliquer peuvent être trouvés dans [PH1], [JR1] et [AA1]. Dans tous les cas, l'identification de l'espace à ôter par projection consiste à :

- constituer un ensemble **D** de spectres représentatif de l'effet des variations de *G* (grandeur d'influence) sans faire varier *Y* (grandeur d'intérêt). Pour les grandeurs d'influence que l'on peut maîtriser, on réalisera cette collection au moyen d'un plan d'expériences (cf. premier exemple, ci-dessous). Pour les autres, ces spectres pourront être estimés par le calcul (Cf deuxième exemple, ci-dessous).
- Identifier une base dans l'espace décrit par **D**. Cette opération peut se réaliser aisément au moyen d'une décomposition en valeur singulière (SVD) ou simplement en retenant les premiers vecteurs propres d'une Analyse en Composantes Principales (ACP) sur **D** (non centrée, non normée). Cette base **P** contient *k* vecteurs orthonormés de l'espace spectral, de dimension *p*.
- Projeter la base d'étalonnage orthogonalement à **P** :

$$\mathbf{X} = \mathbf{X} (\mathbf{I} - \mathbf{P}^T\mathbf{P})$$

- Étalonner un nouveau modèle sur la base orthogonalisée.

Premier exemple : L'orthogonalisation explicite est appliquée sur les données de l'exemple 1, présenté au paragraphe 1.3.1 (cf. détails de la procédure dans [JR1]). Les spectres du jeu de données E_1 sont regroupés par pomme. Ceci donne 10 matrices de 8 spectres du même fruit acquis à différentes températures. Ces 10 matrices sont ensuite centrées individuellement. De la sorte, dans chaque matrice ne subsiste que les variations inter-températures, l'information relative au fruit ayant disparu dans le centrage. L'ensemble de ces 80 spectres est rassemblé dans la matrice **D**, puis les *k* vecteurs de la base **B** sont calculés par une SVD sur **D**. Pour choisir conjointement le nombre *k* de dimensions à enlever par projection et le nombre de variables latentes du nouveau modèle, des validations croisées sont réalisées sur le jeu E_1 avec des valeurs de *k* allant de 1 à 6 et un nombre de variables latentes *lv* allant de 1 à 20. Ceci produit une surface de SECV, fonction de *k* et de *lv*, comme reporté en figure 10, gauche. Le choix de *k*=4 et *lv*=12 est effectué ; le modèle en résultant appliqué sur E_2 donne les prédictions reportées sur la figure 10, droite. Les effets de la température sont en très grande partie corrigés.

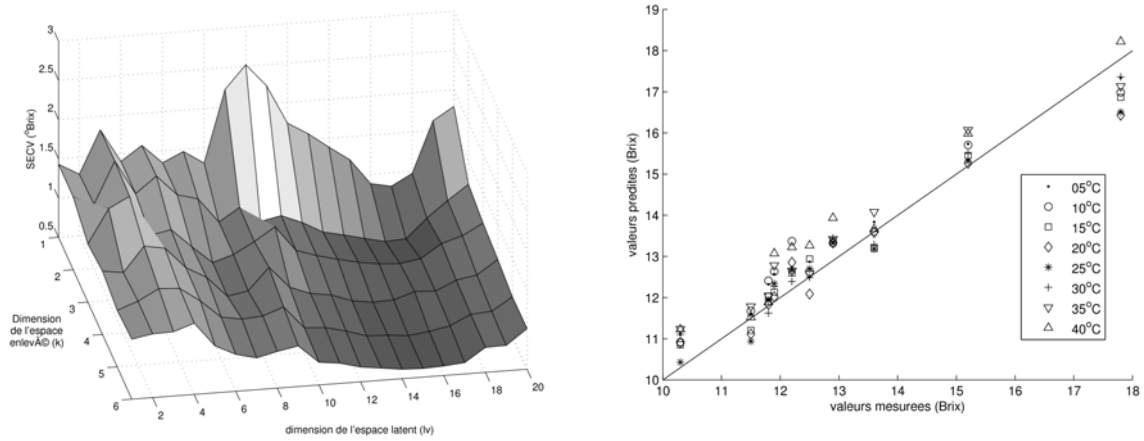


Figure 10 - Application de l'orthogonalisation explicite aux données de l'exemple 1. À gauche, erreur de validation croisée sur le jeu E_1 en fonction des dimensions de l'espace soustrait et de l'espace latent. À droite, prédictions réalisées sur le jeu de données E_2 avec un modèle corrigé de 4 dimensions et étalonné sur E_0 avec 12 variables latentes.

Deuxième exemple : L'orthogonalisation explicite est appliquée sur les données de l'exemple 2, présenté au paragraphe 1.3.2 (Cf détails de la procédure dans [JR2]). Comme les grandeurs d'influence responsables des effets de bloc, apparaissant d'un millésime à l'autre, sont inconnues, les spectres de la matrice **D** doivent être synthétisés par le calcul. Pour ce faire, on suppose que l'on dispose du spectre et de la valeur de la réponse de quelques échantillons ($\mathbf{X}_c$, $\mathbf{y}_c$) de l'ensemble de test, posant le problème de robustesse. Pour chaque valeur y de $\mathbf{y}_c$, on applique un noyau centré sur y sur le vecteur $\mathbf{y}_0$ de la base d'étalonnage, afin de trouver la combinaison linéaire **c** des individus de $\mathbf{y}_0$ qui donne la valeur y. Cette combinaison linéaire est ensuite appliquée sur les spectres $\mathbf{X}_0$, ce qui donne une estimation du spectre qui aurait dû être mesuré pour la valeur y. La différence entre le spectre estimé et le spectre mesuré est adopté comme représentatif de l'influence à corriger, et stocké dans **D**. L'algorithme de cette méthode est donné ci-après :

1. soit **x** le spectre d'un individu acquis dans des conditions perturbées, et y la réponse correspondante
2. trouver la combinaison linéaire **c** des n_0 individus de la base d'étalonnage telle que $\mathbf{c}^T\mathbf{y_0} = y$ *(par exemple au moyen d'un noyau)*
3. appliquer **c** à $\mathbf{X}_0$, pour produire une estimation du spectre qui aurait dû être mesuré en l'absence de perturbation : $\mathbf{x}^* = \mathbf{c}^T\mathbf{X}_0$
4. calculer la différence entre les spectres mesuré et estimé : $\mathbf{d} = \mathbf{x} - \mathbf{x}^*$

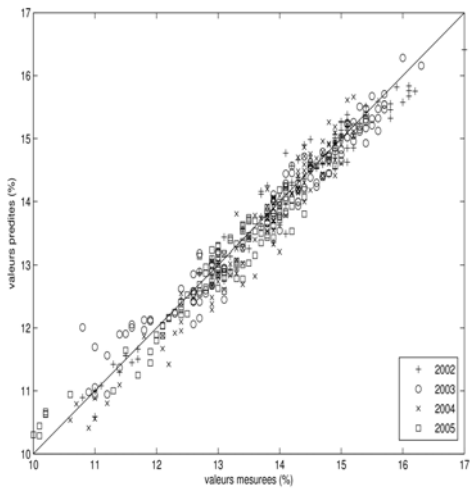
La figure 11 montre le résultat de l'application de cette méthode sur les données de l'exemple 2, en prenant pour le recalage les 3 premiers échantillons de chaque année de test. Les biais et les pentes qui étaient observés avant correction (figure 4) ont pratiquement disparu.

L'orthogonalisation explicite est une méthode très récente, qui commence à être intégrée dans certains logiciels dédiés à la spectrométrie (WinEasy, paquet logiciel de FOSS et The Unscrambler, logiciel d'analyse de données de CAMO). Les avantages de cette méthode, outre l'amélioration de la robustesse, sont assez nombreux :

- Le modèle devient indépendant de G, donc
- Le modèle fonctionne aussi en l'absence de perturbation
- La correction est « embarquée » dans le modèle, il n'y a pas besoin

d'orthogonaliser les signaux des nouvelles mesures

- Plusieurs influences peuvent être corrigées simultanément
- La partie corrigée peut être interprétée
- On peut traiter des bases d'étalonnage existantes, indépendamment du logiciel d'étalonnage



Année	2002	2003	2004	2005
Biais (%)	0,03	-0,04	-0,1	-0,08
Pente	0.96	0,93	1,1	0,92

Figure 11 - Application de l'orthogonalisation explicite aux données de l'exemple 2. À gauche, prédiction sur les jeux F_{2002}, F_{2003}, F_{2004} et F_{2005}. À droite, biais et pentes de chacune de ces prédictions.

Conclusion

Dans cet article, nous avons vu que la robustesse des étalonnages prenait un caractère particulier dans le cas hautement multivarié des signaux continus, tels que les spectres. La très grande dimension de l'espace de mesure laisse en effet une grande latitude à l'expression des grandeurs d'influence, si aucune précaution n'est prise. La stratégie de prise en compte de ce problème est déclinée, selon les propriétés de contrôlabilité et de mesurabilité de la grandeur d'influence.

Dans le cas où on dispose de la valeur de la grandeur d'influence au moment de l'utilisation du modèle d'étalonnage, trois méthodes ont été décrites : la correction a priori, qui consiste à modifier la mesure pour correspondre aux conditions de l'étalonnage ; la correction du modèle, qui consiste à adapter le modèle aux nouvelles conditions ; la correction a posteriori, qui consiste à corriger les sorties du modèle.

Lorsque la valeur de la grandeur d'influence n'est pas disponible au moment de l'utilisation du modèle d'étalonnage, il faut construire un modèle robuste, soit implicitement, en soignant la construction de la base d'étalonnage, soit explicitement, en ôtant de l'espace de mesure les dimensions responsables de la non robustesse du modèle. Ces méthodes, basées sur l'orthogonalisation de sous espaces vectoriels sont présentées à la fin de cet article.

Author-produced version of the article published in Techniques de l'Ingénieur, 2010, SL 265, 11 p.

Pour en savoir plus

La robustesse des étalonnages multidimensionnels

Application aux données spectrales

par **Jean-Michel ROGER**

Ingénieur en Chef du Génie Rural des Eaux et des Forêts

Chercheur

À lire également dans nos bases

[DB1] BERTRAND D. : Étalonnage multidimensionnel : application aux données spectrales ; Les techniques de l'ingénieur ; 10 mars 2005 ; Référence P264
[MD1] DÉSENFANT, M. ; PRIEL, M. et RIVIER, C. : Évaluation des incertitudes des résultats d'analyse. Les techniques de l'ingénieur ; 10 décembre 2005 ; Référence P105

Sources bibliographiques

[AA1] ANDREW, A. and FEARN, T. Transfer by orthogonal projection: making near-infrared calibrations robust to between-instrument variation. Chemometrics and Intelligent Laboratory Systems, 72(1):51–56, June 2004.
[HM1] MARTENS, H. et NAES, T. Multivariate Calibration. Wiley, New
[JR1] ROGER, J.M., CHAUCHARD, et BELLON-MUREL, V. EPO-PLS External Parameter Orthogonalisation of PLS : Application to temperature independent measurement of sugar content of intact fruits. Chemometrics and Intelligent Laboratory Systems, 66-2:191–204, 2003.
[JR2] ROGER, J.M., CHAUCHARD, F., et WILLIAMS, P. Removing the block effects in calibration by means of dynamic orthogonal projection. application to the year effect correction for wheat protein prediction. Journal of Near Infra Red Spectroscopy, 16:311–315, 2008.
[MS1] SEASHOLTZ, M.B. et KOWALSKI, B.R. The parsimony principle applied to
[PG1] GORRAY, P.A. General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method. Anal. Chem., 62:570–573, 1990.
[PH1] HANSEN, P.W. Pre-processing method minimizing the need for reference analysis. J. Chemometrics, 15:123–131, 2001.
[RB1] BARNES R.J., DHANOA, M. S. and LISTER S.J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl. Spectrosc., 43-5:772–777, 1989.
[SP1] PREYS, S., ROGER, J.M., BOULET, J.C. Robust calibration using orthogonal projection and experimental design. application to the correction of the light scattering effect on turbid NIR spectra. Chemometrics and Intelligent Laboratory Systems, 91:28–33, 2008.
[TD1] DAVIS, T. et FEARN, T. Back to basics : removing multiplicative effects (1). Spectroscopy Europe, 19(4):24–28, 2007.
[YW1] WAND, Y., VELTKAMP, D.J., KOWALSKI, R. Multivariate Instrument Standardization, Analytical Chemistry 63 (1991) 2750-2758.
multivariate calibration. Anal. Chim. Acta, 277:165–177, 1993.
York. 2005