Revue des méthodes d'estimation de la validité des prédictions

Hélio-SPIR

30 Septembre 2011, Montpellier

Sébastien Preys, Jordane Lallemand - ONDALYS

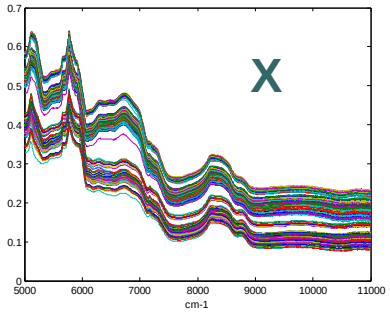
ondalys

Plan

Modèle d'étalonnage multivarié

ETALONNAGE

Développement modèle



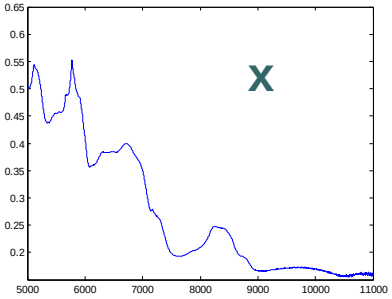
+ **y** → Modèle **b**

$$y = \mathbf{x}^T \mathbf{b} + \varepsilon$$

incertitude globale/moyenne $RMSEP = \sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$

PREDICTION

Utilisation modèle

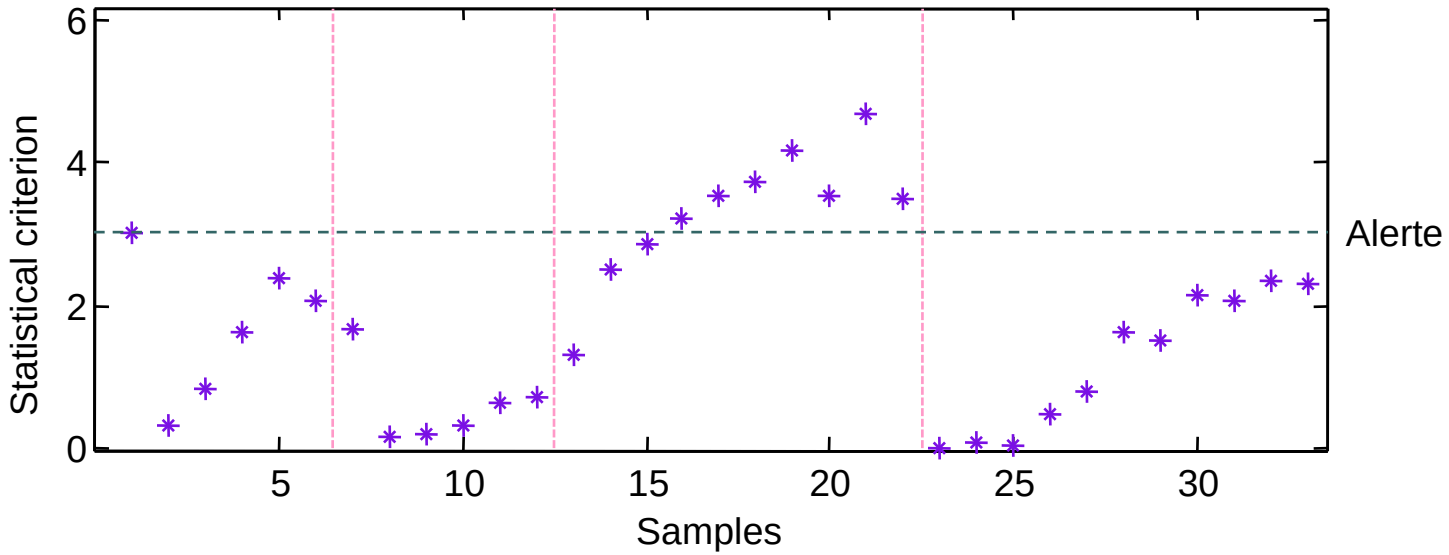


+ Modèle **b** → $\hat{y}$

quelle incertitude individuelle ?

Intérêt : outil de diagnostic en temps réel 1

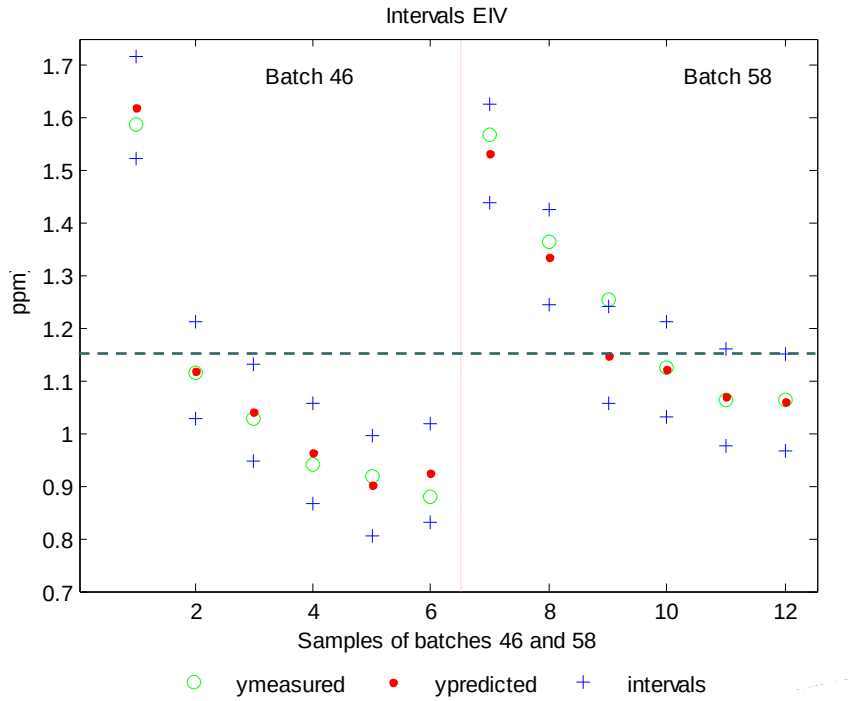
- Monitoring et détection des dérives



Intérêt : outil de diagnostic en temps réel 2

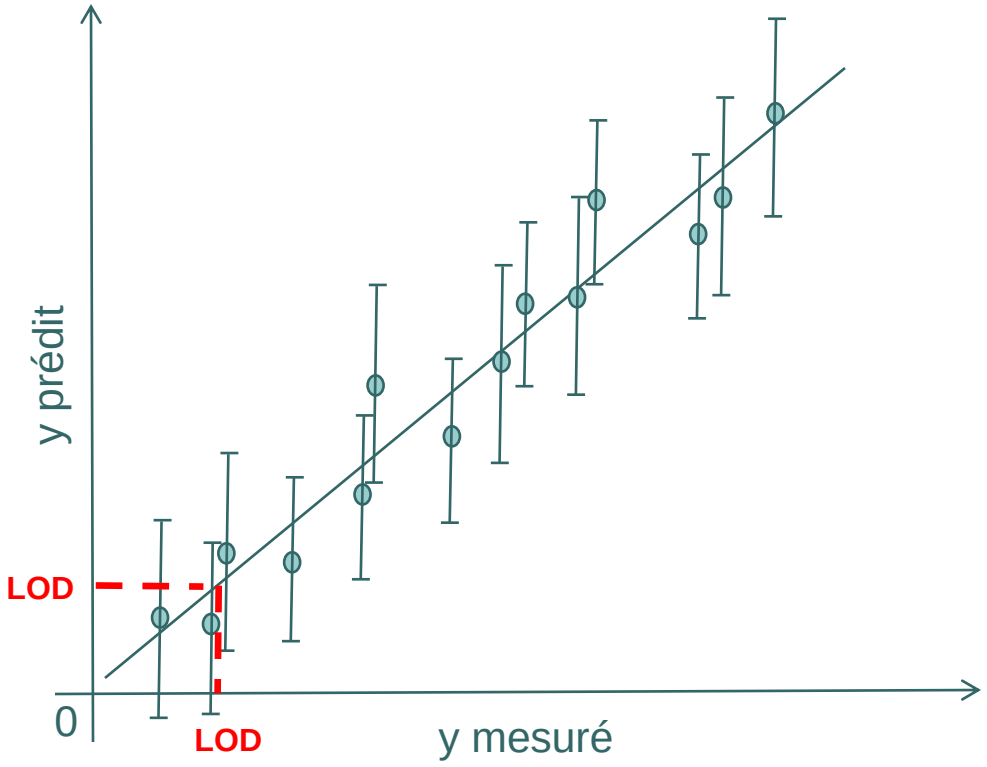
- Monitoring et prise de décision

  Ex.: Batch end-point



Intérêt : limite de détection (LOD)

- Calcul de la LOD



Plan

I. Contexte

II. Expressions analytiques

III. Méthodes de ré-échantillonnage

IV. Conclusions

Cas de la MLR (OLS)

- Pas ou peu de colinéarité dans **X**

Modèle

$$y = \mathbf{x}^{\mathbf{T}}\mathbf{b} + \varepsilon$$

Matrice de variance-covariance du vecteur de régression

$$\boldsymbol{\Sigma}_{\hat{\mathbf{b}}} = \hat{\sigma}_{\varepsilon}^{2}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}$$

Variance résiduelle du modèle = erreur globale du modèle

Estimation de la variance de la valeur de prédiction

$$\hat{\sigma}^{2}(\hat{y}_i) = \hat{\sigma}_{\varepsilon}^{2}\,\mathbf{x}_i^{\mathbf{T}}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{x}_i = \hat{\sigma}_{\varepsilon}^{2}h_i$$

Levier

Intervalle de confiance

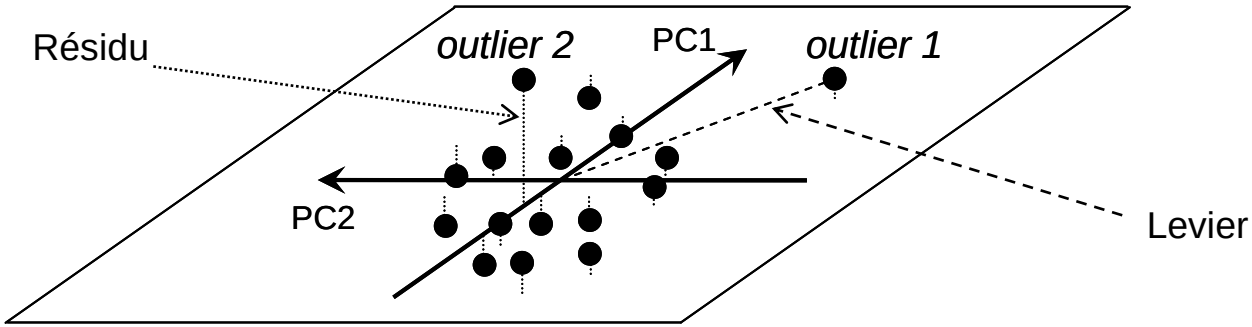
$$\hat{y}_i \pm t.\hat{\sigma}(\hat{y}_i)$$

Statistique de Student

Levier

- Levier (*leverage*) = distance au centre du modèle = T^2 (distance d'Hotelling) ~ MD (*Mahalanobis Distance*)

- Résidu = distance au modèle = Q



Cas de la PLS

- Colinéarité dans **X** → modèle pour **X** → expression pour la MLR trop optimiste…

Modèle

$$y = \mathbf{x}^{\mathbf{T}}\mathbf{b} + \varepsilon$$

Matrice de variance-covariance du vecteur de régression

$$\boldsymbol{\Sigma}_{\hat{\mathbf{b}}} = \hat{\sigma}_{\varepsilon}^{2}(\mathbf{R}\mathbf{R}^{\mathbf{T}})$$

R (p x k) matrice de poids

Variance résiduelle du modèle = erreur globale du modèle

Estimation de la variance de la valeur de prédiction

$$\hat{\sigma}^{2}(\hat{y}_i) = \hat{\sigma}_{\varepsilon}^{2}(1 + \frac{1}{n} + h_i)$$

Levier

Höskuldsson, A. (1988). "PLS regression methods." Journal of Chemometrics 2(3): 211-228.

ASTM (2000). Standard practices for infrared multivariate quantitative analysis - E 1655: 592-619.

Modèle EIV (Errors In Variables)

- Propagation des erreurs → prise en compte des erreurs de mesures dans **X** et dans **y**

PHASE D'ETALONNAGE — PHASE DE PREDICTION

$$\hat{\sigma}^2(\hat{y}_i) = (\frac{1}{n} + h_i)(\hat{\sigma}^2_{\varepsilon} + \hat{\sigma}^2_{\delta y} + \|\mathbf{b}\|^2 \hat{\sigma}^2_{\delta X}) + \hat{\sigma}^2_{\varepsilon_i} + \|\mathbf{b}\|^2 \hat{\sigma}^2_{\delta Xi}$$

Variance de l'erreur de mesure de référence

Variance de l'erreur de mesure spectrale

Variance résiduelle du modèle pour la phase de prédiction

Variance de l'erreur de mesure spectrale pour la phase de prédiction

Hypothèses : « iid noise » (independently and identically distributed)

- Bruit sans structure de corrélation
- Homoscedasticité

Faber, K. and B. R. Kowalski (1996). "Prediction error in least squares regression: Further critique on the deviation used in The Unscrambler." Chemometrics and Intelligent Laboratory Systems 34(2): 283-292.

Modèle EIV simplifié

- Simplification dans le cas où les erreurs dans les phases d'étalonnage et de prédiction sont proches

Rappel : EIV

$$\hat{\sigma}^2(\hat{y}_i) = (\frac{1}{n} + h_i)(\underbrace{\hat{\sigma}_\varepsilon^2 + \hat{\sigma}_{\delta y}^2 + \|\mathbf{b}\|^2 \hat{\sigma}_{\delta X}^2}_{\sim MSEC}) + \hat{\sigma}_{\varepsilon_i}^2 + \|\mathbf{b}\|^2 \hat{\sigma}_{\delta Xi}^2$$

EIV simplifié

$$\hat{\sigma}^2(\hat{y}_i) = (1 + \frac{1}{n} + h_i) MSEC - \hat{\sigma}_{\delta y}^2$$

Hypothèses : plus de « iid noise »

- Bruit corrélé
- Hétéroscedasticité
- Pas valable si biais important
- Pas valable si interférence non modélisée (résidu important)

Rappel : Höskuldsson et ASTM $\hat{\sigma}^2(\hat{y}_i) = \hat{\sigma}_\varepsilon^2(1 + \frac{1}{n} + h_i)$

Hodges, S.D. and P.G. Moore (1972). "Data uncertainties and least squares regression". Journal of the Royal Statistical Society. Series C (Applied Statistics) 21 (2): 185-195.

Faber, N. M. and R. Bro (2002). "Standard error of prediction for multiway PLS: 1. Background and a simulation study." Chemometrics and Intelligent Laboratory Systems 61(1-2): 133-149.

Expression de The Unscrambler ®

- Prise en compte du levier ET du résidu

Variance résiduelle de l'éch. de prédiction

Prediction Error Sum of Squares

1ère version (H.Martens)

$$\hat{\sigma}^2(\hat{y}_i) = \left(\frac{V_{xi,pr}}{V_{x\_tot,val}} + \frac{1}{n} + h_i \right) \frac{PRESS}{2n}$$

Variance résiduelle moyenne en validation

Nombre de composantes du modèle

2ème version (1995)

$$\hat{\sigma}^2(\hat{y}_i) = \underbrace{\left(2\left(1 - \frac{(A+1)}{n} \right) \right)}_{\text{coefficient}} \left(\frac{V_{xi,pr}}{V_{x\_tot,val}} + \frac{1}{n} + h_i \right) \frac{PRESS}{2n}$$

De Vries, S. and C. J.F. Ter Braak (1995). "Prediction error in partial least squares regression: a critique on the deviation used in The Unscrambler." Chemometrics and Intelligent Laboratory Systems 30(2): 239-245.

Plan

I. Contexte

II. Expressions analytiques

III. Méthodes de ré-échantillonnage

IV. Conclusions

Principe général

- Méthodes empiriques
- Génération artificielle de plusieurs populations d'étalonnage à partir de celle d'origine (réelle)

↳ Génération de plusieurs modèles

↳ Génération de plusieurs valeurs de prédiction

↳ Ecart-type ou variance de la valeur de prédiction = incertitude $\hat{\sigma}^2(\hat{y}_i)$

- Pas d'hypothèses à vérifier, mais calcul intensif

Ré-échantillonage sur objets

- **Jack-knife = validation croisée**
  - Tirage aléatoire sans remise
  - 2≤m≤n jeux de données = nombre de segments
  - Ex.: leave-one-out
  - Attention au données clusterisées !

- **Bootstrap**
  - Tirage aléatoire avec remise
  - Ex.: m=1000
  - Attention au données clusterisées !

Efron, B. (1979). "Bootstrap methods : another look at the jackknife." Annals of Statistics 7: 1-26.

Ré-échantillonage sur résidus

- **Bootstrap de résidus**
  - 1er modèle
  - Tirage aléatoire avec remise des résidus de prédiction
  - Addition de ces résidus aux y prédits $y_i^b = \hat{y}_i + e_i^b$
  - Calcul d'un 2ème modèle, etc…

Faber, N. M. (2002). "Uncertainty estimation for multivariate regression coefficients." Chemometrics and Intelligent Laboratory Systems 64(2): 169-179.

- **Addition de bruit (*Noise addition*)**
  - Principe similaire au bootstrap de résidus
  - 1er modèle
  - Addition de résidus distribués normalement aux y
  - Calcul d'un 2ème modèle, etc…

$$y_i^b = y_i + RMSEC.F(0,1)$$

F. Javier del Río, J. R., F. Xavier Rius, (2001). "Prediction intervals in linear regression taking into account errors on both axes." Journal of Chemometrics 15(10): 773-788.

Plan

I. Contexte

II. Expressions analytiques

III. Méthodes de ré-échantillonnage

IV. Conclusions

Conclusions

Expressions analytiques

$$\hat{\sigma}^2(\hat{y}_i) = \hat{\sigma}_\varepsilon^2(1+\frac{1}{n}+h_i)$$ Höskuldsson, ASTM

$$(\frac{1}{n}+h_i)(\hat{\sigma}_\varepsilon^2+\hat{\sigma}_{\delta y}^2+\|\mathbf{b}\|^2\hat{\sigma}_{\delta X}^2)+\hat{\sigma}_{\varepsilon_i}^2+\|\mathbf{b}\|^2\hat{\sigma}_{\delta Xi}^2$$ Faber (1996)

$$(1+\frac{1}{n}+h_i)MSEC-\hat{\sigma}_{\delta y}^2$$ Faber (2002)

$$\left(2\left(1-\frac{(A+1)}{n}\right)\right)\left(\frac{V_{xi,pr}}{V_{x\_tot,val}}+\frac{1}{n}+h_i\right)\frac{PRESS}{2n}$$ Unscrambler ®

Distance de Mahalanobis (MD)

Résidu (Q) ?

Méthodes de ré-échantillonnage

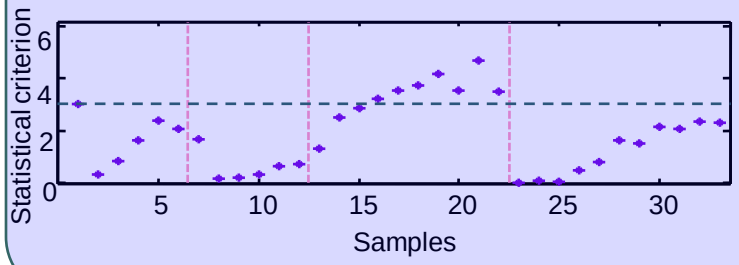
Jack-knife

Bootstrap sur objets

Bootstrap sur résidus

Addition de bruit

VALIDATION



monitoring

outliers

Outils d'Aide à la Décision

LOD

interférences