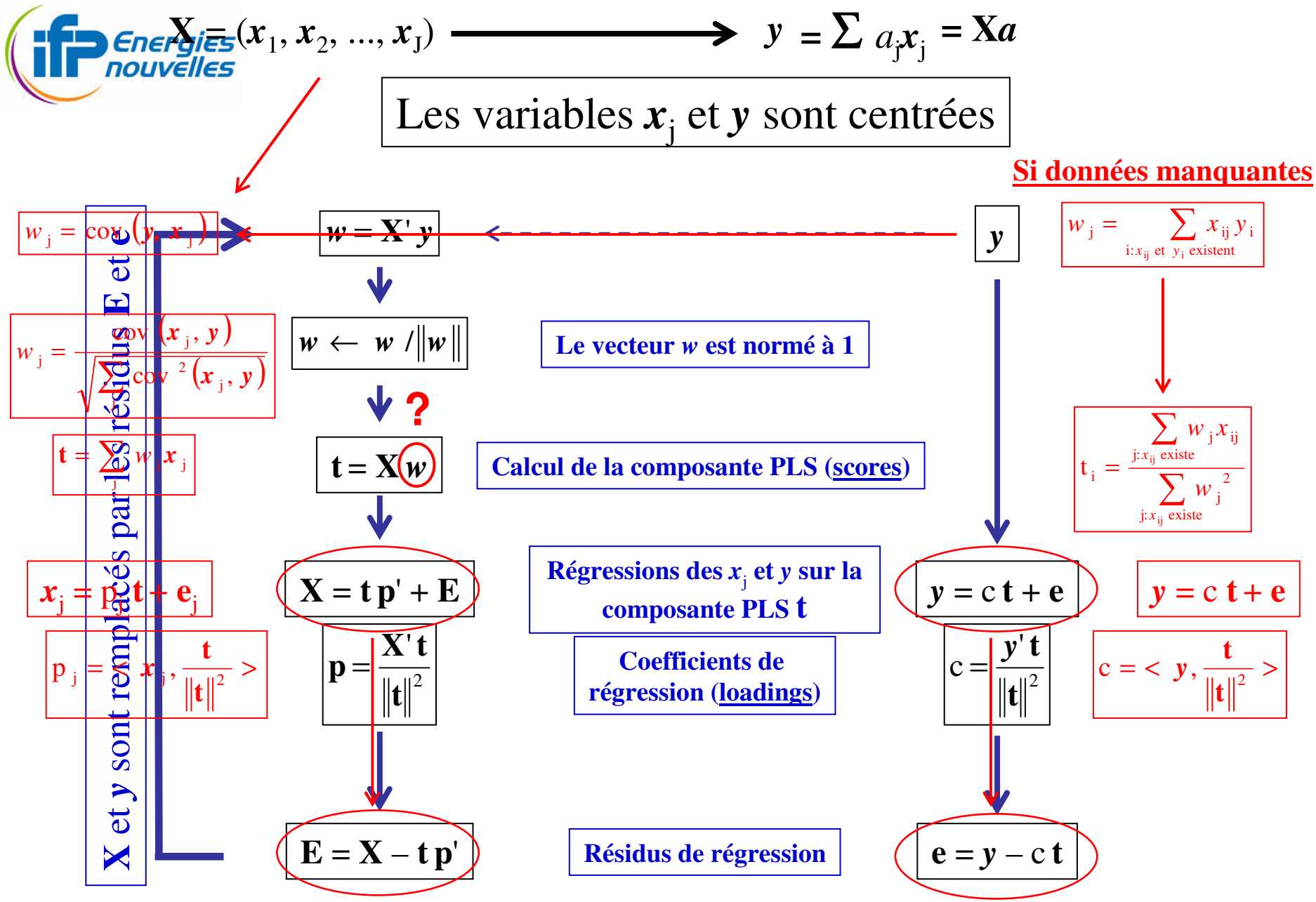
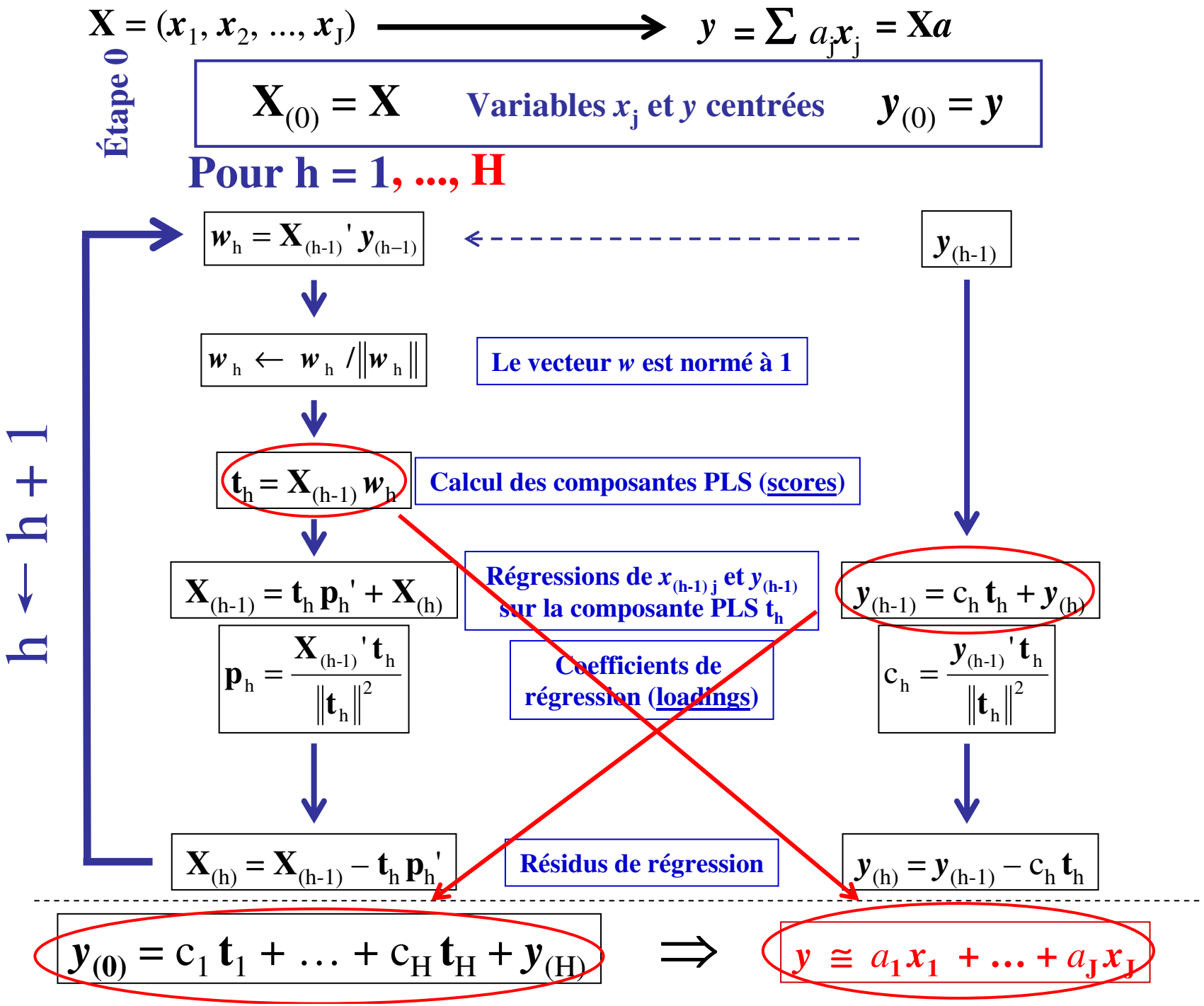
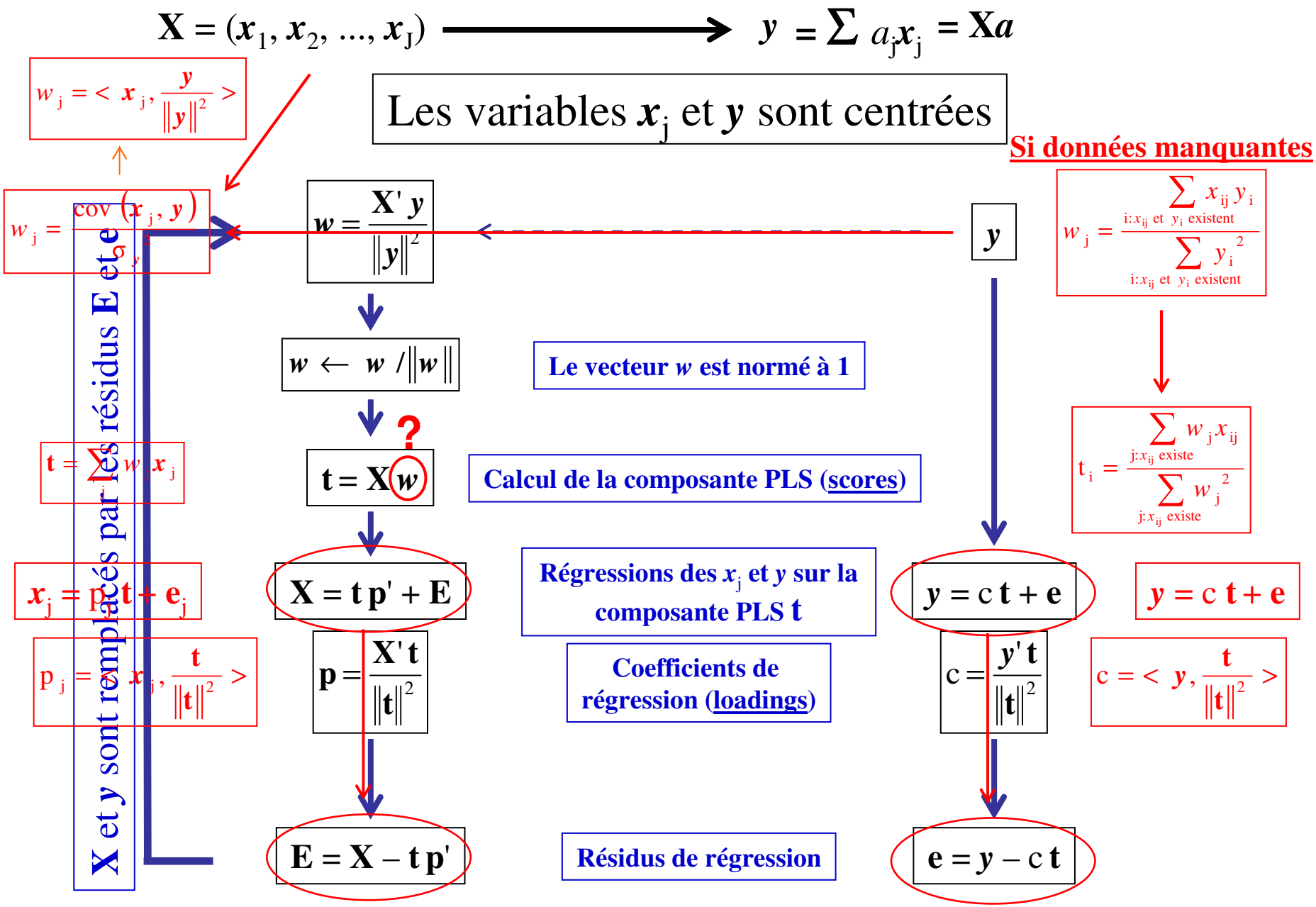
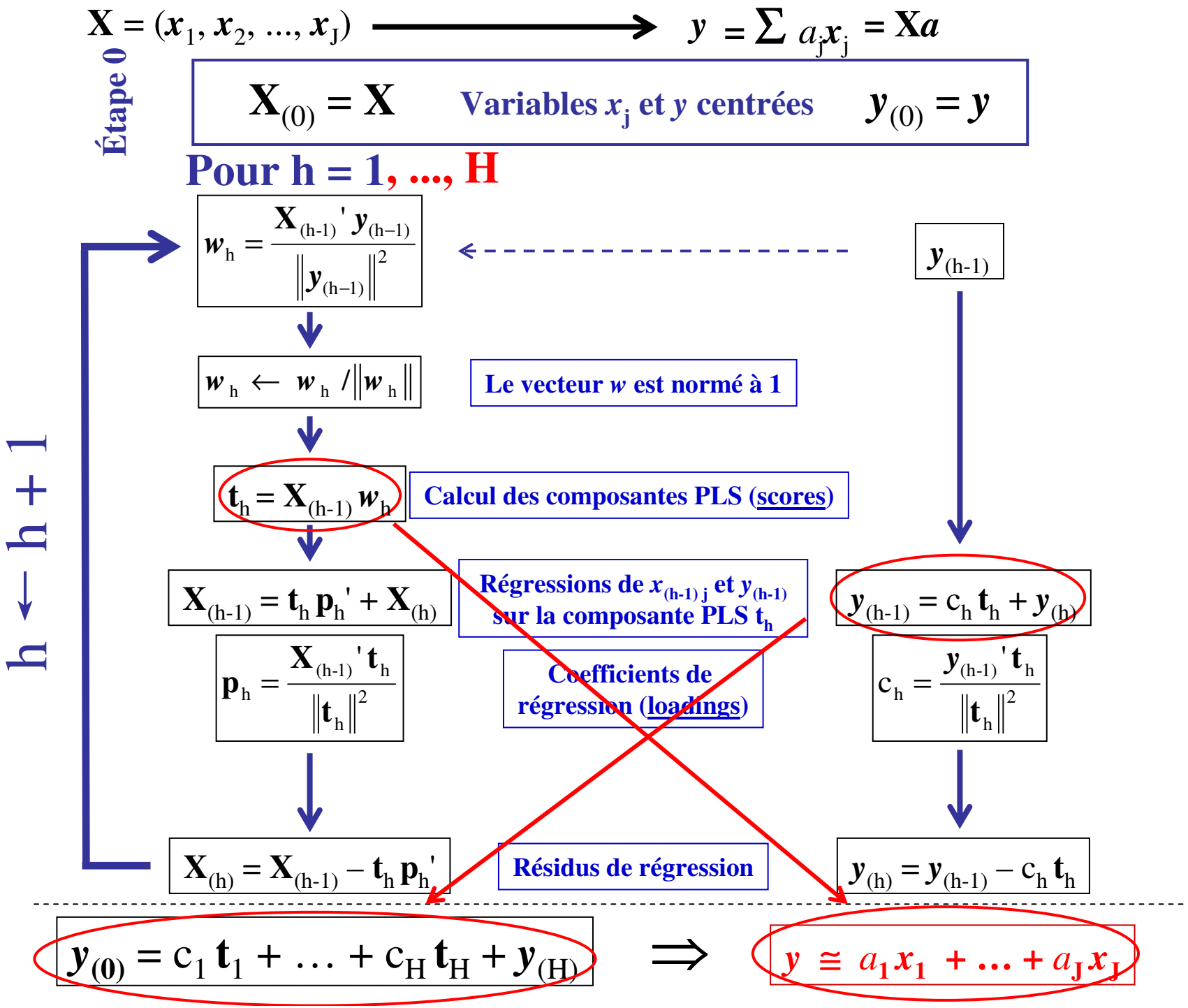
Algorithme PLS1


$X = (x_1, x_2, \ldots, x_J) \longrightarrow y = \sum a_j x_j = Xa$
Les variables x_j et y sont centrées
Si données manquantes
$w_j = \mathrm{cov}(y, x_j)$
$w = X'y$
y
$w_j = \sum_{i:x_{ij} \text{ et } y_i \text{ existent}} x_{ij} y_i$
$w_j = \frac{\mathrm{cov}(x_j, y)}{\sqrt{\sum \mathrm{cov}^2(x_j, y)}}$
$w \leftarrow w / \|w\|$
Le vecteur w est normé à 1
?
$t = \sum w_j x_j$
$t = Xw$
Calcul de la composante PLS (scores)
$t_i = \frac{\sum_{j:x_{ij} \text{ existe}} w_j x_{ij}}{\sum_{j:x_{ij} \text{ existe}} w_j^2}$
$x_j = p_j t + e_j$
$X = t p' + E$
Régressions des x_j et y sur la composante PLS t
$y = c\,t + e$
$y = c\,t + e$
$p_j = < x_j, \frac{t}{\|t\|^2} >$
$p = \frac{X't}{\|t\|^2}$
Coefficients de régression (loadings)
$c = \frac{y't}{\|t\|^2}$
$c = < y, \frac{t}{\|t\|^2} >$
X et y sont remplacés par les résidus E et e
$E = X - t p'$
Résidus de régression
$e = y - c\,t$


$\mathbf{X} = (x_1, x_2, \ldots, x_J)$
$y = \sum a_j x_j = \mathbf{X}a$
Étape 0
$\mathbf{X}_{(0)} = \mathbf{X}$ Variables x_j et y centrées $y_{(0)} = y$
Pour h = 1, ..., H
$w_h = \mathbf{X}_{(h-1)}' y_{(h-1)}$
$y_{(h-1)}$
$w_h \leftarrow w_h / \|w_h\|$
Le vecteur w est normé à 1
$\mathbf{t}_h = \mathbf{X}_{(h-1)} w_h$
Calcul des composantes PLS (scores)
$\mathbf{X}_{(h-1)} = \mathbf{t}_h \mathbf{p}_h' + \mathbf{X}_{(h)}$
Régressions de $x_{(h-1)\,j}$ et $y_{(h-1)}$ sur la composante PLS $\mathbf{t}_h$
$y_{(h-1)} = c_h \mathbf{t}_h + y_{(h)}$
$\mathbf{p}_h = \frac{\mathbf{X}_{(h-1)}' \mathbf{t}_h}{\|\mathbf{t}_h\|^2}$
Coefficients de régression (loadings)
$c_h = \frac{y_{(h-1)}' \mathbf{t}_h}{\|\mathbf{t}_h\|^2}$
$\mathbf{X}_{(h)} = \mathbf{X}_{(h-1)} - \mathbf{t}_h \mathbf{p}_h'$
Résidus de régression
$y_{(h)} = y_{(h-1)} - c_h \mathbf{t}_h$
h ← h + 1
$y_{(0)} = c_1 \mathbf{t}_1 + \ldots + c_H \mathbf{t}_H + y_{(H)}$
$\Rightarrow$
$y \cong a_1 x_1 + \ldots + a_J x_J$

Algorithme PLS1 avec réduction de y

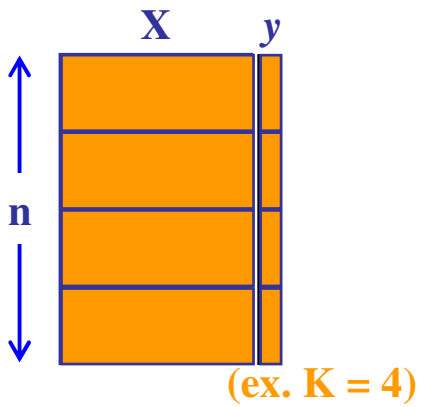

$\mathbf{X} = (x_1, x_2, \ldots, x_J) \longrightarrow y = \sum a_j x_j = \mathbf{X}a$
Les variables x_j et y sont centrées
Si données manquantes
$w_j = < x_j, \frac{y}{\|y\|^2} >$
$w_j = \frac{\mathrm{cov}(x_j, y)}{\sigma_y}$
$w = \frac{\mathbf{X}'y}{\|y\|^2}$
y
$w_j = \frac{\sum_{i: x_{ij} \text{ et } y_i \text{ existent}} x_{ij} y_i}{\sum_{i: x_{ij} \text{ et } y_i \text{ existent}} y_i^2}$
$w \leftarrow w / \|w\|$
Le vecteur w est normé à 1
$t = \sum w_j x_j$
$t = \mathbf{X}w$?
Calcul de la composante PLS (scores)
$t_i = \frac{\sum_{j: x_{ij} \text{ existe}} w_j x_{ij}}{\sum_{j: x_{ij} \text{ existe}} w_j^2}$
$x_j = p_j t + e_j$
$\mathbf{X} = \mathbf{t}\mathbf{p}' + \mathbf{E}$
Régressions des x_j et y sur la composante PLS $\mathbf{t}$
$y = c\,\mathbf{t} + \mathbf{e}$
$y = c\,\mathbf{t} + \mathbf{e}$
$p_j = < x_j, \frac{\mathbf{t}}{\|\mathbf{t}\|^2} >$
$\mathbf{p} = \frac{\mathbf{X}'\mathbf{t}}{\|\mathbf{t}\|^2}$
Coefficients de régression (loadings)
$c = \frac{y'\mathbf{t}}{\|\mathbf{t}\|^2}$
$c = < y, \frac{\mathbf{t}}{\|\mathbf{t}\|^2} >$
$\mathbf{E} = \mathbf{X} - \mathbf{t}\mathbf{p}'$
Résidus de régression
$\mathbf{e} = y - c\,\mathbf{t}$
$\mathbf{X}$ et y sont remplacés par les résidus $\mathbf{E}$ et $\mathbf{e}$


$\mathbf{X} = (x_1, x_2, \ldots, x_J)$
$y = \sum a_j x_j = \mathbf{X}a$
Étape 0
$\mathbf{X}_{(0)} = \mathbf{X}$ Variables x_j et y centrées $y_{(0)} = y$
Pour h = 1, ..., H
$w_h = \frac{\mathbf{X}_{(h-1)}' y_{(h-1)}}{\|y_{(h-1)}\|^2}$
$y_{(h-1)}$
$w_h \leftarrow w_h / \|w_h\|$
Le vecteur w est normé à 1
$\mathbf{t}_h = \mathbf{X}_{(h-1)} w_h$
Calcul des composantes PLS (scores)
$\mathbf{X}_{(h-1)} = \mathbf{t}_h \mathbf{p}_h' + \mathbf{X}_{(h)}$
Régressions de $x_{(h-1)\,j}$ et $y_{(h-1)}$ sur la composante PLS $\mathbf{t}_h$
$y_{(h-1)} = c_h \mathbf{t}_h + y_{(h)}$
$\mathbf{p}_h = \frac{\mathbf{X}_{(h-1)}' \mathbf{t}_h}{\|\mathbf{t}_h\|^2}$
Coefficients de régression (loadings)
$c_h = \frac{y_{(h-1)}' \mathbf{t}_h}{\|\mathbf{t}_h\|^2}$
$\mathbf{X}_{(h)} = \mathbf{X}_{(h-1)} - \mathbf{t}_h \mathbf{p}_h'$
Résidus de régression
$y_{(h)} = y_{(h-1)} - c_h \mathbf{t}_h$
h ← h + 1
$y_{(0)} = c_1 \mathbf{t}_1 + \ldots + c_H \mathbf{t}_H + y_{(H)}$
$\Rightarrow$
$y \cong a_1 x_1 + \ldots + a_J x_J$

Sélection du nombre de composantes PLS

Le nombre H de composantes PLS est déterminé par validation croisée (*leave-one-out* ou *K-folds*) à partir du minimum de la courbe du PRESS (PRediction Error Sum of Squares)

Données d'apprentissage

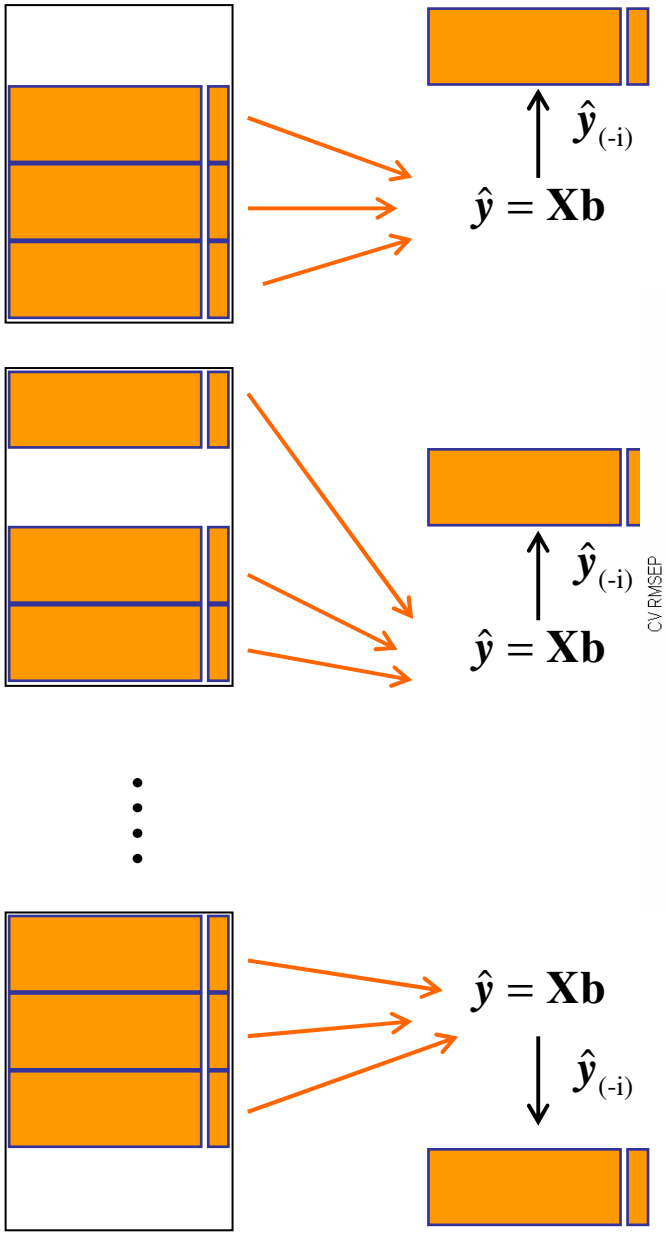


Validation croisée par *K-folds*

On se donne un nombre K

↓

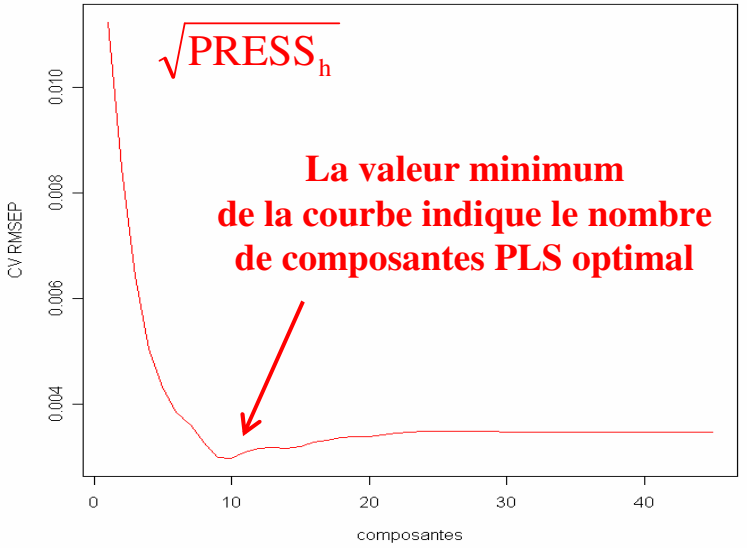
Partitionnement de l'ensemble des données en K ensembles



$$PRESS = \sum_i \left(y_i - \hat{y}_{(-i)}\right)^2$$

$$PRESS_h = \sum_i \left(y_i^{(h)} - \hat{y}_{(-i)}^{(h)}\right)^2$$

h = numéro de la composante PLS



K = n

↓

Leave-one-out

Avantages et limites

Avantages

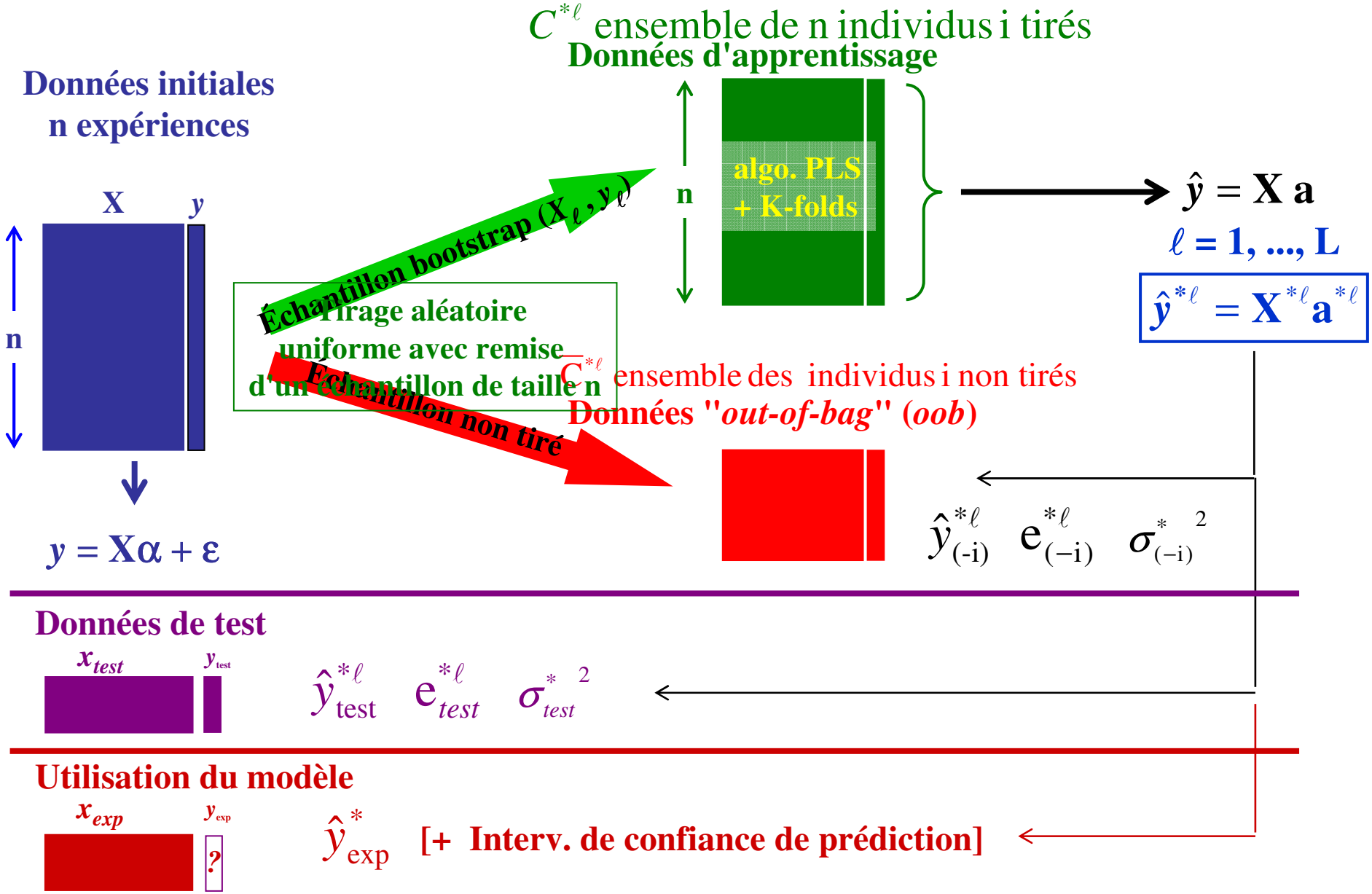
- Le nombre des expériences peut être inférieur à celui des variables explicatives $\boldsymbol{x}$
- Les coefficients du modèle respectent les signes (et intensités) des corrélations des variables explicatives $\boldsymbol{x}$ avec la variable à expliquer $\boldsymbol{y}$
  - Le coefficient $\boldsymbol{a}_j$ peut être interprété comme étant la contribution de $\boldsymbol{x}_j$ à la construction de $\boldsymbol{y}$
- Méthode multidimensionnelle descriptive
  - Représentation des individus par des points dans l'espace orthonormé des composantes PLS
  - Représentation des variables $\boldsymbol{x}$ et $\boldsymbol{y}$ en tant que vecteurs par leurs corrélations avec les composantes PLS (→ cercle de corrélations)
- On peut calculer :
  - $\mathbf{R}_y^2$: la part de variance de $\boldsymbol{y}$ expliquée par le modèle
  - $\boldsymbol{\sigma}_x^2$: variance explicative des $\boldsymbol{x}$ (i.e. du nuage de points dans l'espace des composantes PLS)

Limites

- Les calculs statistiques propres aux MC ne sont pas applicables à la régression PLS
  - Pas de tests statistiques sur le modèle ou sur le LOF (équivalents du test de Fisher)
  - Pas de tests statistiques sur les coefficients (équivalents du test de Student)
  - Pas d'intervalle de confiance pour :
    - les coefficients du modèle
    - les valeurs prédites par le modèle
- Pas de possibilité de réaliser directement des plans d'expériences pour la PLS

Algorithme PLS bootsrap pour la sélection de variables

Construction d'un modèle par apprentissage

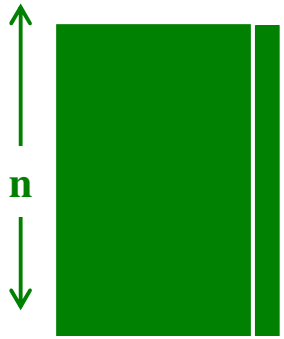


Données d'apprentissage

Pour $\ell = 1, \ldots,$ L bootstraps

$C^{*\ell}$ ensemble des n individus i tirés

Données d'apprentissage



Modèle PLS $\hat{\mathbf{y}}^{*\ell} = \mathbf{X}^{*\ell}\mathbf{a}^{*\ell}$

$$\text{coefficients } \mathbf{a}^{*\ell} = (a_1^{*\ell}, a_2^{*\ell}, \ldots, a_J^{*\ell})^T$$

Pour toute variable x_j on a $\left\{ a_j^{*\ell},\ \ell = 1, \ldots, L \right\}$

La variable x_j sera éliminée si *Prob* ($a_j = 0$) $\geq \alpha$

où α est un seuil fixé

Données *out-of-bag*

Pour $\ell = 1, \ldots, L$ bootstraps

$\overline{C}^{*\ell}$ ensemble des individus i non tirés

Données *out-of-bag* (*oob*)

Prédiction : $\hat{y}_{(-i)}^{*\ell} = (\mathbf{a}^{*\ell})^T \boldsymbol{x}_i^{*\ell}$

Erreur de prédiction : $e_{(-i)}^{*\ell} = \hat{y}_{(-i)}^{*\ell} - y_i$

Q^2_{oob} sur données oob : $Q_{oob}^{*\ell 2} = Cor^2(\hat{\boldsymbol{y}}^{*\ell}, \boldsymbol{y})$

Erreur quadratique moyenne de prédiction *out-of-bag* : $EQM_{oob}^{*\ell} = \sqrt{\frac{1}{|\overline{C}^{*\ell}|} \sum_{i \in \overline{C}^{*\ell}} (y_{(-i)}^{*\ell} - y_i)^2}$

Indicateurs globaux de la qualité de prédiction du modèle

Indicateur ponctuel (en chaque point i) de la qualité de prédiction du modèle

Pour une expérience i

Variance de prédiction

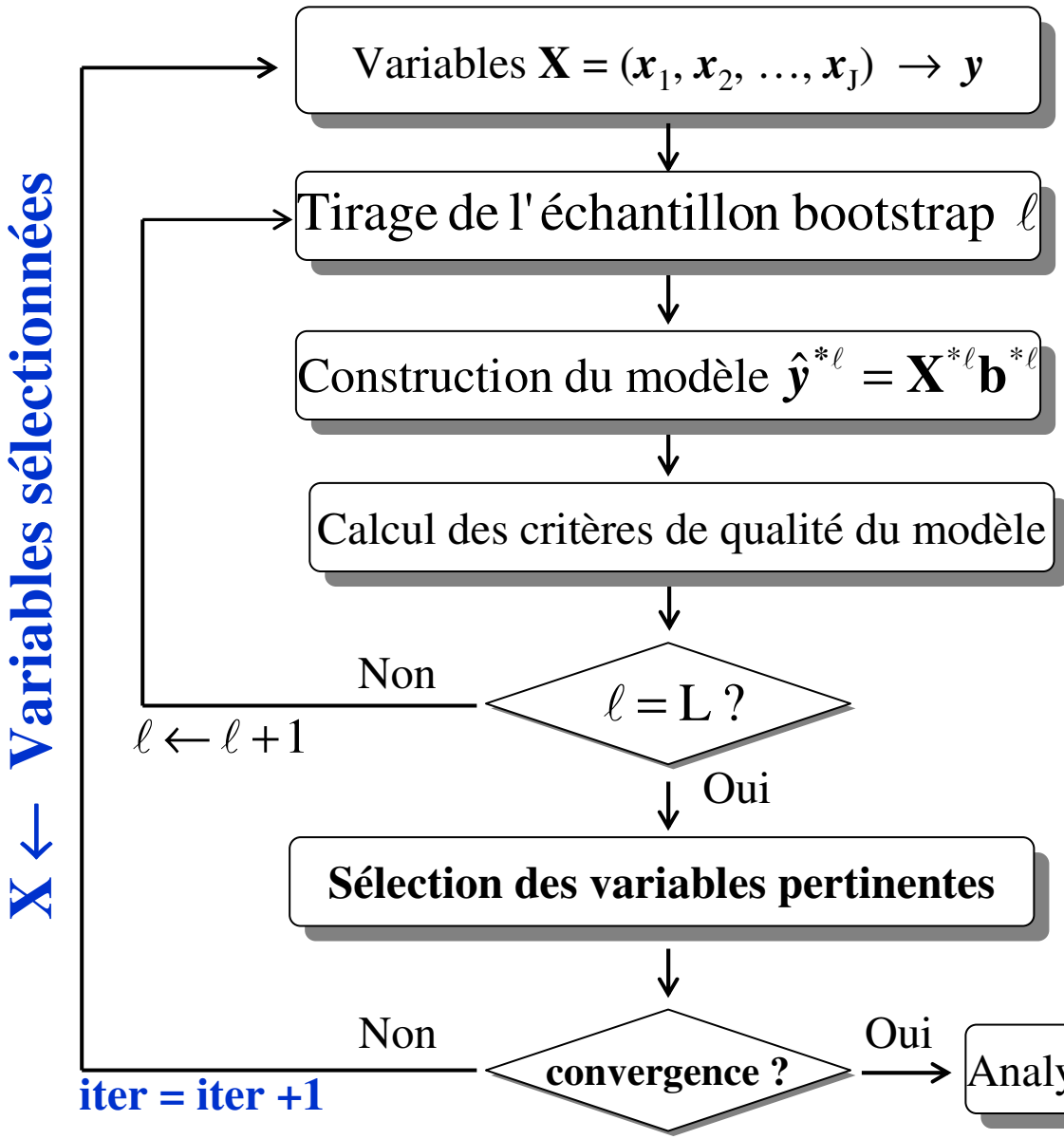
$$\sigma_i^{*2} = \frac{1}{L} \sum_{\ell=1}^{L} \left(y_i^{*\ell} - \bar{\hat{y}}_i \right)^2$$

Moyenne des prédictions en i

$$\bar{\hat{y}}_i = \frac{1}{L} \sum_{\ell=1}^{L} y_i^{*\ell}$$

Algorithme PLS-boostrap

L = nombre de bootstraps
$\alpha_{optimal}$ = seuil fixé



A chaque itération

Distributions relatives au modèle

$\{\mathbf{a}^{*\ell} ;\ \ell = 1,\ldots, L\}$

$\{Q_{oob}{}^{*\ell 2} ;\ \ell = 1,\ldots, L\}$

$\{EQM_{oob}{}^{*\ell} ;\ \ell = 1,\ldots, L\}$

Distributions relatives aux individus

$\{\hat{y}^{*\ell}_{(-i)} ;\ \ell = 1, \ldots, L\}$

$\{e^{*\ell}_{(-i)} ;\ \ell = 1, \ldots, L\}$

$variance : \sigma^{*2}_{(-i)}$

Composantes PLS $t_1, t_2, \ldots, t_H$

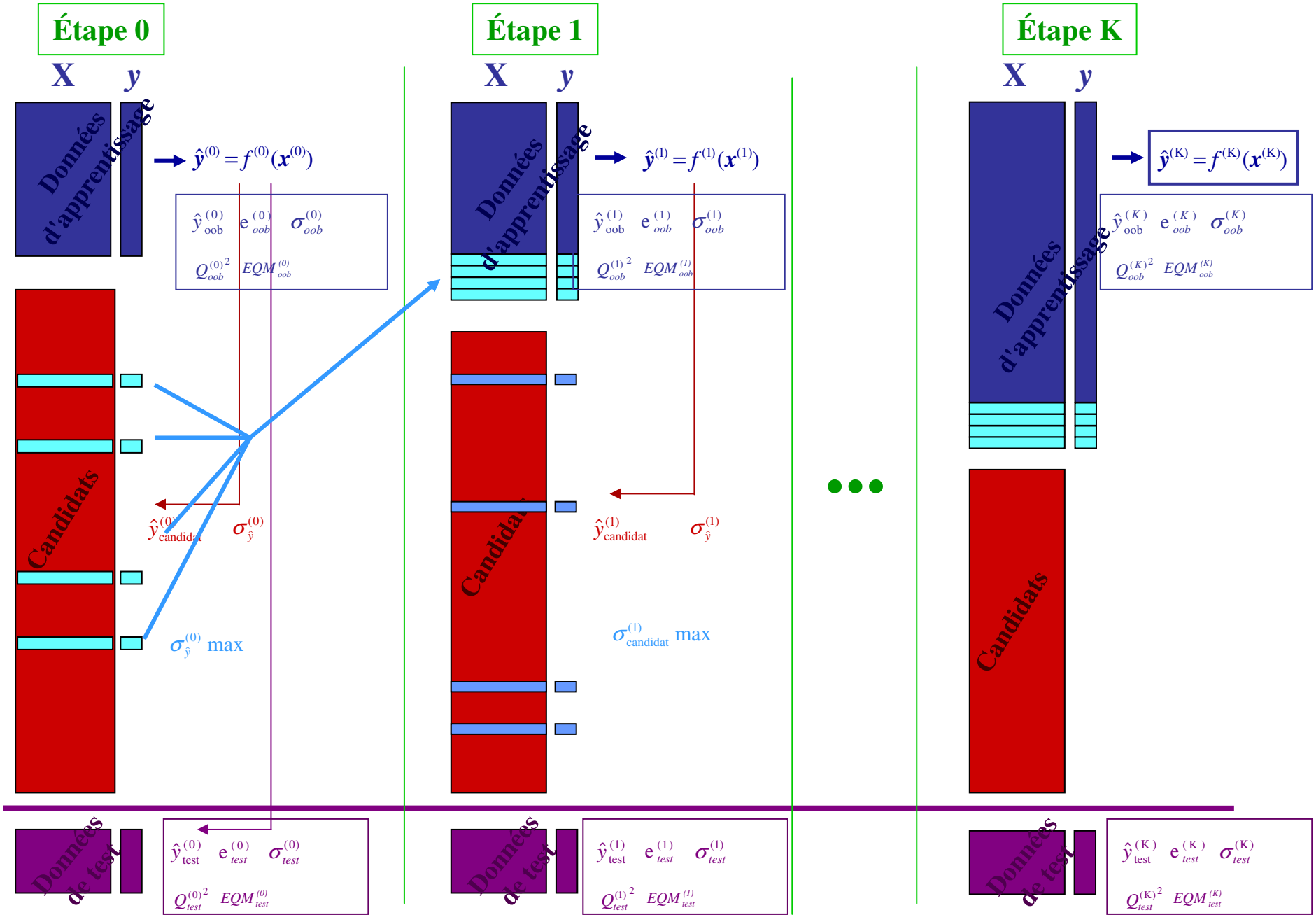
$\{t^{*\ell}_{ih} ;\ \ell = 1, \ldots, L\}$

Prédictions sur individus de test x_{test}

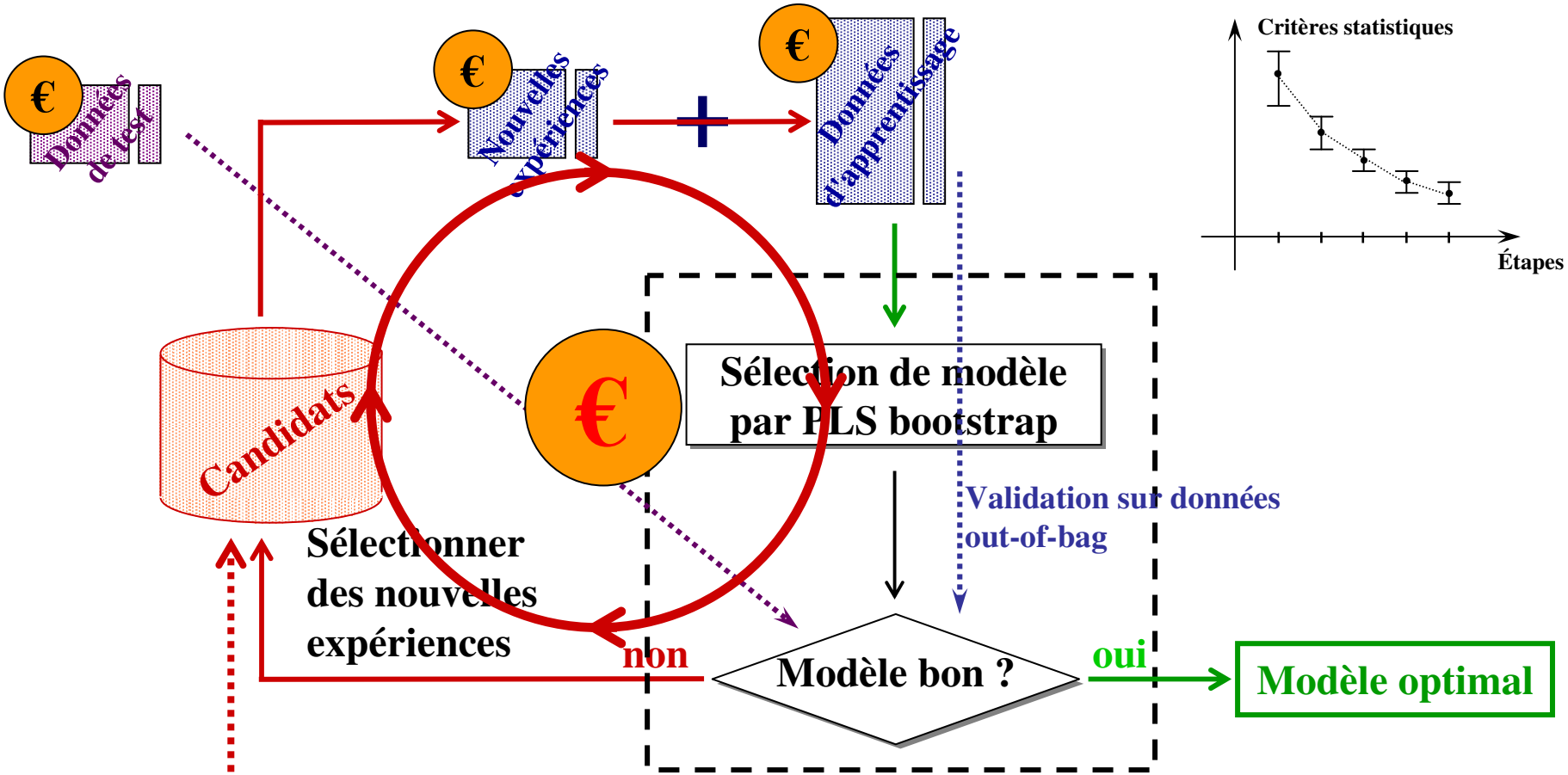
$\{\hat{y}^{*\ell}_{test} ;\ \ell = 1, \ldots, L\}$

Planification expérimentale adaptative pour régression PLS

Planification expérimentale adaptative pour régression PLS



Algorithme de planification adaptative



- Nouvelles expériences à réaliser
- BdD expérimentale
- Simulateur numérique

Références bibliographiques

REGRESSION PLS / Sélection de modèle pour régression PLS

- Wold H. (1966), Estimation of principal components and related models by iterative least squares. In P. Krishnaiah, editors. *Multivariate Analysis*, Academic Press, 391–420.
- Wold S., Esbensen K. and Geladi P. (1987), Principal component analysis *Chemometrics and Intelligent Laboratory Systems*, **2**, 37–52.
- Tenenhaus M. (1998), Ed. Technips, Paris
- Mevik B., Wehrens R. (2007), The pls Package: Principal Component and Partial Least Squares Regression in R, Journal of Statistical Software, January 2007, Volume 18, Issue 2., http://www.jstatsoft.org/
- Lazraq A., Cléroux R., Gauchy J.-P. (2003), Selecting both latent and explanatory variables in PLS1 regression model, Chemometrics and Intelligent Laboratory Systems 66 (2003) 117-126.
- Gauchi J.-P., Chagnon P. (2001), Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data, Chemometrics and Intelligent Laboratory Systems 58 (2001) 171-193.
- Faraj A., Noçairi H., Constant M. (2007), Sélection de modèle PLS par ré-échantillonnage bootstrap in *Classification : points de vue croisés*, Revue des Nouvelles Technologies de l'Information, RNTI-C-2, Ed. M. Nadif.
- Bastien P., Vinzi Esposito V., Tenehaus M. (2005), PLS generalised linear regression, Computational Statistics and Data Analysis 48 (2005) 17-46, Elsevier

PLANS D'EXPERIENCES POUR SURFACE DE REPONSE

- Droesbeke J.J., Fine J. et Saporta G. Éd (1997), Plans d'expériences : Applications à l'entreprise, Éditions Technip, Paris
- Lewis G.A., Mathieu D., Phan-Tan-Luu R.(1999). - Pharmaceutical experimental design, Marcel Dekker Inc., New York
- Myers R.H., Montgomery D.C. (2002), Response Surface Methodology : Process and Product Optimization Using Designed Experiments, J. Wiley Ed., N-Y

BOOTSTRAP

- Efron B., Tibshirani R. (1993), An introduction to the Bootstrap, Chapman and Hall, London.

Logiciels

- NEMROD-W (LPRAI) : http:// www.nemrodw.com
- Design Expert : http://www.statease.com/
- http://www.r-project.org/ : site du projet cran [cf. librairie pls : Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR)]