Une nouvelle formule d'estimation de l'incertitude des prédictions réalisées par des modèles linéaires

Application à la spectrométrie PIR

E Fernandez-Ahumada, JM Roger, B Palagos

jean-michel.roger@cemagref.fr

Plan

- Introduction
- Théorie
- Matériel & Méthodes
- Résultats
- Conclusion

Introduction

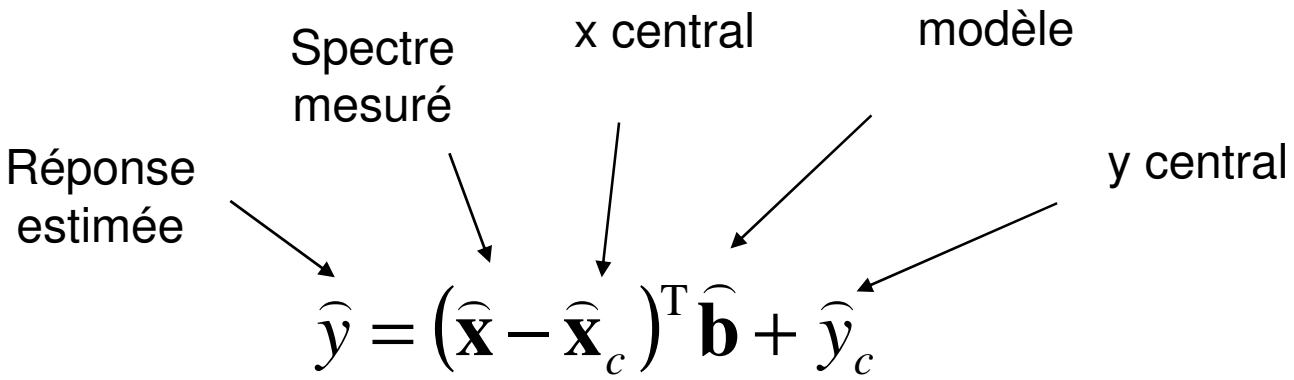
- Il existe deux types de méthodes pour estimer l'incertitude de prédiction d'un modèle
  - Ré-échantillonage (p.ex. bootstrap)
    - \+ Moins d'hypothèses donc plus exact
    - \- Donne seulement une valeur
    - \- Gourmand en temps de calcul
  - Expression analytique de la propagation d'erreur
    - \- Plus d'hypothèses donc moins exact
    - \+ Donne une décomposition de l'incertitude

Introduction

- Il existe deux types de méthodes pour estimer l'incertitude de prédiction
  - Ré-échantillonage (p.ex. bootstrap)
    - + Moins d'hypothèses donc plus exact
    - - Donne seulement une valeur
    - - Gourmand en temps de calcul
  - Expression analytique de la propagation d'erreur
    - - Plus d'hypothèses donc moins exact
    - + Donne une décomposition de l'incertitude

Théorie

Supposons qu'un modèle a été étalonné, sur N échantillons :

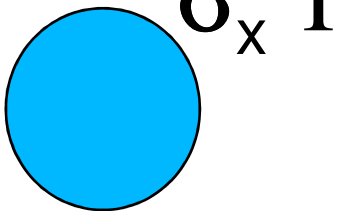


noté : $\hat{y} = \hat{\mathbf{z}}^{\mathrm{T}} \hat{\mathbf{b}} + \hat{y}_c$

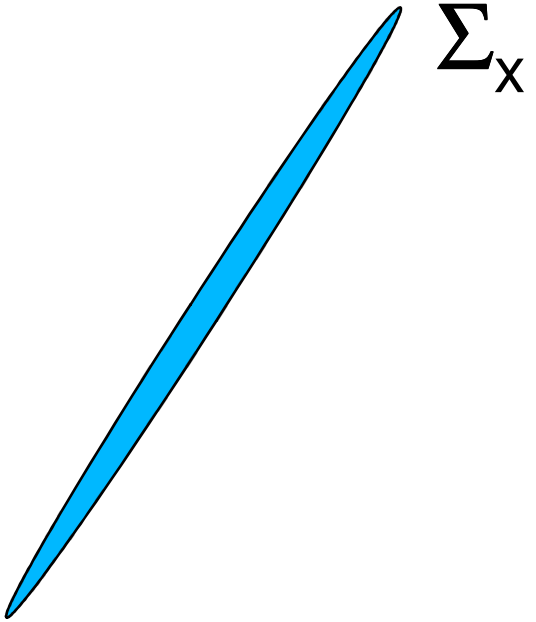
Question : $\mathrm{Var}(\hat{y}) = \mathrm{F}(\delta\mathbf{x}, \sigma^2_{LAB}, N, \ldots)$

Comment l'erreur de mesure se propage-t-elle dans le modèle ?

Deux hypothèses extrêmes de représentation de l'espace dans lequel évolue δx



Si on suppose que les variables de x sont indépendantes et identiquement distribuées (i.i.d.)



Si on tient compte de la corrélation entre les variables de x

Théorie

- L'expression classique suppose un δx i.i.d. et néglige les seconds ordres

$$\mathrm{Var}\left(\hat{y}\right) = \left(1 + \frac{1}{N}\right)\left\|\mathbf{b}\right\|^2 \sigma_x^2 + \mathbf{z}^T \boldsymbol{\Sigma}_b \mathbf{z} + \frac{\sigma_{LAB}^2}{N}$$

- La nouvelle expression relâche cette hypothèse :

$$\mathrm{Var}\left(\hat{y}\right) = \left(1 + \frac{1}{N}\right)\mathbf{b}^T \boldsymbol{\Sigma}_x \mathbf{b} + \mathbf{z}^T \boldsymbol{\Sigma}_b \mathbf{z} + \frac{\sigma_{LAB}^2}{N} + \left(1 + \frac{1}{N}\right)\mathrm{tr}\left(\boldsymbol{\Sigma}_x \boldsymbol{\Sigma}_b\right)$$

Théorie : Terme 1

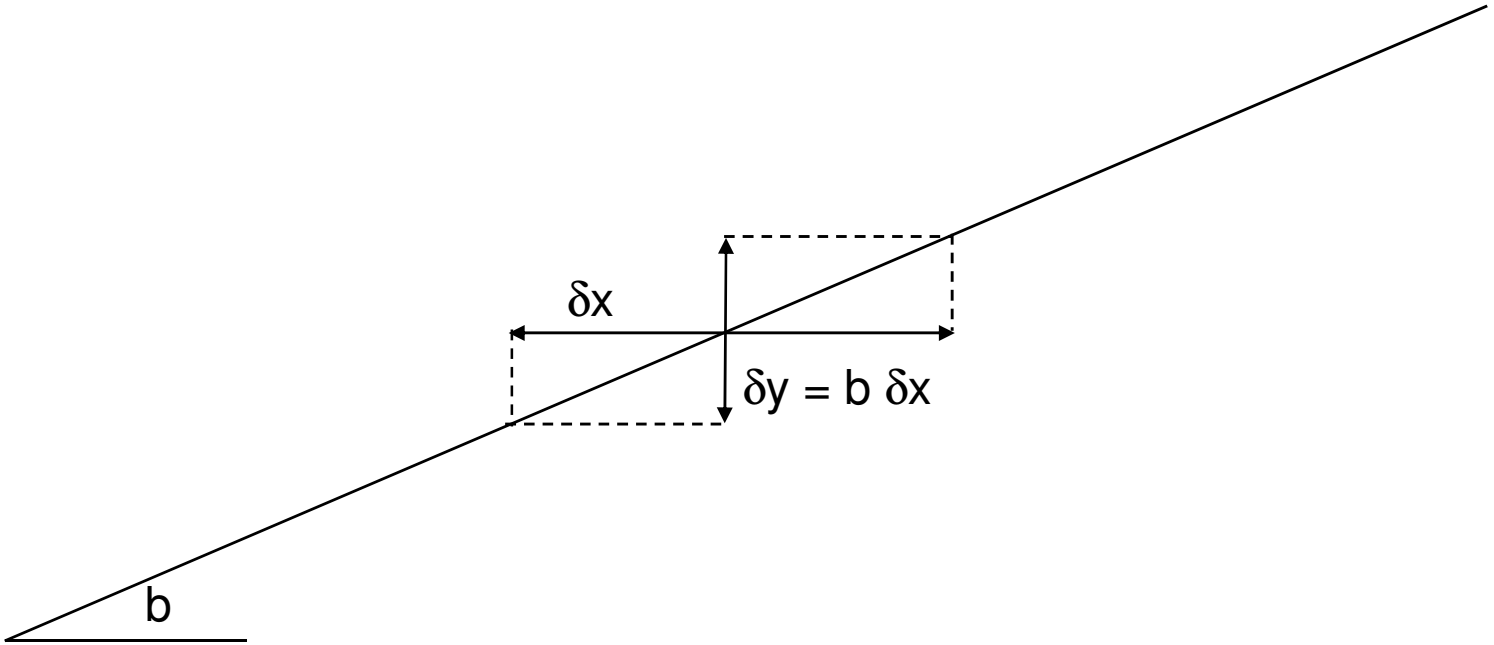
$$\mathrm{Var}(\hat{y}) = \left(1+\frac{1}{N}\right)\|\mathbf{b}\|^2\sigma_x^2 + \mathbf{z}^T\boldsymbol{\Sigma}_b\mathbf{z} + \frac{\sigma_{LAB}^2}{N}$$

Traduit l'amplification de l'erreur de mesure par le modèle

- L'expression classique ne tient pas compte de la structure de l'erreur et peut de ce fait être très pessimiste
- L'expression nouvelle tient compte de la structure de l'erreur et doit fournir une estimation plus réaliste

$$\mathrm{Var}(\hat{y}) = \left(1+\frac{1}{N}\right)\mathbf{b}^T\boldsymbol{\Sigma}_x\mathbf{b} + \mathbf{z}^T\boldsymbol{\Sigma}_b\mathbf{z} + \frac{\sigma_{LAB}^2}{N} + \left(1+\frac{1}{N}\right)\mathrm{tr}\left(\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_b\right)$$

Vue simplifiée en dimension 1



Théorie : Terme 2

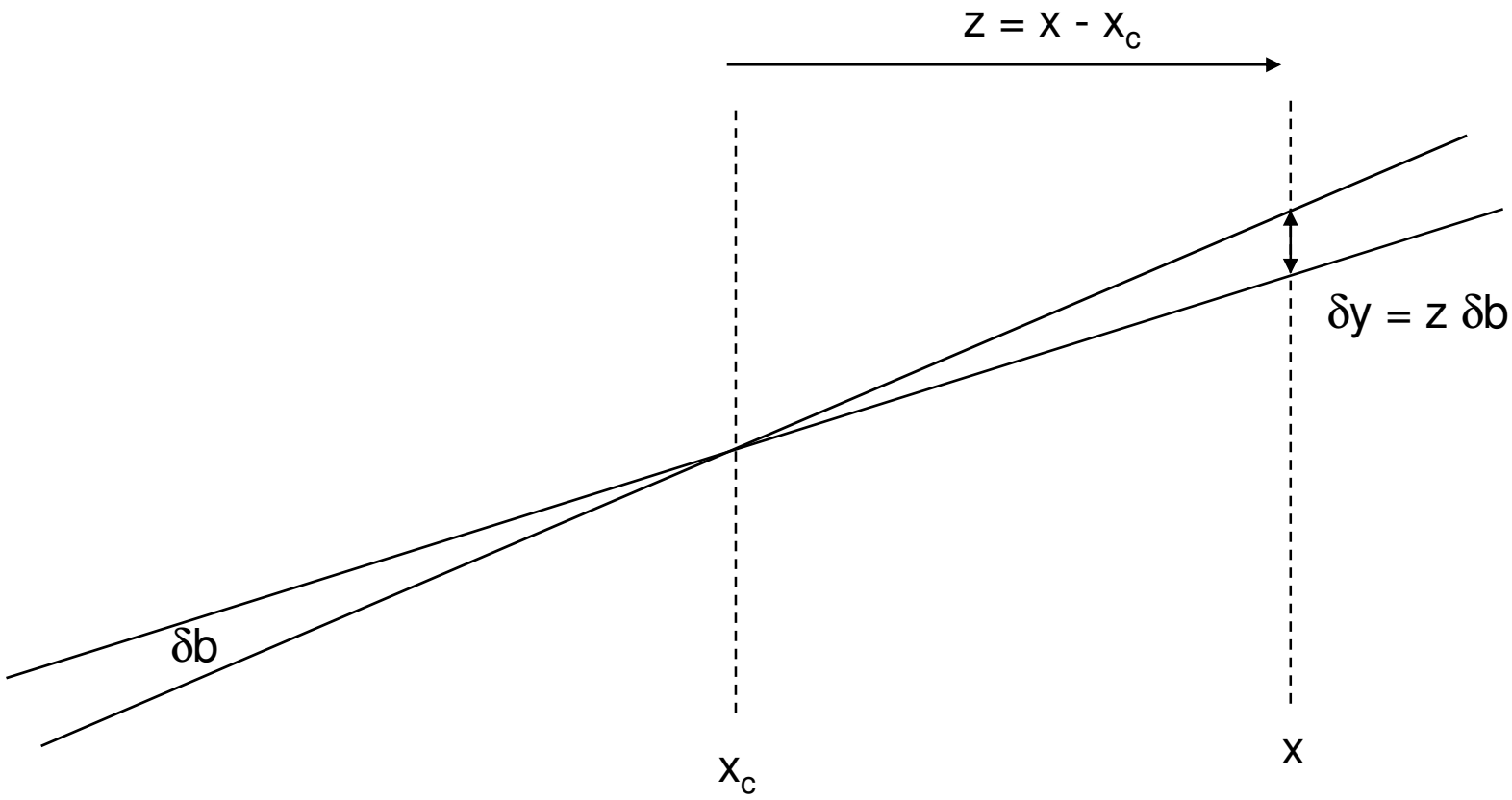
$$\mathrm{Var}(\hat{y}) = \left(1+\frac{1}{N}\right)\|\mathbf{b}\|^2\sigma_x^2 + \mathbf{z}^T\boldsymbol{\Sigma}_b\mathbf{z} + \frac{\sigma_{LAB}^2}{N}$$

Amplification de l'incertitude du modèle par le spectre

- Identique dans les deux expressions
- Relié à la notion de « leverage » : plus l'individu est loin du centre, plus il sera sensible à l'incertitude du modèle

$$\mathrm{Var}(\hat{y}) = \left(1+\frac{1}{N}\right)\mathbf{b}^T\boldsymbol{\Sigma}_x\mathbf{b} + \mathbf{z}^T\boldsymbol{\Sigma}_b\mathbf{z} + \frac{\sigma_{LAB}^2}{N} + \left(1+\frac{1}{N}\right)\mathrm{tr}(\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_b)$$

Vue simplifiée en dimension 1



Théorie : Terme 3

$$\mathrm{Var}(\hat{y}) = \left(1+\frac{1}{N}\right)\|\mathbf{b}\|^2\sigma_x^2 + \mathbf{z}^T\boldsymbol{\Sigma}_b\mathbf{z} + \frac{\sigma_{LAB}^2}{N}$$

Influence directe de la variance labo (SEL)

- Identique dans les deux expressions
- SEL peut agir aussi, à travers **b**, dans les autres termes.

$$\mathrm{Var}(\hat{y}) = \left(1+\frac{1}{N}\right)\mathbf{b}^T\boldsymbol{\Sigma}_x\mathbf{b} + \mathbf{z}^T\boldsymbol{\Sigma}_b\mathbf{z} + \frac{\sigma_{LAB}^2}{N} + \left(1+\frac{1}{N}\right)\mathrm{tr}\left(\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_b\right)$$

Théorie : Terme 4

$$\mathrm{Var}(\hat{y}) = \left(1+\frac{1}{N}\right)\|\mathbf{b}\|^2\sigma_x^2 + \mathbf{z}^T\boldsymbol{\Sigma}_b\mathbf{z} + \frac{\sigma_{LAB}^2}{N}$$

Ce terme est relié à la partie commune des espaces d'incertitude de x et de b

- Présent seulement dans la nouvelle expression
- Vient d'un second ordre : $\mathbf{Var}(\delta\mathbf{z}^T\delta\mathbf{b})$

$$\mathrm{Var}(\hat{y}) = \left(1+\frac{1}{N}\right)\mathbf{b}^T\boldsymbol{\Sigma}_x\mathbf{b} + \mathbf{z}^T\boldsymbol{\Sigma}_b\mathbf{z} + \frac{\sigma_{LAB}^2}{N} + \left(1+\frac{1}{N}\right)\mathrm{tr}\left(\boldsymbol{\Sigma}_x\boldsymbol{\Sigma}_b\right)$$

Matériel et méthodes

Data:

- Pour 385 échantillons d'aliments pour bétail
  - 1 spectre PIR (moyenne de 10 répétitions)
  - 1 valeur de taux de protéines (% massique)

Modèle:

- PLS associée à différents prétraitements
  - Spectres bruts, Detrend, Normalisation (SNV)
  - Dérivée seconde (Sav. & Gol.)

L'incertitude « vraie » a été estimée par bootstrap

Matériel et méthodes

- Les erreurs spectrales (δx) suivantes sont considérées :
  - Pour chaque spectre x, un spectre idéal x* a été calculé comme la moyenne de ceux ayant un y voisin
  - L'erreur spectrale a été calculée comme la différence entre les deux : $\delta x = x - x^*$

- Il s'agit d'une erreur spécifique à chaque échantillon, irréductible et responsable pour partie des résidus

- L'incertitude du modèle (Σ_b) a été estimée par un processus de cross-validation

Résultats

	$\mathrm{Var}(\hat{y})$	Raw 17 LV	Detrend 16 LV	2nd Der 18 LV	SNV 16 LV	2nd+SNV 15 LV	SNV+2nd 14 LV
Classical	Term 1	1200	69	46	260	13	21
	Term 2	0.13	0.14	0.15	0.16	0.12	0.14
	SUM	**1200**	**69**	**46**	**260**	**13**	**21**
New	Term 1	1.90	1.74	1.61	2.31	1.95	1.92
	Term 2	0.13	0.14	0.15	0.16	0.12	0.14
	Term 4	0.26	0.28	0.34	0.33	0.30	0.38
	SUM	**2.28**	**2.16**	**2.10**	**2.79**	**2.36**	**2.44**
	MSEC	1.67	1.65	1.40	1.81	1.40	1.60
	True (BS)	**2.22**	**2.24**	**2.17**	**2.57**	**2.07**	**2.47**

Résultats

	$\mathrm{Var}(\hat{y})$	Raw 17 LV	Detrend 16 LV	2nd Der 18 LV	SNV 16 LV	2nd+SNV 15 LV	SNV+2nd 14 LV
Classical	Term 1	1200	69	46	260	13	21
	Term 2	0.13	0.14	0.15	0.16	0.12	0.14
	SUM						**21**
New	Term 1						1.92
	Term 2						0.14
	Term 4						0.38
	SUM	**2.28**	**2.16**	**2.10**	**2.79**	**2.36**	**2.44**
	MSEC	1.67	1.65	1.40	1.81	1.40	1.60
	True (BS)	**2.22**	**2.24**	**2.17**	**2.57**	**2.07**	**2.47**

La variance « vraie » est très stable

Les prétraitements influent peu
Peut être inutiles dans ce cas ?

Résultats

	$\mathrm{Var}(\hat{y})$	Raw 17 LV	Detrend 16 LV	2nd Der 18 LV	SNV 16 LV	2nd+SNV 15 LV	SNV+2nd 14 LV
Classical	Term 1	1200	69	46	260	13	21
	Term 2	0.13	0.14	0.15	0.16	0.12	0.14
	SUM	**1200**	**69**	**46**	**260**	**13**	**21**
New	Term 1						
	Term 2						
	Term 4	0.26		.34	0.33	0.30	0.38
	SUM	**2.28**	**2.16**	**10**	**2.79**	**2.36**	**2.44**
	MSEC	1.67	1.65	1.40	1.81	1.40	1.60
	True (BS)	**2.22**	**2.24**	**2.17**	**2.57**	**2.07**	**2.47**

Les évolutions de la variance « vraie » et du MSEC sont cohérentes

Résultats

	$\mathrm{Var}(\hat{y})$	Raw 17 LV	Detrend 16 LV	2nd Der 18 LV	SNV 16 LV	2nd+SNV 15 LV	SNV+2nd 14 LV
Classical	Term 1						21
	Term 2						0.14
	SUM						**21**
New	Term 1	1.90	1.74	1.61	.31	1.95	1.92
	Term 2	0.13	0.14	0.1	0.16	0.12	0.14
	Term 4	0.26	0.28	0.34	0.33	0.30	0.38
	SUM	**2.28**	**2.16**	**2.10**	**2.79**	**2.36**	**2.44**
	MSEC	1.67	1.65	1.40	1.81	1.40	1.60
	True (BS)	**2.22**	**2.24**	**2.17**	**2.57**	**2.07**	**2.47**

La nouvelle expression donne des résultats proches de la variance « vraie »

Résultats

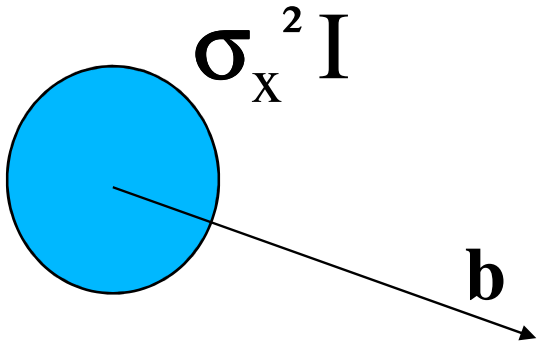
	$\mathrm{Var}(\hat{y})$	Raw 17 LV	Detrend 16 LV	2nd Der 18 LV	SNV 16 LV	2nd+SNV 15 LV	SNV+2nd 14 LV
Classical	Term 1	1200	69	46	260	13	21
	Term 2	0.13	0.14	0.15	0.16	0.12	0.14
	SUM	**1200**	**69**	**46**	**260**	**13**	**21**
New	Term 1	1.90	1.74	1.61			1.92

L'expression classique surestime la variance et s'avère très sensible au prétraitement

À cause de l'hypothèse d'erreur i.i.d. , seule l'intensité de l'erreur agit, pas sa structure

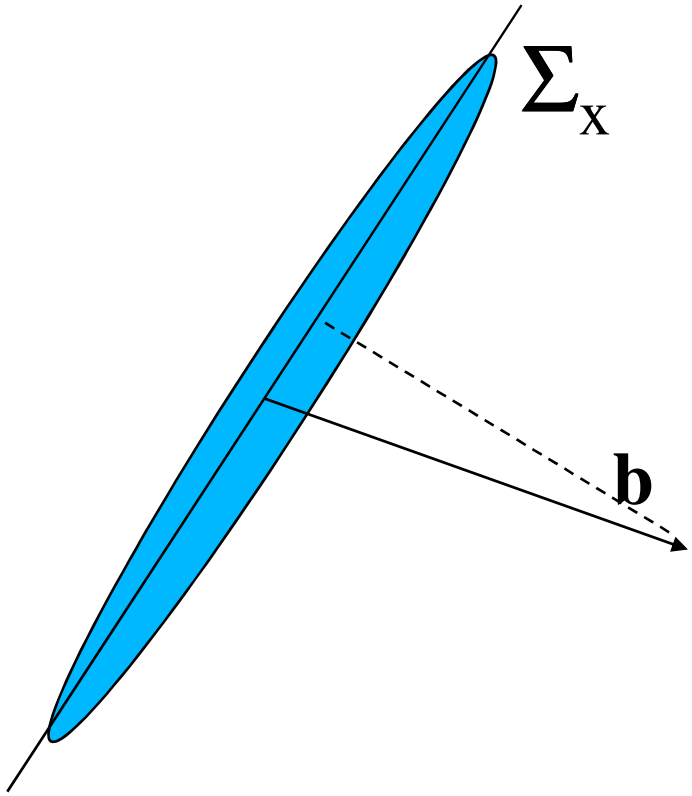
La capacité du modèle à s'auto-orthogonaliser n'est pas prise en compte

Interprétation géométrique



Erreur i.i.d. :

b ne peut pas être orthogonal à δx.



Erreur structurée :

b peut être (quasi)orthogonal à δx,
ce qui réduit le terme 1

Résultats

	$\mathrm{Var}(\hat{y})$	Raw 17 LV	Detrend 16 LV	2nd Der 18 LV	SNV 16 LV	2nd+SNV 15 LV	SNV+2nd 14 LV
NeW	Term 1	1.90	1.74	1.61	2.31	1.95	1.92
	Term 2	0.13	0.14	0.15	0.16	0.12	0.14
	Term 4	0.26	0.28	0.34	0.33	0.30	0.38
	SUM	**2.28**	**2.16**	**2.10**	**2.79**	**2.36**	**2.44**
	MSEC	1.67	1.65	1.40	1.81	1.40	1.60
	True (BS)	**2.22**	**2.24**	**2.17**	**2.57**	**2.07**	**2.47**

Le terme additionnel (terme 4) n'est pas négligeable
Ici, de 14% à 21% de la variance totale

...ultats

Interprétation des termes :

- Si SNV est appliqué en 1er, un effet non linéaire est ajouté, qui affecte Σ_x et Σ_b.
- Tous les termes augmentent

				2nd Der 18 LV	SNV 16 LV	2nd+SNV 15 LV	SNV+2nd 14 LV
				46	260	13	21
				0.15	0.16	0.12	0.14
				46	**260**	**13**	**21**
New	Term 1	1.90	1.74	1.61	2.31	1.95	1.92
	Term 2	0.13	0.14	0.15	0.16	0.12	0.14
	Term 4	0.26	0.28	0.34	0.33	0.30	0.38
	SUM	**2.28**	**2.16**	**2.10**	**2.79**	**2.36**	**2.44**
	MSEC	1.67	1.65	1.40	1.81	1.40	1.60
	True (BS)	**2.22**	**2.24**	**2.17**	**2.57**	**2.07**	**2.47**

sultats

La dérivée 2nde donne la plus petite incertitude, car elle supprime une erreur additive (une ligne de base)

				2nd Der 18 LV	SNV 16 LV	2nd+SNV 15 LV	SNV+2nd 14 LV
Classical	Term 1	1200		46	260	13	21
	Term 2	0.13	0.14	0.15	0.16	0.12	0.14
	SUM	**1200**	**69**	**46**	**260**	**13**	**21**
New	Term 1	1.90	1.74	1.61	2.31	1.95	1.92
	Term 2	0.13	0.14	0.15	0.16	0.12	0.14
	Term 4	0.26	0.28	0.34	0.33	0.30	0.38
	SUM	**2.28**	**2.16**	**2.10**	**2.79**	**2.36**	**2.44**

- Σ_x devient moins important, ce qui réduit le terme 1
- Mais le modèle est plus complexe (+2LV) et probablement moins stable, ce qui augmente les termes 2 et 4

Résultats

	$\mathrm{Var}(\hat{y})$	Raw 17 LV	Detrend	2nd Der	SNV	2nd SNV	SNV 2nd
Classical	Term 1	1200					
	Term 2	0.13					
	SUM	**1200**	**69**	**46**	**260**	**13**	**21**
New	Term 1	1.90	1.74	1.61	2.31	1.95	1.92
	Term 2	0.13	0.14	0.15	0.16	0.12	0.14
	Term 4	0.26	0.28	0.34	0.33	0.30	0.38
	SUM	**2.28**	**2.16**				
	MSEC	1.67	1.65				
	True (BS)	**2.22**	**2.24**				

Raw et Detrend produisent des modèles très stables, au regard des termes 2 and 4

Le terme 4 est très bas parce que le modèle est simple.
Le sur-apprentissage est faible ; il y a peu de dépendance entre Σ_x et Σ_b.

Conclusion

- Une nouvelle expression de l'incertitude des prédictions a été proposée
- Elle fournit des résultats proches du bootstrap
- Elle peut fournir des informations sur les différentes sources d'incertitude
- De nouveaux indices de qualité des modèles pourraient en être dérivés

Merci de votre attention