Risque et Choix de Modèle en Apprentissage Exemples

Philippe Besse

Université de Toulouse
INSA – Dpt GMM
Institut de Mathématiques – ESP
UMR CNRS 5219

Objectifs

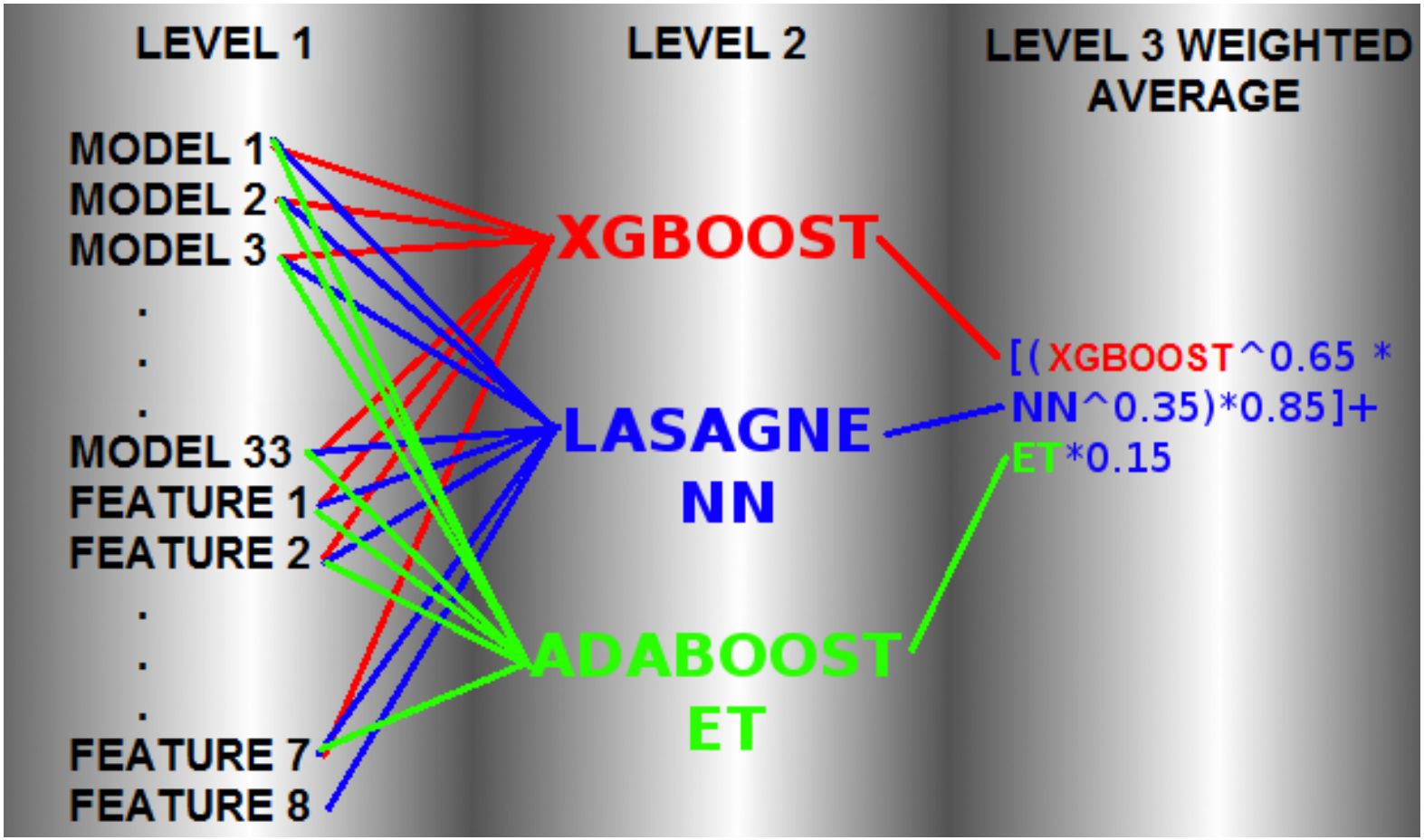
- Facteurs de risque épidémiologiques
- Facteur génétique ou biomarqueurs
- Reconnaissance de forme (caractères)
- Adaptation statistique en prévision météo (pic d'ozone)
- Score d'appétence ou d'attrition en GRC
- Méta modèle ou réduction de modèle physique
- Détection défaillance ou fraude (atypique)
- ...
- Minimiser une erreur de prévision ou risque

Statistique *vs.* Apprentissage Statistique *vs.* Machine

- Explorer ou vérifier, représenter, décrire
- Expliquer ou tester une influence
- Prévoir et sélectionner, interpréter
- Prévision "brute"

Pour quel but ?

- Publication académique (Benchmarks -UCI)
- Solution "industrielle" (ch. chapitre 12)
- Concours de type Kaggle



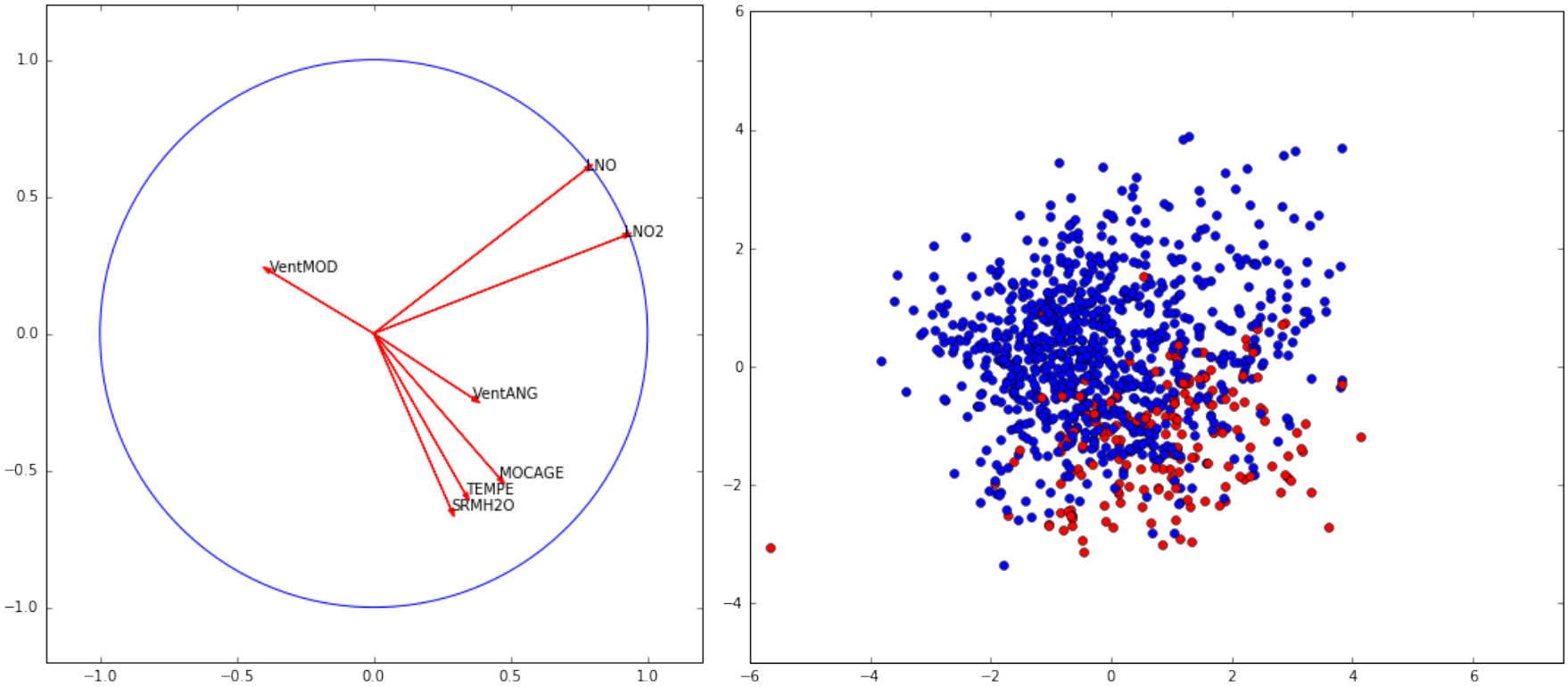
Concours Kaggle : Identify people who have a high degree of Psychopathy based on Twitter usage.

Stratégie de l'apprentissage

1. Extraction avec ou sans échantillonnage, SQL ou NO
2. Exploration, visualisation, nettoyage, transformations...
3. Données manquantes : suppression, imputation
4. Partition de l'échantillon : apprentissage, (validation), test)
5. Pour méthode in $\{k$-nn,rn, tree, RF, svm... $\}$
   - Estimation d'un modèle fonction de q (complexité)
   - Optimisation du (des) paramètres q (validation)
6. Comparaison des méthodes par erreur de prévision sur échantillon test ou meilleur compromis
7. Itération éventuelle (plusieurs échantillons test)
8. Choix de la méthode (prévision, interprétabilité).
9. ré-estimation du modèle, exploitation

Contenu

- Environnement technologique : R, Python, Notebooks
- Critères d'évaluation et leurs modes d'estimation
- Méthodes "oubliées" : modèle gaussien , binomial, analyse discriminante, k-nn
- Exemples illustratifs
  - Jouet : Nuages gaussiens
  - Adaptation statistique : pic d'ozone
    - Régression quantitative concentration O3
    - Discrimination et dépassement de seuil
    - MOCAGE, NO2, NO3, H2O, Température, vent, jour, station
  - Spectrométrie proche infra-rouge (NIR)
  - GRC : Appétence pour la carte Visa Premier
  - ...



Ozone : premier plan de l'ACP réduite (47%)

Logiciels de fouille de données

- Clementine (SPSS)
- Enterprise Miner (SAS)
- Insightfull Miner (Splus)
- KXEN, SPAD,
- Statistica Data Miner
- Statsoft, Tanagra, Weka
- ...

R *vs.* Python

- Langage R, librairies, `caret`
- Langage Python, `pandas, scikit-learn`
- Comparaison
  - Mémoire
  - Parallélisation
  - Classe `Data Frame`
  - Sélection de modèle linéaire général
  - Élagage d'un arbre

Reproductibilité

- Donoho, 2015
- Chaîne de traitements (*pipeline*) automatisée
- Production automatique de rapports
  - R : `sweave, knitr`
  - Python : `pweave`
- Tutoriels : Notebook IPython ou Jupyter (Python, R, Julia...)
- `https://github.com/wikistat/`

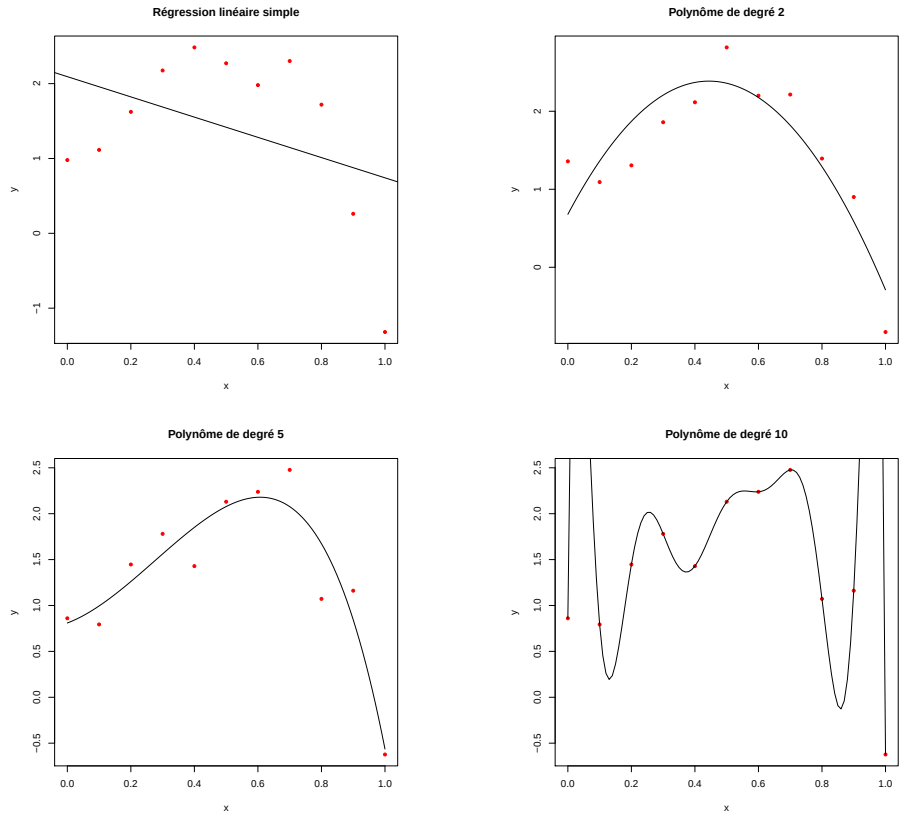
Cf. Notebook

Risque empirique ou qualité d'ajustement

- $\boldsymbol{D}_n$ observation d'un n-échantillon
  $\boldsymbol{D}_n = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\}$ de loi conjointe inconnue P
- $\boldsymbol{x}$ observation de la variable $\boldsymbol{X}$
- $\boldsymbol{D}_n$ est appelé *échantillon d'apprentissage*
- Une *règle de prévision* (ou prédicteur) est une fonction (mesurable) $f : \mathcal{X} \to \mathcal{Y}$, $\boldsymbol{x} \to f(\boldsymbol{x})$
- Une fonction $c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ est une *fonction de perte* si $c(y, y) = 0$ et $c(y, y') > 0$ pour $y \neq y'$

$$\widehat{R_n}(\widehat{f}(\boldsymbol{D}_n), \boldsymbol{D}_n) = \frac{1}{n} \sum_{i=1}^{n} c(y_i, \widehat{f}(\boldsymbol{D}_n)(\boldsymbol{x}_i))$$

- Estimation biaisée, par optimisme



Régression polynomiale : Polynômes de degré 1 ($R^2 = 0.03$), *2* ($R^2 = 0.73$), *5* ($R^2 = 0.874$) *et 10* ($R^2 = 1$)

Estimation sans biais sur un échantillon indépendant

- Partition : $\boldsymbol{D}_n = \boldsymbol{d}_{n_1}^{\text{Appr}} \cup \boldsymbol{D}_{n_2}^{\text{Valid}} \cup \boldsymbol{D}_{n_3}^{\text{Test}}$
- $\widehat{R_n}(\widehat{f}(\boldsymbol{D}_{n_1}^{\text{Appr}}), \boldsymbol{D}_{n_1}^{\text{Appr}})$ pour estimer un modèle choisi $\widehat{f}(\boldsymbol{D}_{n_1}^{\text{Appr}})$
- $\widehat{R_n}(\widehat{f}(\boldsymbol{D}_{n_1}^{\text{Appr}}), \boldsymbol{D}_{n_2}^{\text{Valid}})$ pour optimiser un modèle
- $\widehat{R_n}(\widehat{f}, \boldsymbol{D}_{n_3}^{\text{Test}})$ pour comparer les meilleurs modèles

C_p de Mallows (1973)

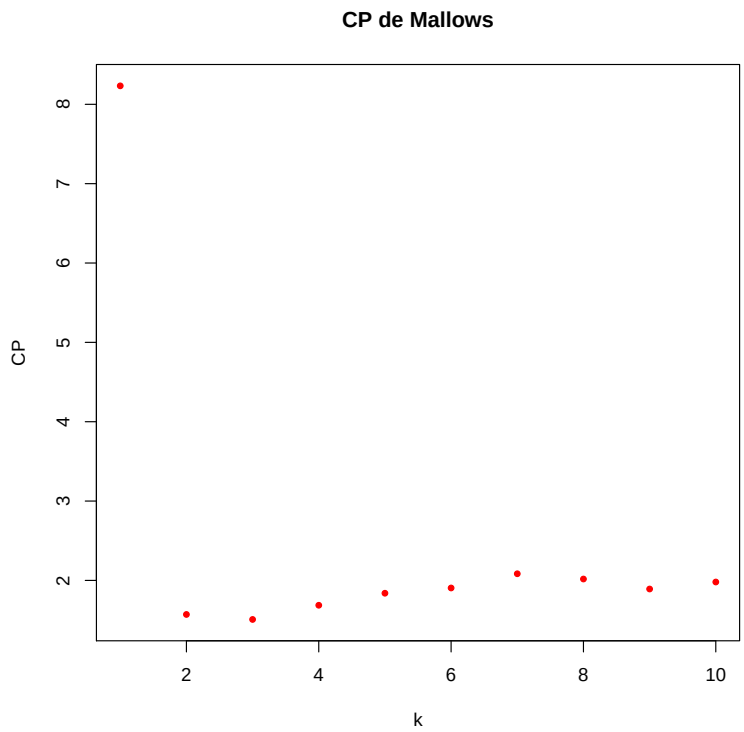
- Décomposition de l'erreur de prévision ou risque quadratique :

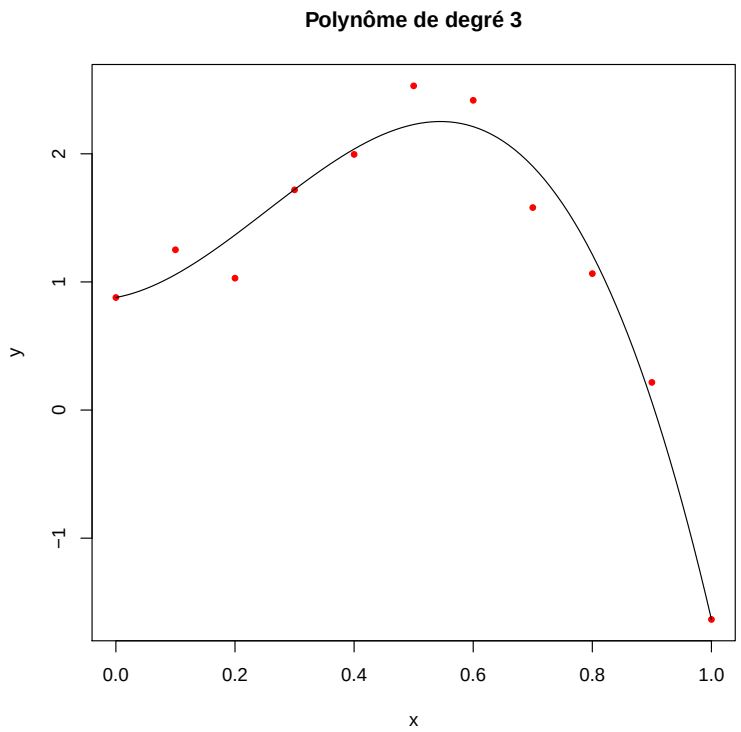
$$\widehat{R_P}(\widehat{f}(\boldsymbol{D}_n)) = \widehat{R_n}(\widehat{f}(\boldsymbol{D}_n), \boldsymbol{D}_n) + \text{Optim}$$

- Estimation normalisée :

$$C_p = \widehat{R_n}(\widehat{f}(\boldsymbol{D}_n), \boldsymbol{D}_n) + 2\frac{d}{n}\widehat{\sigma}^2$$

  - d : nombre de paramètres du modèle
  - n : nombre d'observations
  - s^2 : estimation de la variance de l'erreur par modèle de faible biais





Régression polynomial : C_p de Mallows en fonction du degré du polynôme et modèle de degré 3 sélectionné.

Critère d'Akaïke

- Basé sur la dissemblance de Kullback
- compare la loi de Y et celle de $\widehat{Y}$
  - Suppose que la famille de lois du modèle contient la "vraie" loi de Y
  - Pour tout modèle estimé par minimisation d'une log-vraisemblance $\mathcal{L}$

$$\text{AIC} = -2\mathcal{L} + 2\frac{d}{n}$$

  - Cas gaussien et variance connue : AIC et C_p équivalents
- AIC_c adapté aux petits échantillons gaussiens

$$\text{AIC}_c = -2\mathcal{L} + \frac{n+d}{n-d-2}$$

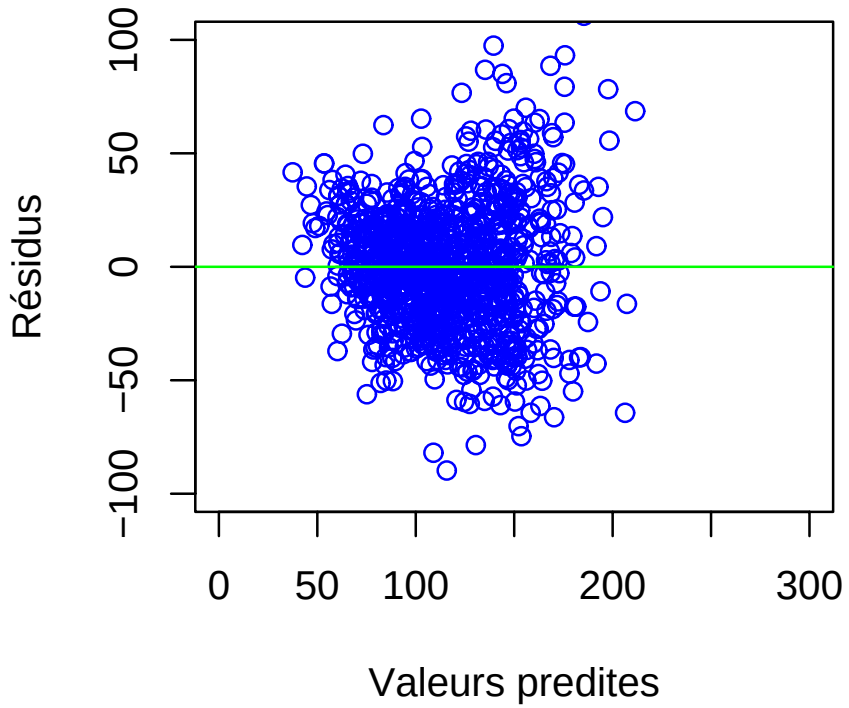
Critère BIC de Schwarz

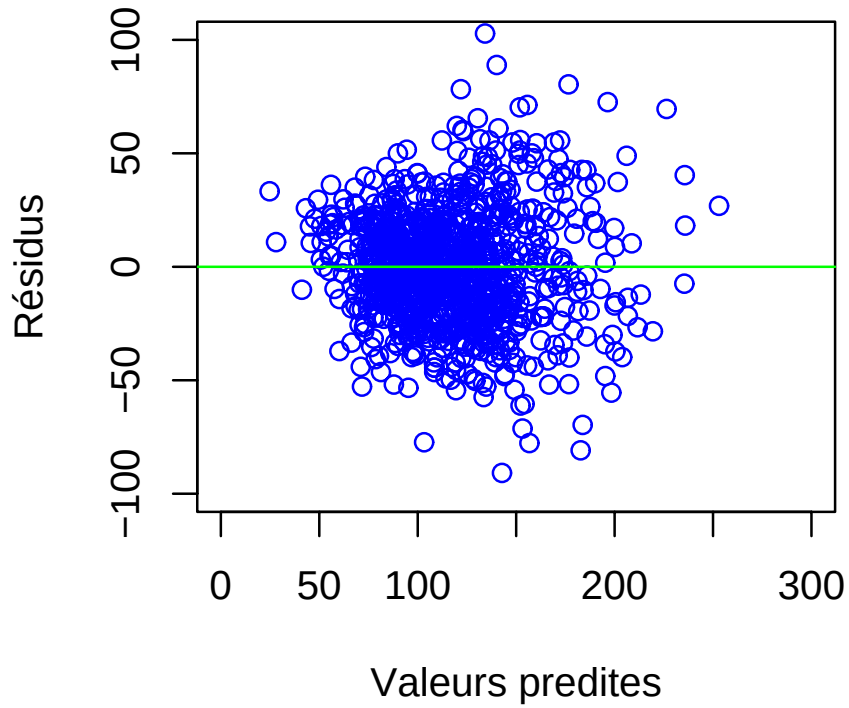
- BIC (Bayesian information criterion)
  - modèle de plus grande probabilité *a posteriori*

$$\text{BIC} = -2\mathcal{L} + \log(n)\frac{d}{n}$$

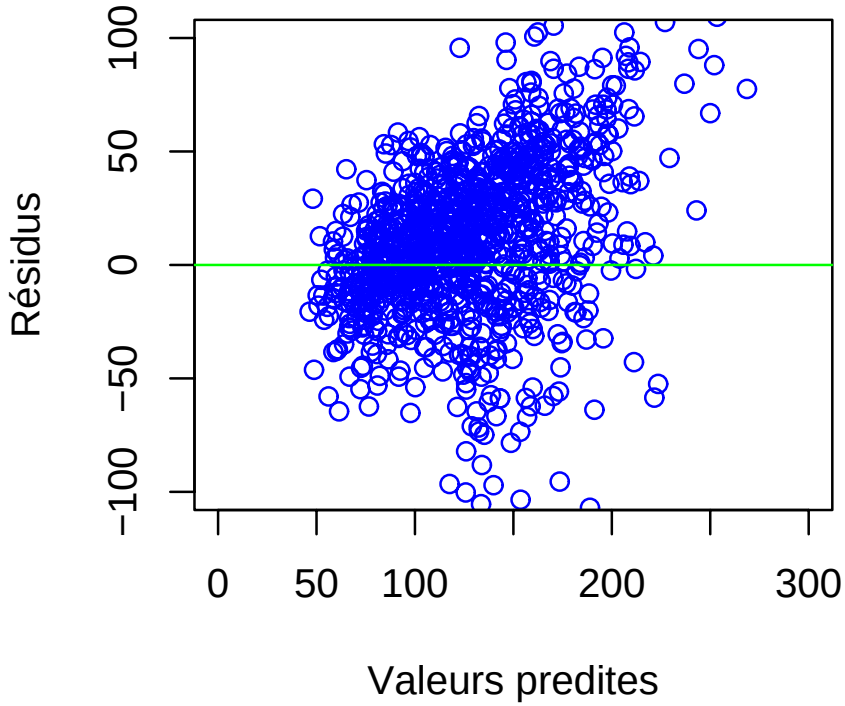
  - Cas gaussien et variance connue : BIC proportionnel à AIC
  - $n > e^2 \approx 7,4$, BIC pénalise plus les modèles complexes
  - Asymptotiquement, la probabilité pour BIC de choisir le bon modèle tend vers 1
  - différent d'AIC qui tend à choisir des modèles trop complexes
  - À Taille fini, BIC risque de se limiter à des modèles trop simples

Cf. Notebook





Ozone : résidus des modèles linéaire et quadratique.



Ozone : résidus du modèle déterministe MOCAGE.

Matrice de confusion pour deux classes

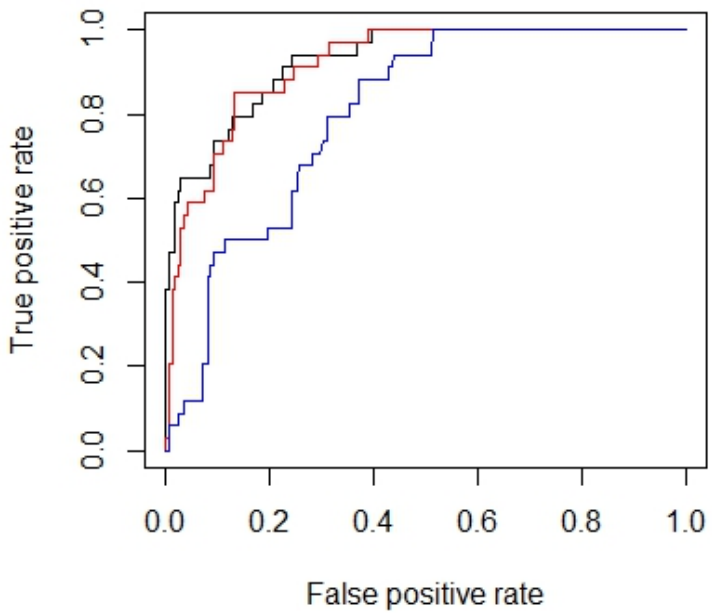
Prévision : Si $\widehat{\pi}_i > s$, $\widehat{y}_i = 1$ sinon $\widehat{y}_i = 0$

Prévision	Observation		Total
	$Y = 1$	$Y = 0$	
$\widehat{y}_i = 1$	$n_{11}(s)$	$n_{10}(s)$	$n_{1+}(s)$
$\widehat{y}_i = 0$	$n_{01}(s)$	$n_{00}(s)$	$n_{0+}(s)$
Total	n_+1	n_+0	n

- Vrais positifs les $n_{11}(s)$ bien classées ($\widehat{y}_i = 1$ et $Y = 1$)
- Vrais négatifs les $n_{00}(s)$ bien classées ($\widehat{y}_i = 0$ et $Y = 0$)
- Faux négatifs les $n_{01}(s)$ mal classées ($\widehat{y}_i = 0$ et $Y = 1$)
- Faux positifs les $n_{10}(s)$ mal classées ($\widehat{y}_i = 1$ et $Y = 0$)
- Le taux d'erreur : $t(s) = \frac{n_{01}(s)+n_{10}(s)}{n}$

Ozone : Courbes ROC.

- Taux de vrais positifs ou *sensibilité* $= \frac{n_{11}(s)}{n_{+1}}$
- Taux de vrais négatifs ou *spécificité* $= \frac{n_{00}(s)}{n_{+0}}$
- Taux de faux positifs $= 1 -$ Spécificité $= \frac{n_{10}(s)}{n_{+0}}$
- AUC : aire sous la courbe



- Sur-échantillonnage
- Score de Pierce : PSS$= H - F = \frac{n_{11}(s)}{n_{+1}(s)} - \frac{n_{10}(s)}{n_{+0}}$
- *Log loss* : LI$= -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m} y_{ik}\log(\widehat{\pi_{ij}})$

Plusieurs classes

- Logistique polytomique ou pas
- Cas général : Risque bayésien associé à δ :

$$R_\delta = \sum_{k=1}^{m} \pi_k \sum_{\ell=1}^{m} c_{\ell \mid k} \int_{\{\mathbf{x} \mid \delta(\mathbf{x})=\mathcal{T}_\ell\}} f_k(\mathbf{x}) d\mathbf{x}$$

Avec

- $c_{\ell \mid k}$: coût du classement dans $\mathcal{T}_\ell$ d'un individu de $\mathcal{T}_k$.
- $\int_{\{\mathbf{x} \mid \delta(\mathbf{x})=\mathcal{T}_\ell\}} f_k(\mathbf{x}) d\mathbf{x}$:
- Probabilité d'affecter $\mathbf{x}$ à $\mathcal{T}_\ell$ alors qu'il est dans $\mathcal{T}_k$.

Analyses discriminantes

- Estimation des densités conditionnelles de chaque classe
- Estimation paramétrique
  - Linéaire : Gaussienne multidimensionnelle avec homoscédasticité
  - Quadratique : Gaussienne multidimensionnelle avec hétéroscédasticité
- Estimation non paramétrique
  - par noyau (fléau de la dimension)
  - k plus proches voisins avec optimisation de k

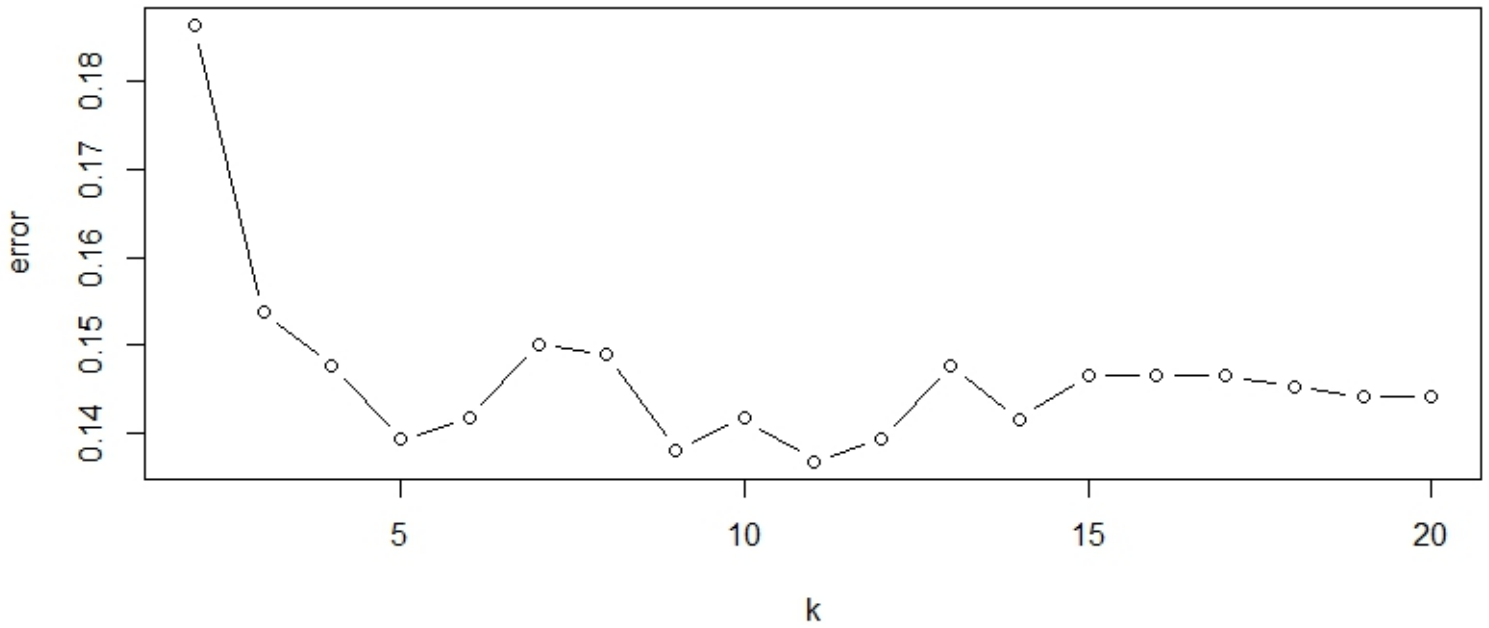
Validation croisée

- Partition aléatoire de l'échantillon en K classes égales
- Soit $\tau : \{1, \ldots, n\} \mapsto \{1, \ldots, K\}$ la fonction d'indexation
- $\widehat{f}^{(-k)}$ estimation de f sans la $k$ième partie de l'échantillon
- Estimation par validation croisée de l'erreur de prévision :

$$\widehat{R_{\text{CV}}} = \frac{1}{n} \sum_{i=1}^{n} l(y_i, \widehat{f}^{(-\tau(i))}(\boldsymbol{x}_i))$$

- Choix de K : n (variance), petit (biais), 10 par défaut
- Utilisation fréquente en choix de modèle :

$$\widehat{\theta} = \arg\min_{\theta} \widehat{R_{\text{CV}}}(\theta)$$



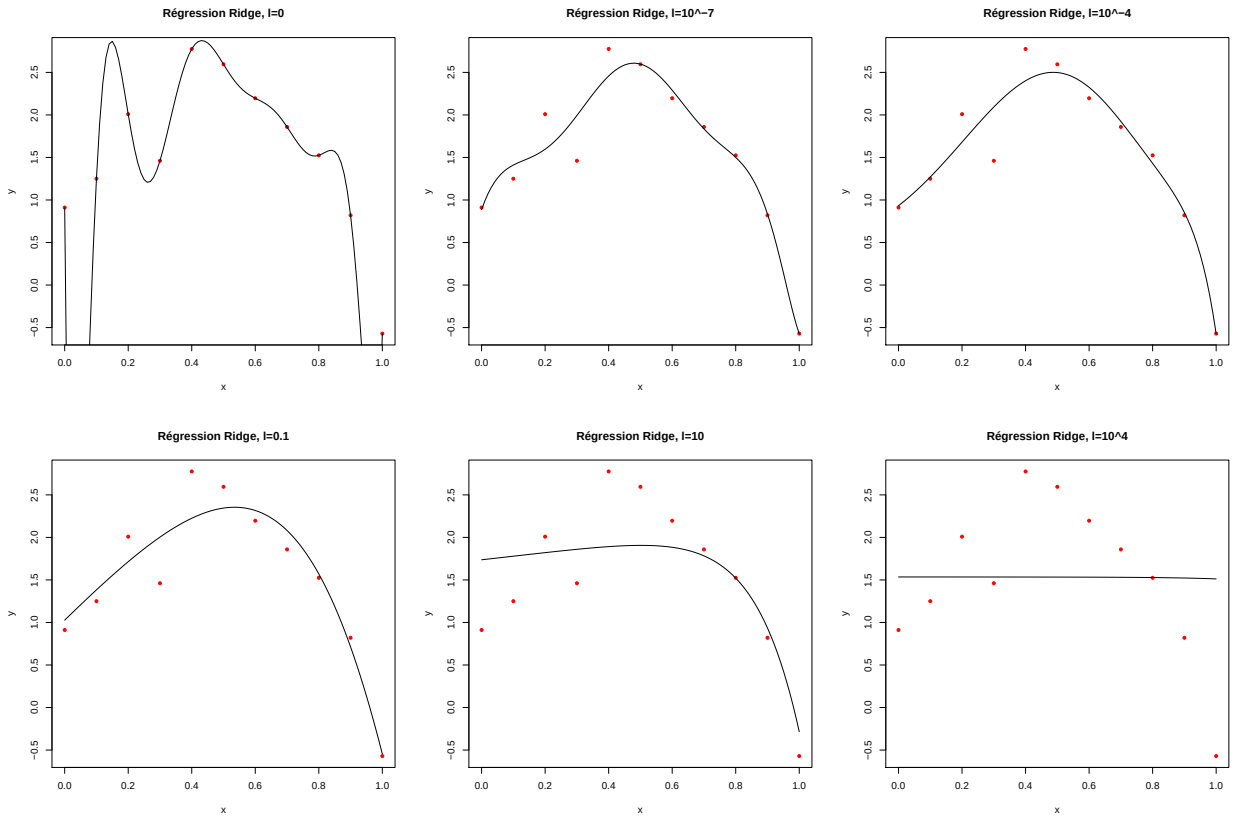
Ozone : Optimisation de k par validation croisée.

Application aux régressions *ridge* et Lasso

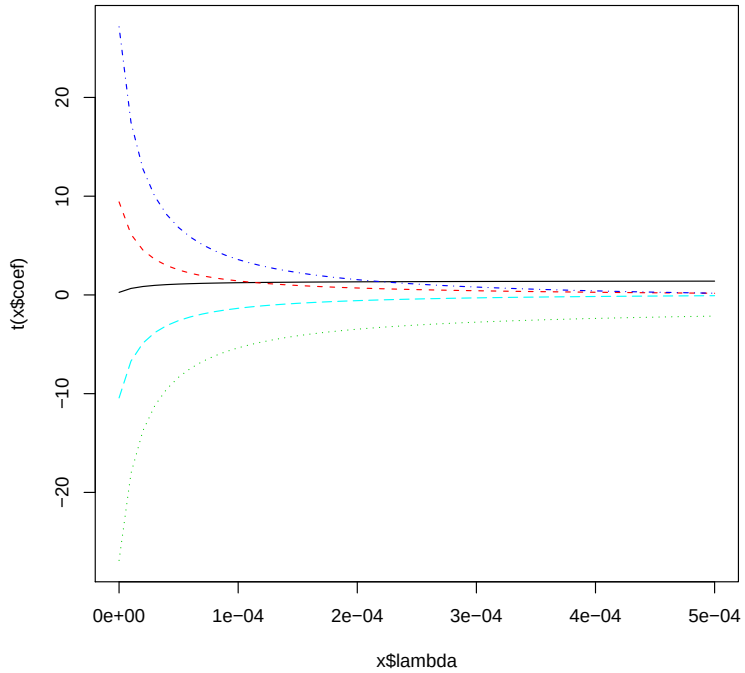
- Régularisation ℓ_2 et ℓ_1

$$\begin{aligned}
\widehat{\beta}_{\text{Ridge}} &= \text{argmin}_{\beta\in\mathbb{R}^{p+1}} \left(\sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p} X_i^{(j)} \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right) \\
&= (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^T\mathbf{Y} \\
\widehat{\beta}_{\text{Lasso}} &= \text{argmin}_{\beta\in\mathbb{R}^{p}} \left(\sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p} X_i^{(j)} \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)
\end{aligned}$$

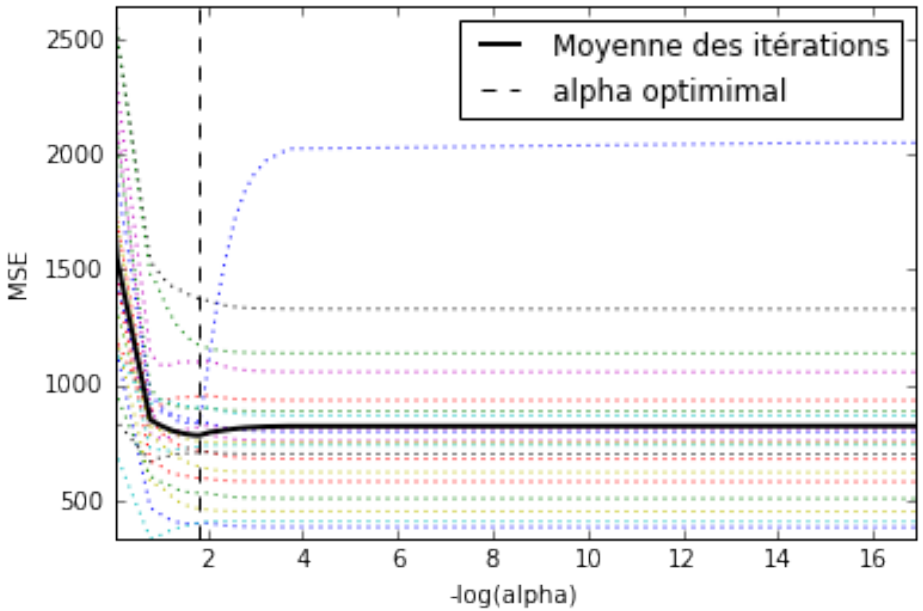
- Optimisation par validation croisée
- Régularisation *Elastic net*



Régression polynomial et régularisation ridge



Régression polynomial : chemin de régularisation des paramètres ridge.



Ozone : optimisation de régularisation lasso par validation croisée.

Estimateur Bootstrap

- Simulation (Monte Carlo) de la distribution d'un estimateur
- Principe : substituer P_n, à la distribution inconnue P
- Tirage avec remise d'un échantillon de même taille

$$\widehat{R_{\text{Boot}}} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{n}\sum_{i=1}^{n} c(y_i, f_{\boldsymbol{z}^{*b}}(\boldsymbol{x}_i))$$

- Biais ?

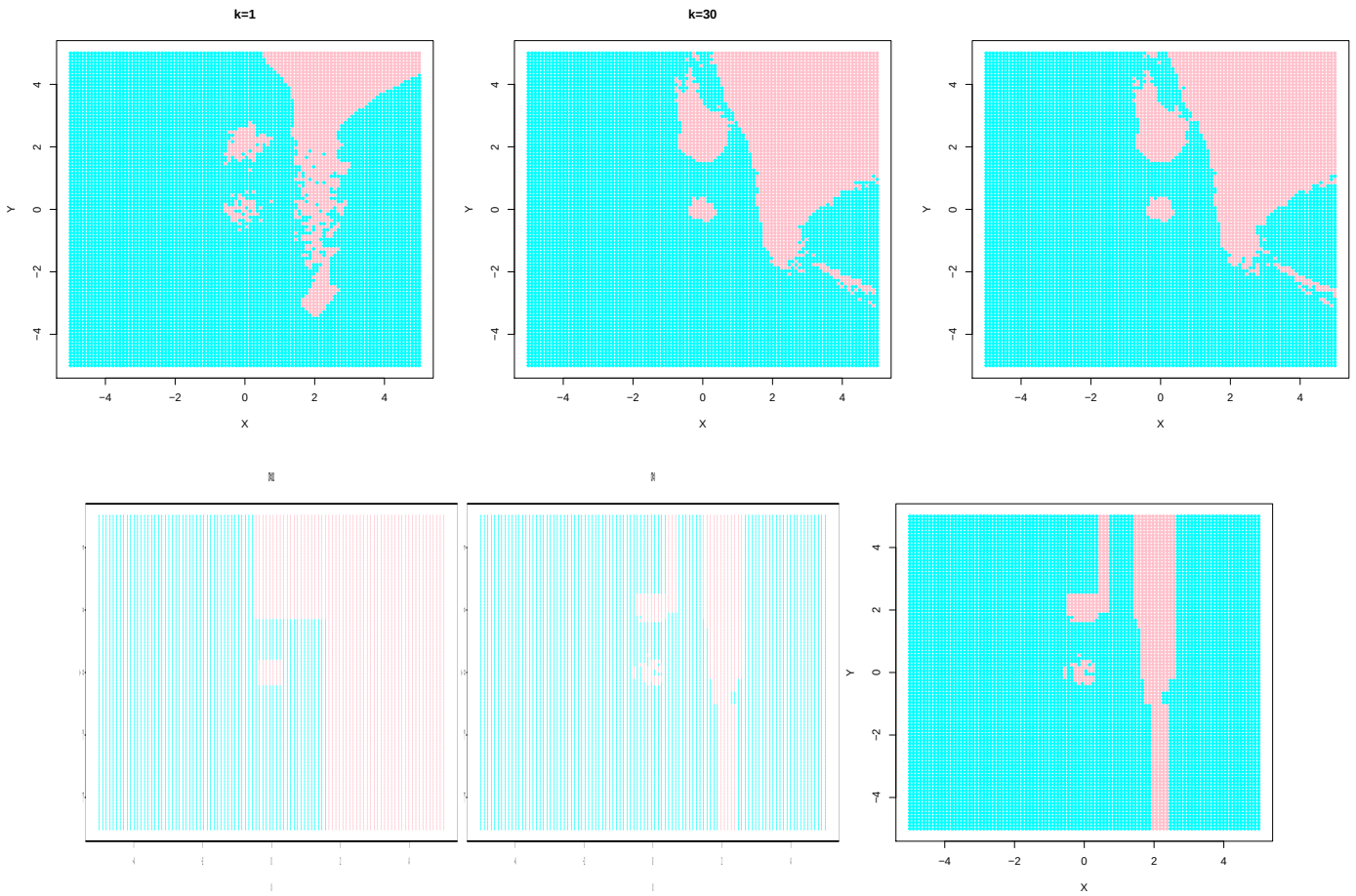
Estimateur bootstrap out-of-bag

Distinguer les observations de l'échantillon bootstrap et les autres

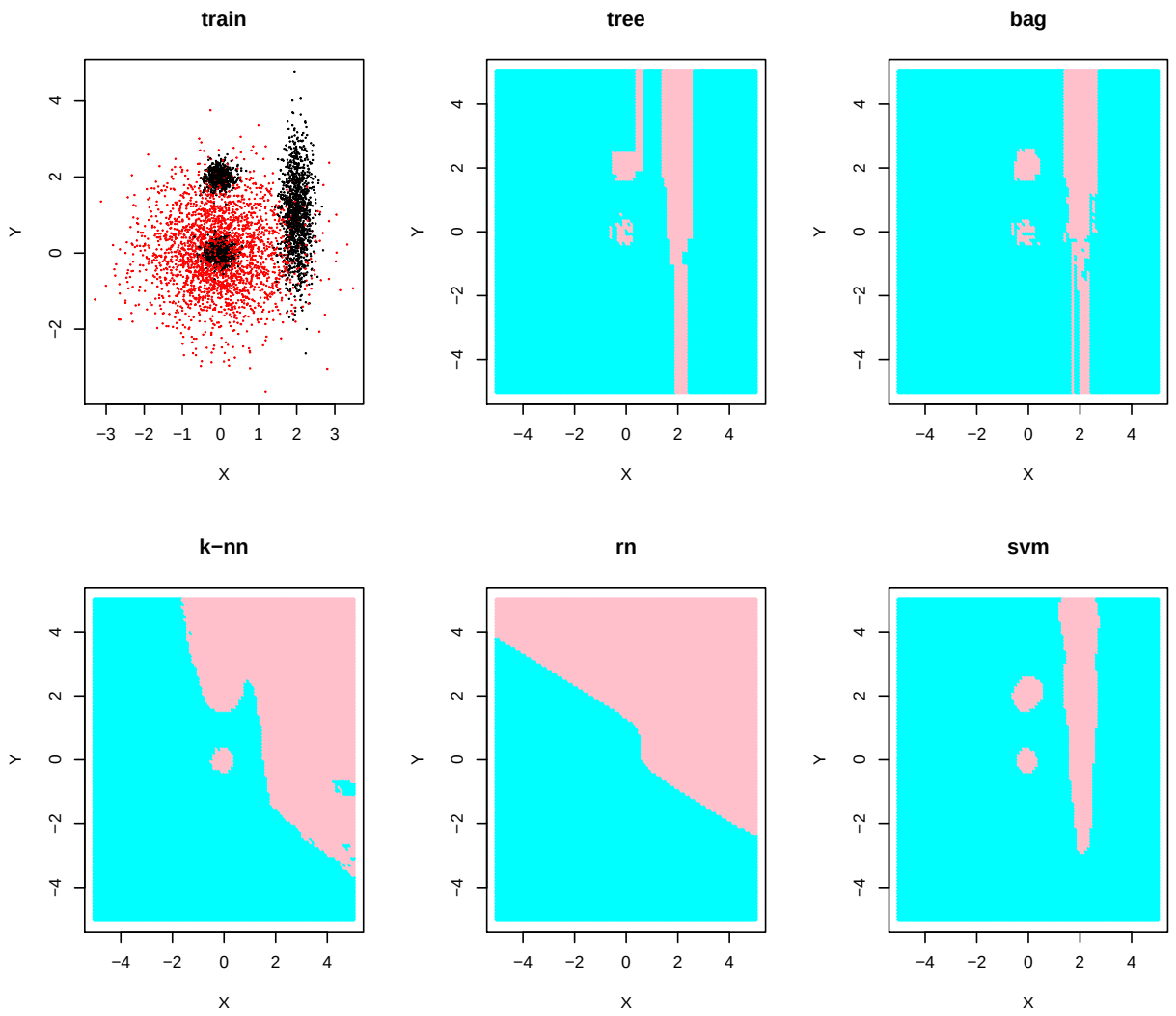
$$\widehat{R_{\text{oob}}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{B_i}\sum_{b\in K_i} c(y_i, f_{\mathcal{Z}^{*b}}(\boldsymbol{x}_i))$$

- K_i est l'ensemble des indices b des échantillons bootstrap ne contenant pas la $i$ème observation à l'issue des B simulations
- $B_i = |K_i|$ est le nombre de ces échantillons
- Biais ?
- $\widehat{R}_{.632} = 0,368 \times \widehat{R}_n(\hat{f}(\boldsymbol{d}^n), \boldsymbol{d}^n) + 0,632 \times \widehat{R}_{\text{oob}}$

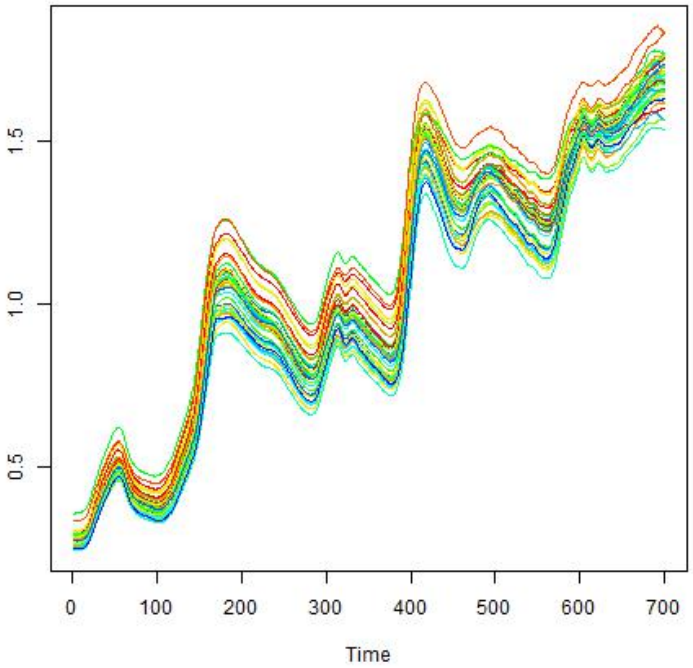
Cf. Notebook

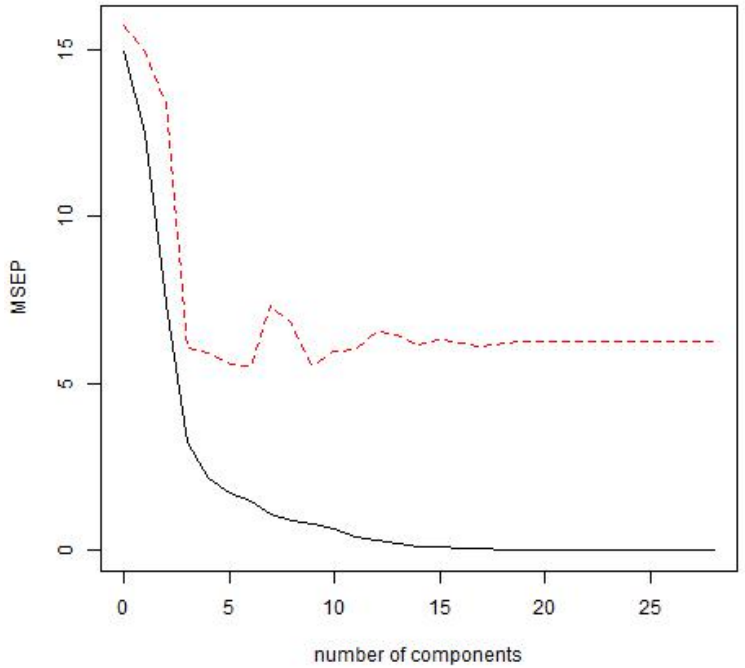


Mélanges gaussiens : k plus proches voisins et arbres de discrimination.



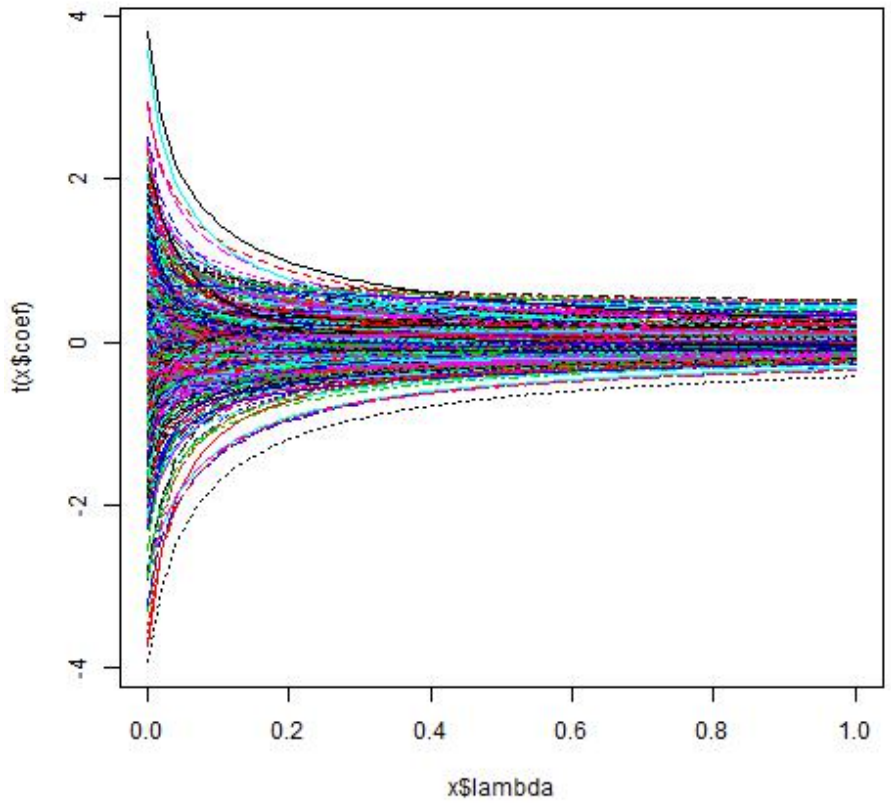
Mélanges gaussiens : Optimisation par validation croisée.

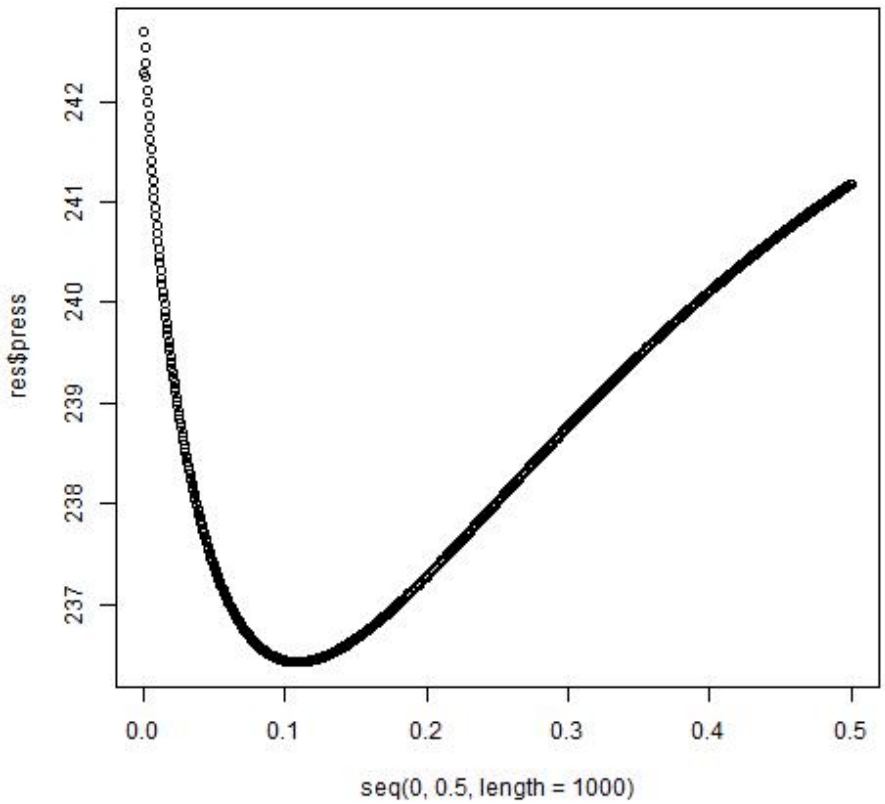




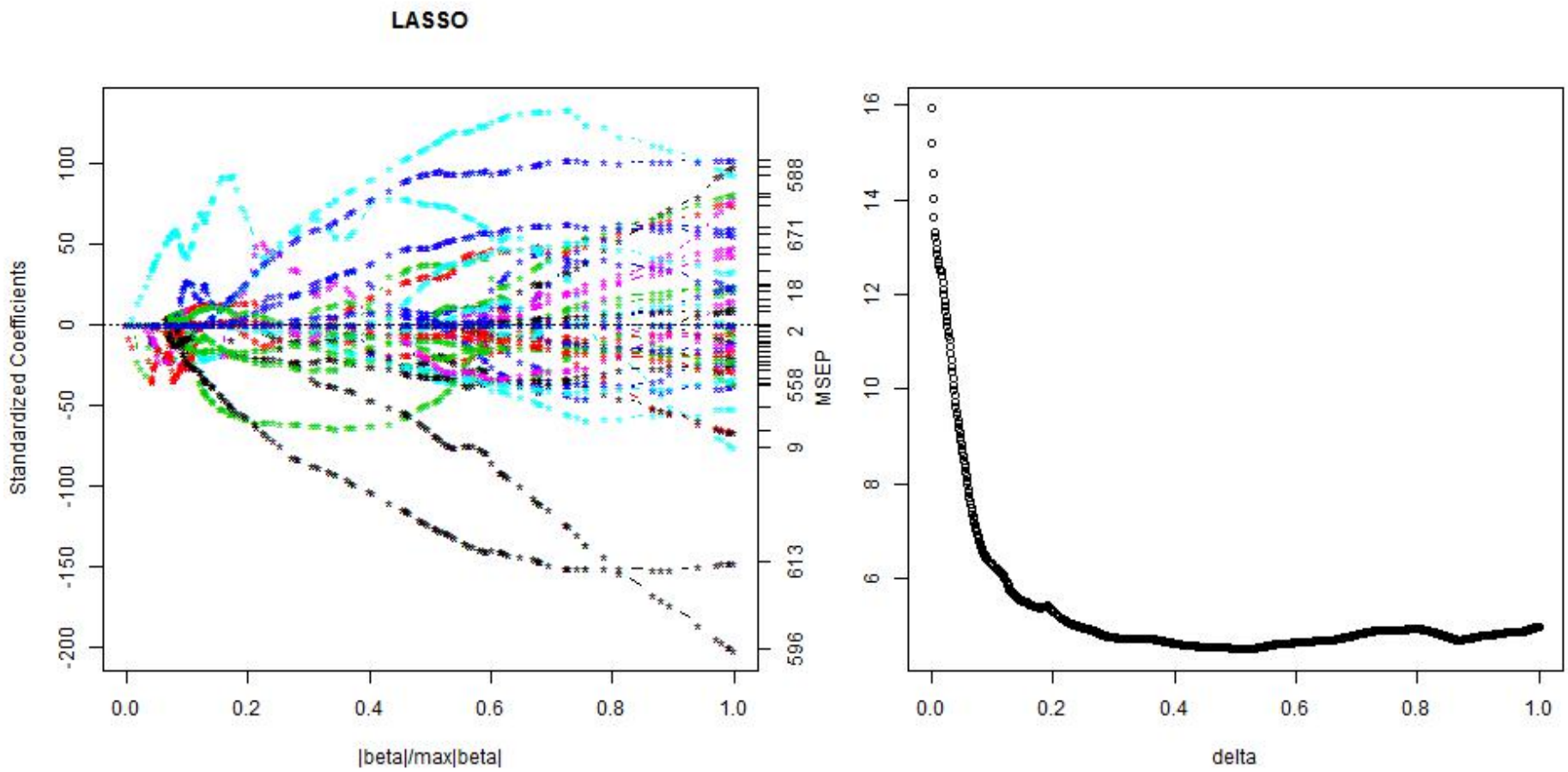
Cookies : Spectres proche infrarouge (NIR) de 72 échantillons de pâtes à gâteaux. $p > n$*. Nb composantes de la PLS*

Régression sur composantes orthogonales, PLS, *sparse* PLS, *sparse* PLS DA (Lê Cao et al. 2009, 2011)

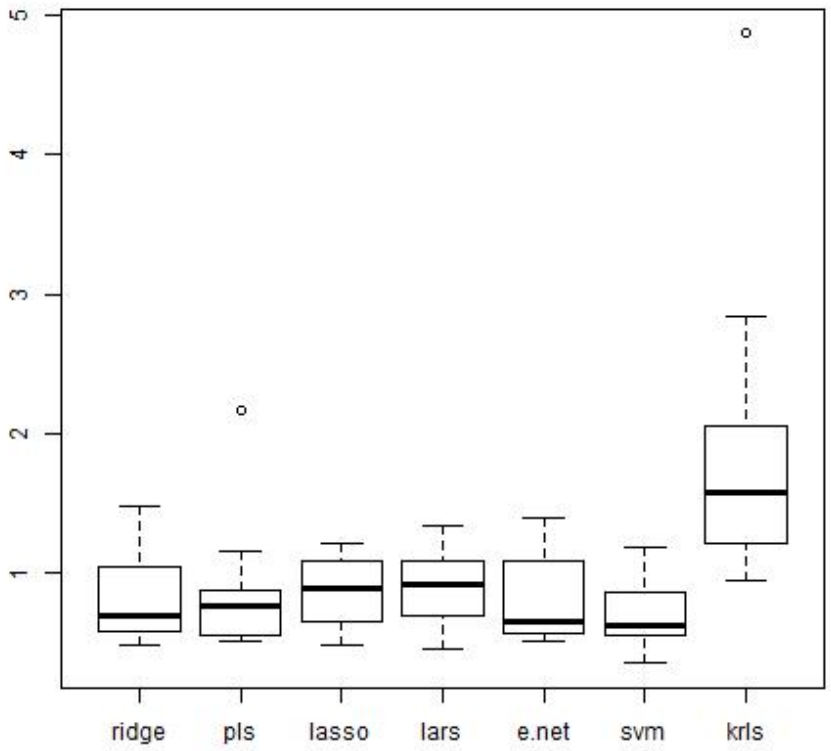




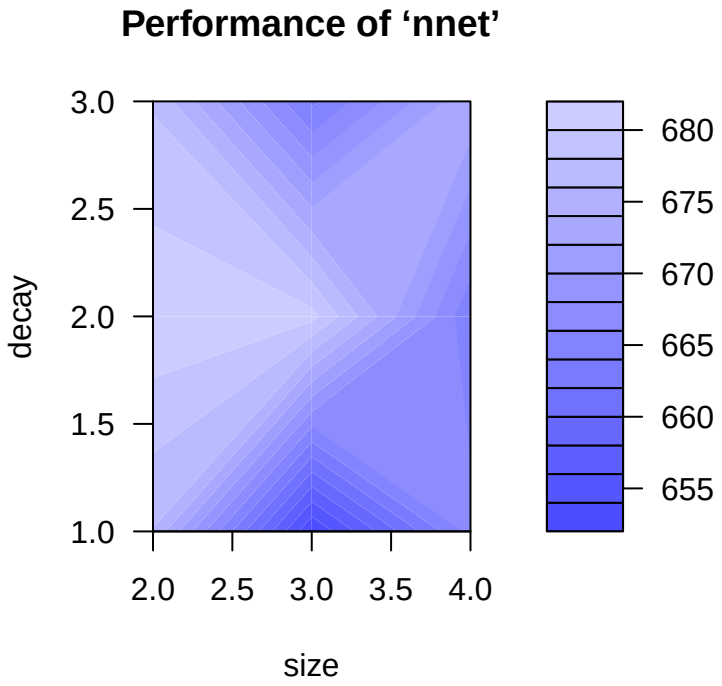
Cookies : chemins de régularisation ridge et optimisation.

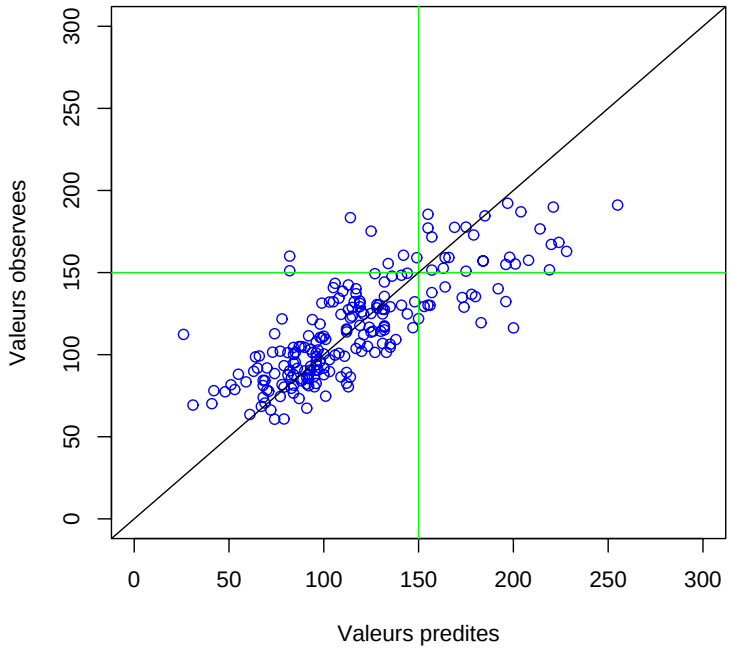


Cookies : chemin de régularisation lasso et optimisation.

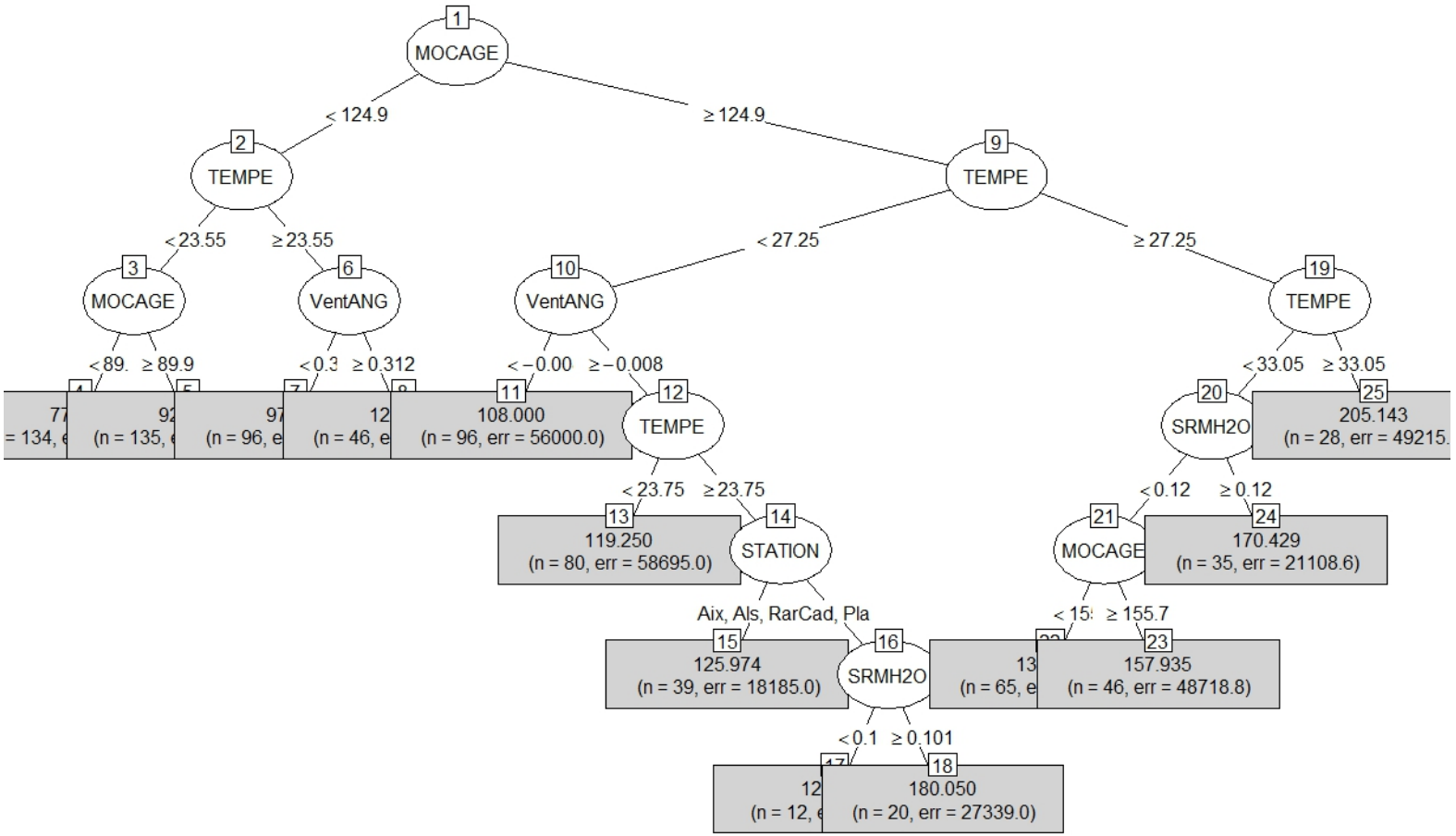


Cookies : Distribution des erreurs de prévision.

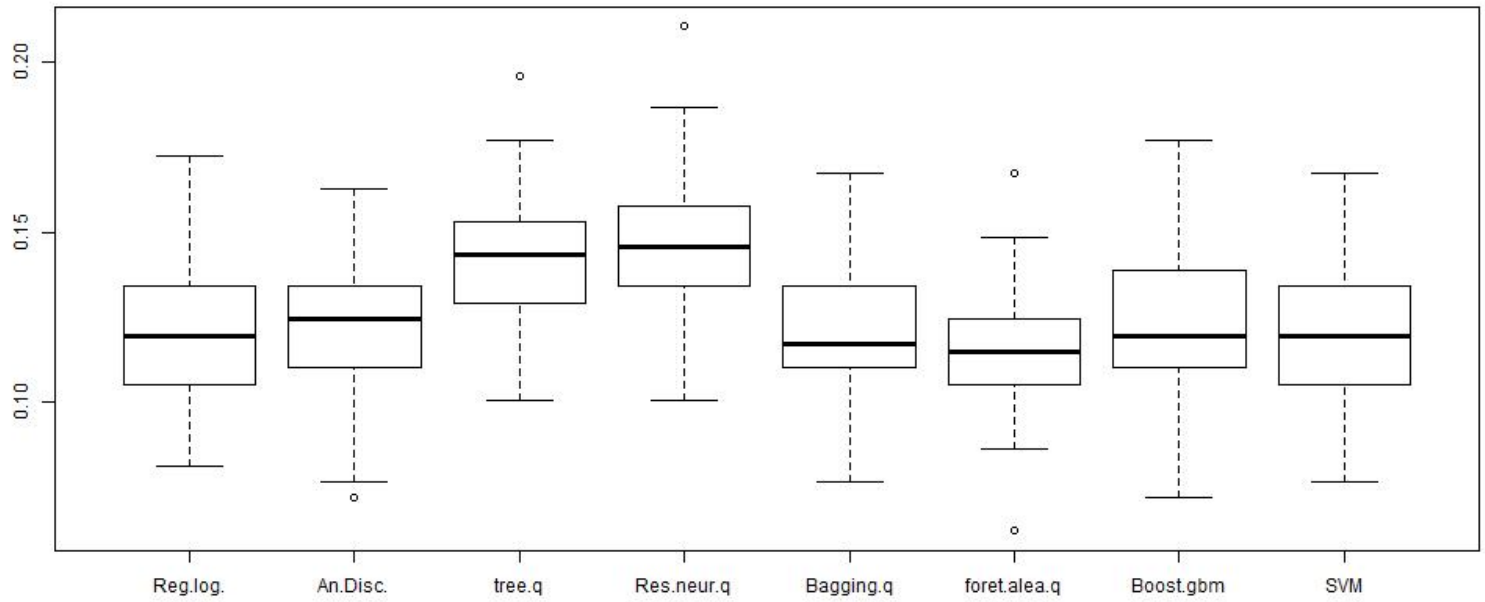




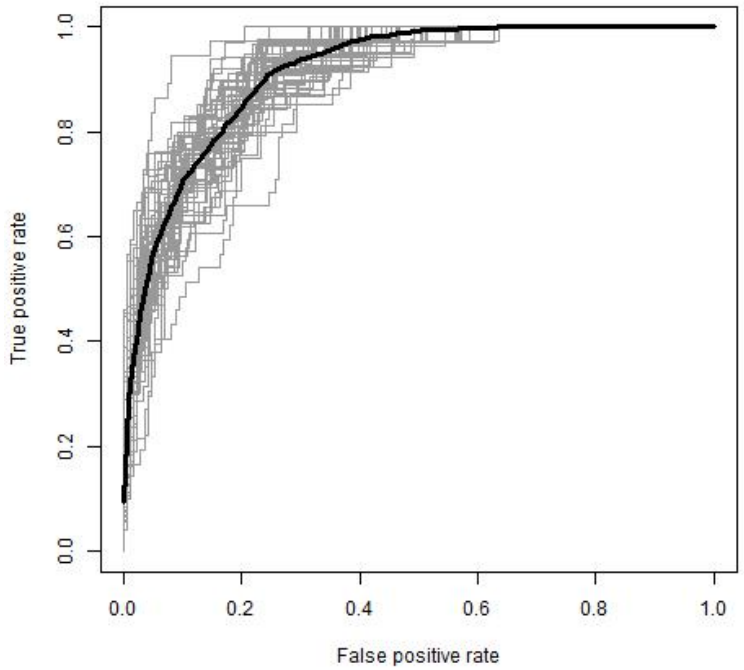
Ozone : optimisation des paramètres d'un réseau neurones et SVM : Valeurs observées fonction des valeurs prédites.

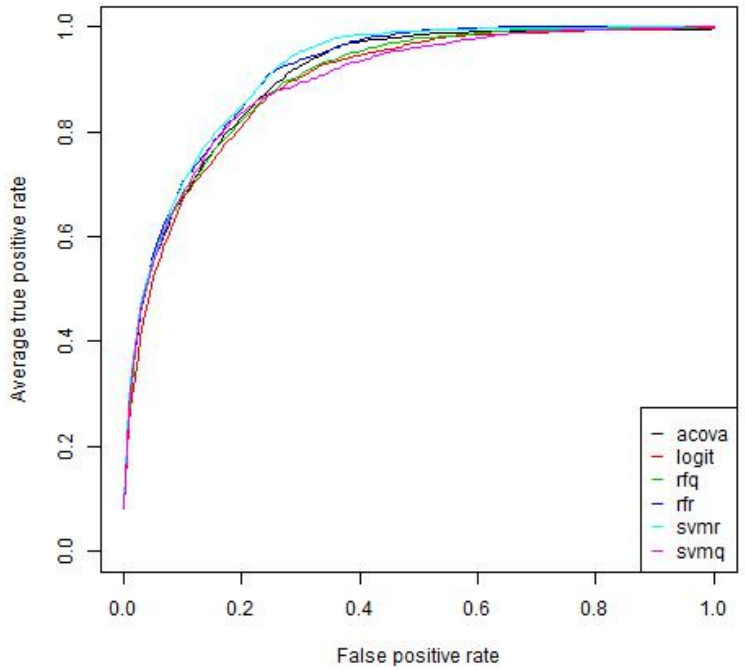


Ozone : Arbre de régression élagué.



Ozone : Diagrammes boîtes des taux d'erreurs pour la prévision des dépassements de seuil.

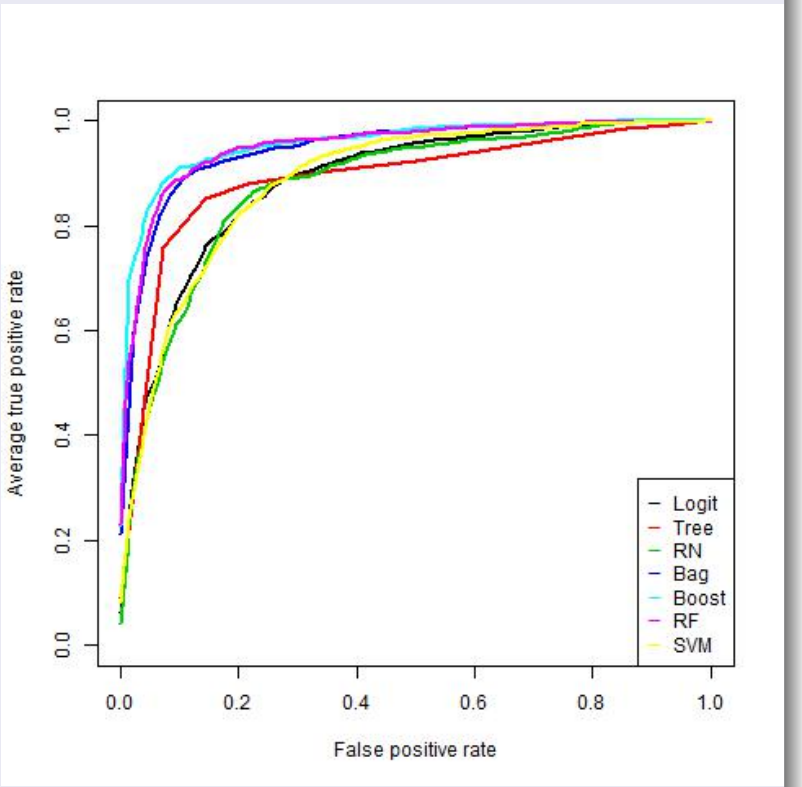




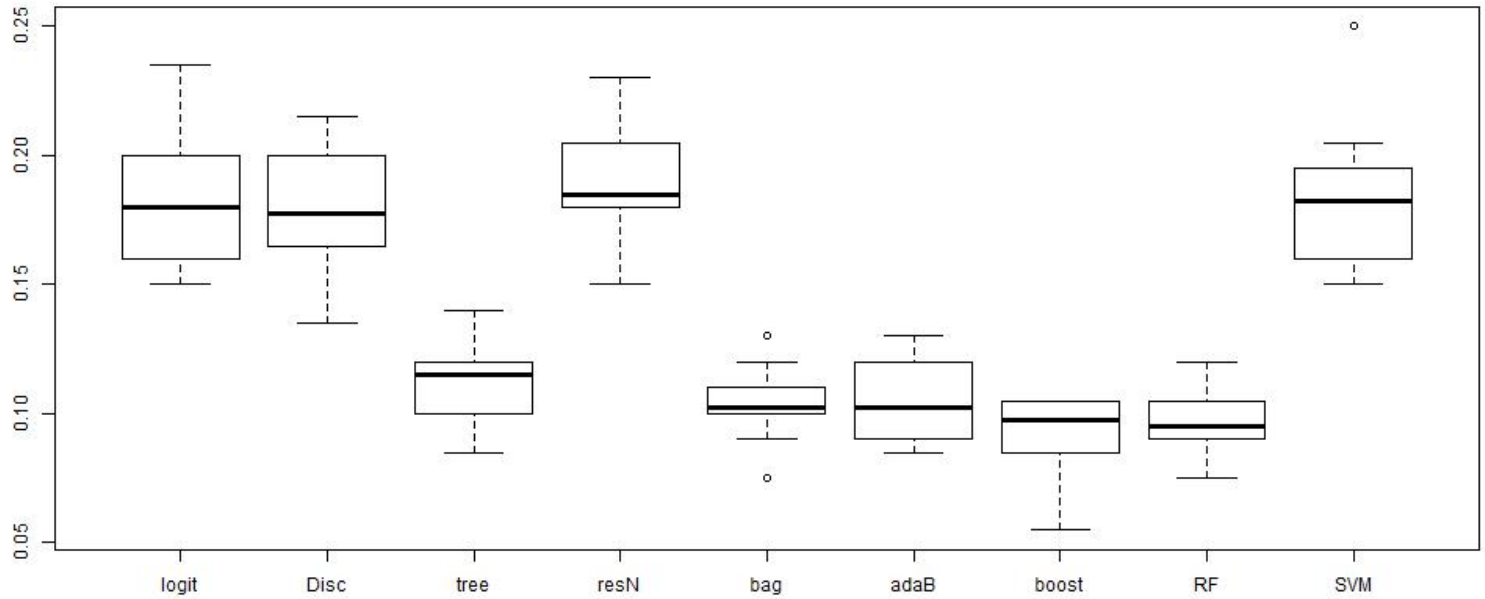
Ozone : courbes ROC (random forest) pour 50 échantillons test et moyennes.

GRC : score d'appétence

- Données bancaires de possession ou on de la Carte VP
- Échantillon "équilibré" de 825 clients
- 32 variables comportementales
- Travail important de préparation



GRC : Courbes ROC



GRC : Diagrammes boîtes des taux d'erreurs.

Discussion

- Pas de méthode universellement meilleure
- Éviter l"acharnement thérapeutique" façon "kaggle"
- Optimisation des paramètres plus ou moins complexe et ou coûteuse
- Ne pas négliger les méthodes simples, interprétables (Hand, 2006 et Donoho, 2015)
- R pour l'interprétation *vs.* Python pour l'efficacité
- Points "oubliés" :
  - Préparation des données cf. chapitre 12
  - Données manquantes (imputation)
  - Classes déséquilibrées
  - Détection d'observation atypiques (OCC).