Détection d'individus atypiques en régression SIR

Hadrien Lorenzo & Jérôme Saracco

hadrien.lorenzo@u-bordeaux.fr
jerome.saracco@math.u-bordeaux.fr

Equipe ASTRAL, INRIA BSO

Mercredi 9 juin 2021

Le modèle semi-paramétrique SIR

Contexte théorique : le « single index model » semi-paramétrique de Duan et Li (1991) tel que

$$y = f(\beta' x) + \epsilon \tag{1}$$

où

y est une variable réponse univariée,
$x \in \mathbb{R}^p$, les covariables, telles que $\mathbb{E}(x) = \mu$ et $\mathbb{V}(x) = \Sigma$,
ϵ est indépendant de x,
f la « fonction de lien » et le paramètre euclidien $\beta \in \mathbb{R}^p$ sont inconnus.

f étant inconnue, β n'est pas totalement identifiable.
Il est par contre possible d'estimer l'espace engendré par β, appelé Espace EDR (« Effective Dimension Reduction »).

Remarque : Le modèle (1) peut être généralisé à un bruit non additif et hétéroscédastique.

Estimation de l'espace EDR et de f

L'estimation du modèle SIR passe par deux étapes

Estimation de l'espace EDR

$$M = \mathbb{V}\left[\mathbb{E}\{x|T(y)\}\right] \quad \Longrightarrow \quad \text{Nom de la méthode « régression inverse par tranches »}$$

Le vecteur propre principal, noté $b \in \mathbb{R}^p$, de $\Sigma^{-1}M$ est une direction EDR.

Estimation de f

Utilisation d'un estimateur non paramétrique de type « kernel-smoothing » sur $(y, b'x)$.

Problème : Cette procédure d'estimation est sensible aux individus atypiques.

Peu de recherche sur SIR « robuste », voir Gather, Hilker et Becker (2002) ou Cook et Critchley (2000) par exemple.

Comment définir un individu atypique ?

Un **individu atypique** (« **outlier** ») est une observation ne suivant pas le modèle statistique proposé.

Attention : Un individu « **borderline** » suit le modèle statistique mais se réalise avec une probabilité « *faible* », en « queue » de distribution.
Donc « **outlier** »≠ « **borderline** ».

Objectif de la présentation : Proposer trois méthodes computationnelles de détection d'individus atypiques dans ce cadre.

Une structure d'estimation commune

Notations :

Un jeu de données $S = \{(x_i, y_i),\ i = 1, \ldots, n\}$.

Trois méthodes (**MONO**, **TTR** et **BOOT**) suivent trois étapes :

Etape 1 : Estimation(s) de (b, f).

Etape 2 : Estimation d'erreurs de prédiction.

Etape 3 : Classification **normal**/**outlier**(/**borderline**).

$H = 10$ tranches sont utilisées pour SIR,

Noyau Gaussien et fenêtre obtenue par CV pour le « kernel smoothing ».

Etape 1 : Estimation(s) de *(b,f)*

MONO Un seul couple (b, f) estimé sur S entier.

TTR R couples (b, f), chacun construit sur un sous-échantillon d'apprentissage de S de taille « $0.9n$ ».

BOOT B couples (b, f), chacun construit sur un sous-échantillon bootstrap de S.

Etape 2 : Estimation d'erreurs de prédiction.

Estimer des y_i, en déduire une erreur e_i d'estimation des y_i (qui sera utilisée pour classer les observations en **normal**/**outlier**(/**borderline**))

MONO Estimation sur l'échantillon complet S (IB pour « In Bag »).

TTR Estimation sur les R échantillons de test (OOB pour « Out Of Bag »).

BOOT Estimation sur les B échantillons d'apprentissage (IB).

Etape 2 : Estimation d'erreurs de prédiction.

Estimer des y_i, en déduire une erreur e_i d'estimation des y_i (qui sera utilisée pour classer les observations en **normal**/**outlier**(/**borderline**))

MONO Estimation sur l'échantillon complet S (IB pour « In Bag »).

TTR Estimation sur les R échantillons de test (OOB pour « Out Of Bag »).

BOOT Estimation sur les B échantillons d'apprentissage (IB).

Idée : Convenir qu'une observation **normale** sera toujours bien prédite, qu'une observation **borderline** sera uniquement bien prédite si elle est présente (au moins une fois) dans l'échantillon d'apprentissage, alors qu'un **outlier** sera, quant à lui, toujours mal prédit.

Etape 2 : Estimation d'erreurs de prédiction.

Estimer des y_i, en déduire une erreur e_i d'estimation des y_i (qui sera utilisée pour classer les observations en **normal**/**outlier**(/**borderline**))

MONO Estimation sur l'échantillon complet S (IB pour « In Bag »).

TTR Estimation sur les R échantillons de test (OOB pour « Out Of Bag »).

BOOT Estimation sur les B échantillons d'apprentissage (IB).

Idée : Convenir qu'une observation **normale** sera toujours bien prédite, qu'une observation **borderline** sera uniquement bien prédite si elle est présente (au moins une fois) dans l'échantillon d'apprentissage, alors qu'un **outlier** sera, quant à lui, toujours mal prédit.

$\Longrightarrow$ Nécessité d'introduire une dynamique d'apprentissage pour différencier **outlier**/**borderline**.

Etape 3 : Classification en **normal**/**outlier**(/**borderline**).

Notation : Règle sur boxplot :
$R_0(\bullet) = \{\bullet > Q_3(\bullet) + 1.5(Q_3(\bullet) - Q_1(\bullet))\}$.

MONO $R_0(\{e_i\}_i)$.

TTR Calcul des erreurs moyennes (OOB) sur les R réplications $\bar{e}_i$ et détection de rupture. Utilisation du package **R** `changepoint`, segmentation binaire (un seul point de rupture).

BOOT Calcul des erreurs moyennes (IB) sur les B sous-échantillons bootstrap $\bar{\bar{e}}_i$ et discrimination

→ **Outlier** : $R_0(\{\log(\bar{\bar{e}}_i)\}_i)$

→ **Borderline** : $R_0(\{(\bar{\bar{e}}_i\}_i)$

BOOT seulement différencie **outlier** et **borderline**.

$R = B = 2000$ suffit largement dans les cas rencontrés.

Shéma de simulation

$$y = \frac{(x'\beta)^3}{100} + \epsilon, \tag{2}$$

où

- $\beta = (2, 2, 1, -2, -3, 0, \ldots, 0)' \in \mathbb{R}^p$,
- $x \sim \mathbb{U}_{[-2;2]^p}$,
- $\epsilon \sim \mathcal{N}(0, \sigma^2 = 0.25)$ et $\epsilon \perp\!\!\!\perp x$.

$n = \tilde{n} + \tilde{\tilde{n}}$ avec

- $\tilde{n} = 200$ observations **normales** $\{(x_i, y_i),\ i = 1, \ldots, \tilde{n}\} \sim (2)$ avec $p \in \{5, 20\}$.
- $\tilde{\tilde{n}} = 10$ observations **outlier** : $\forall i = \tilde{n} + 1, \ldots, \tilde{n} + \tilde{\tilde{n}}$,
  - $\rightarrow$ $x_i \sim \mathbb{U}_{[-2;2]^p}$,
  - $\rightarrow$ $y_i \sim \mathbb{U}_{\text{Supp}(y_{1,\ldots,\tilde{n}})}$.

Méthode **MONO**

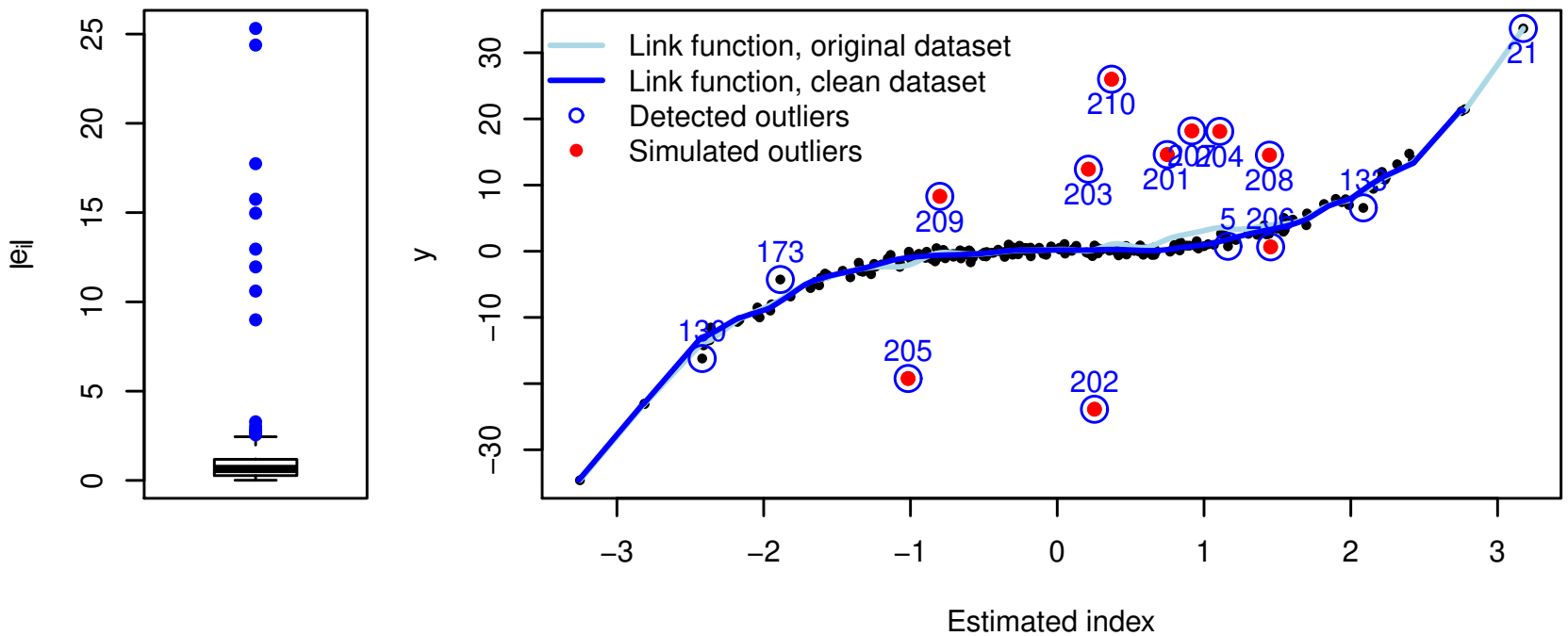


FIGURE – Exemple d'application de la méthode MONO.

Méthode **TTR**

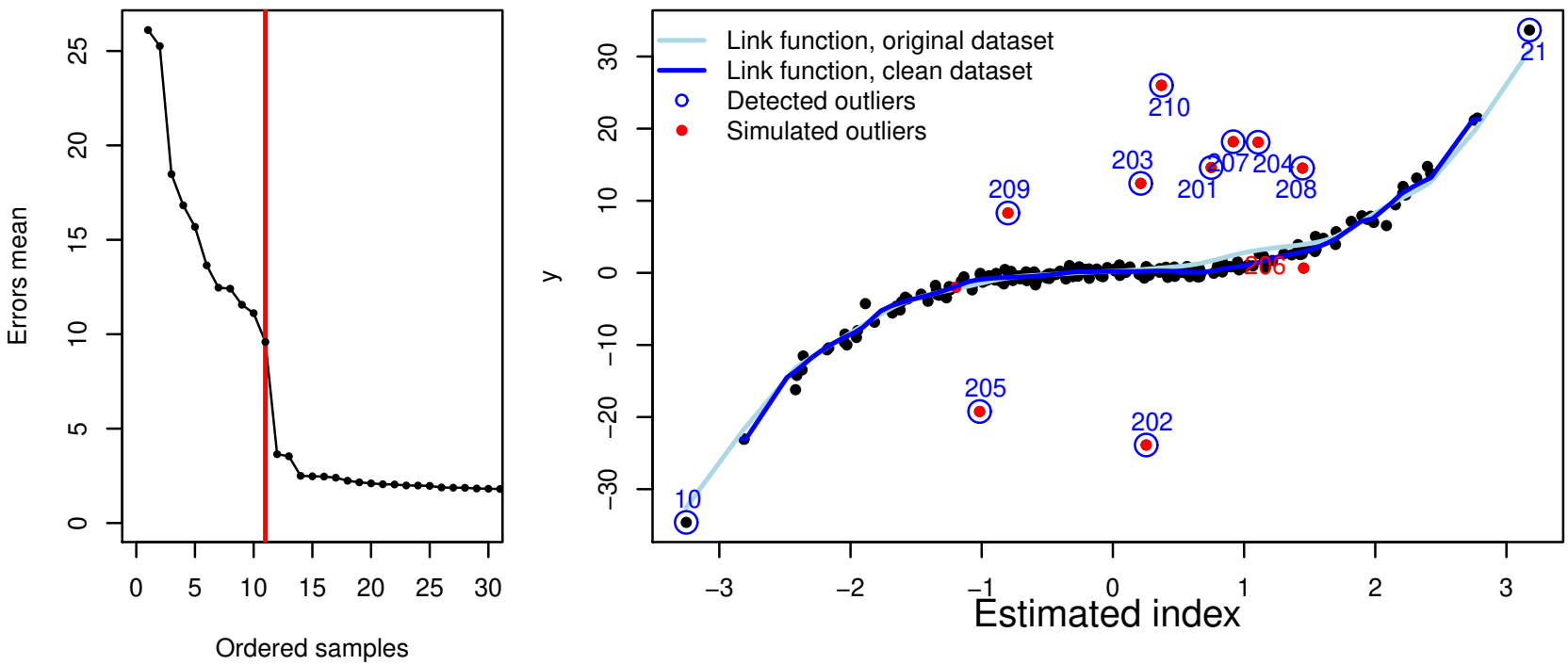


FIGURE – Exemple d'application de la méthode **TTR**.

Observations 10 et 21 plutôt **borderline** mais labélisées **outlier**.

Méthode **BOOT**

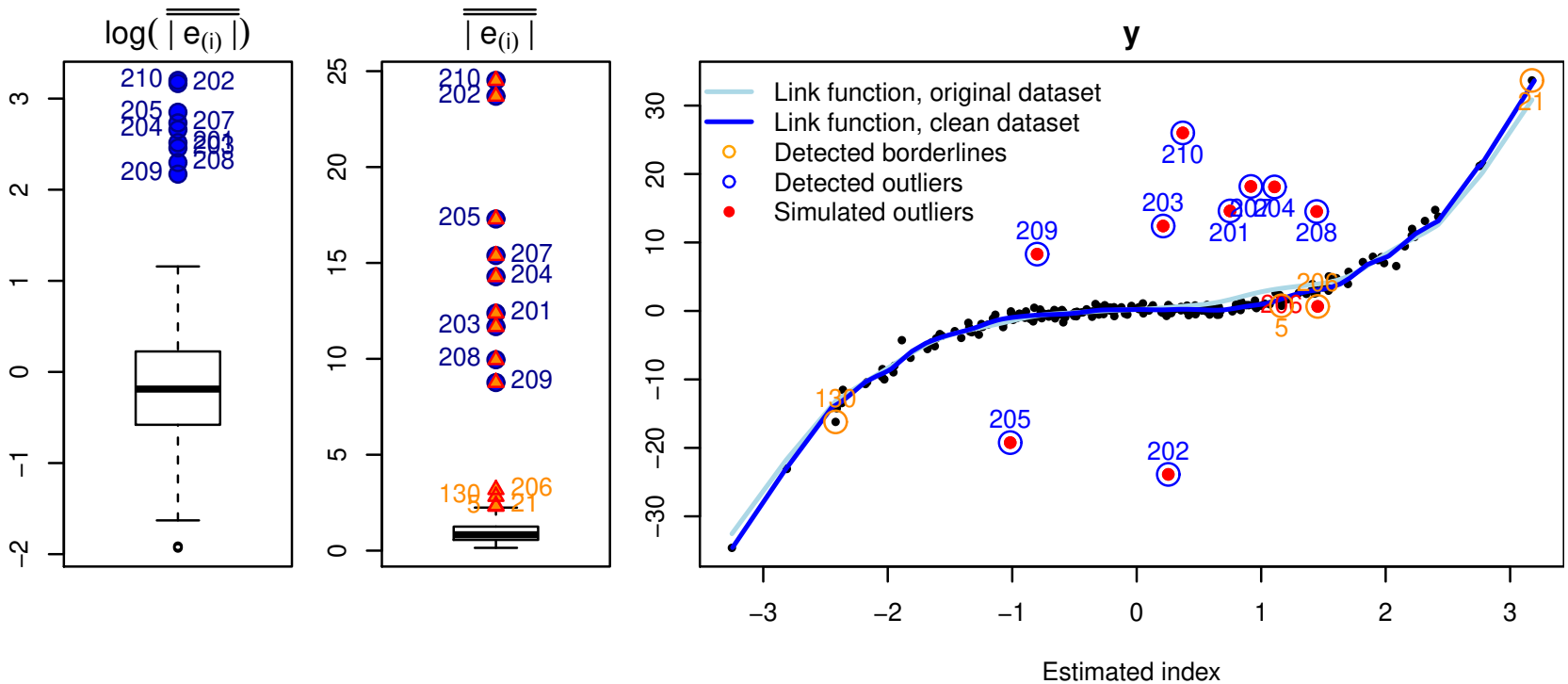
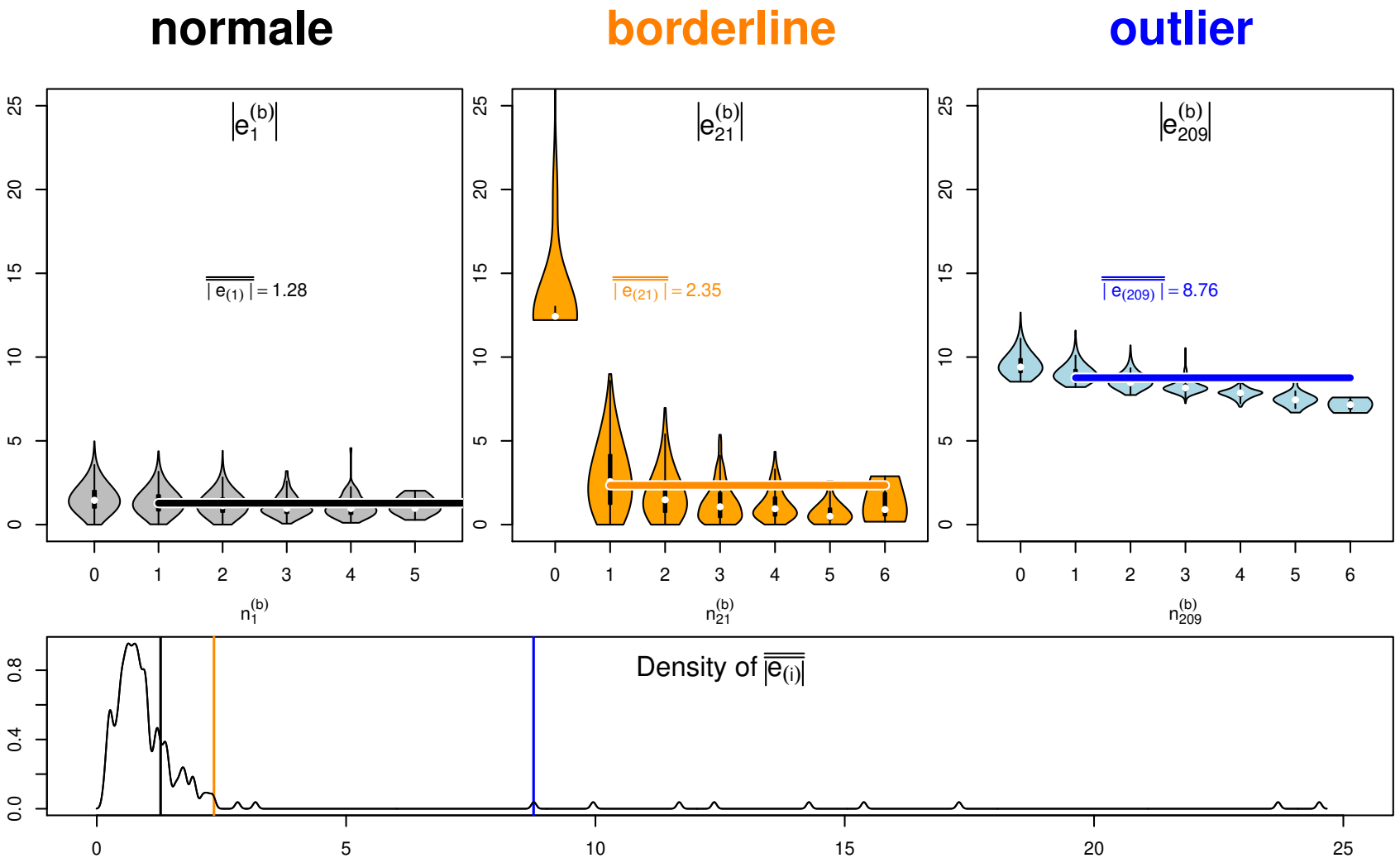


FIGURE – Exemple d'application de la méthode **BOOT**.

Autour de la « dynamique d'apprentissage »



Exemple sur le jeu de données « ozone »

Voir par exemple Cornillon et al. (2012).
$n = 112$, $p = 10$, $R = B = 1000$.
Concentrations en ozone à Rennes (variable à prédire) par des mesures météorologiques (durant l'été).

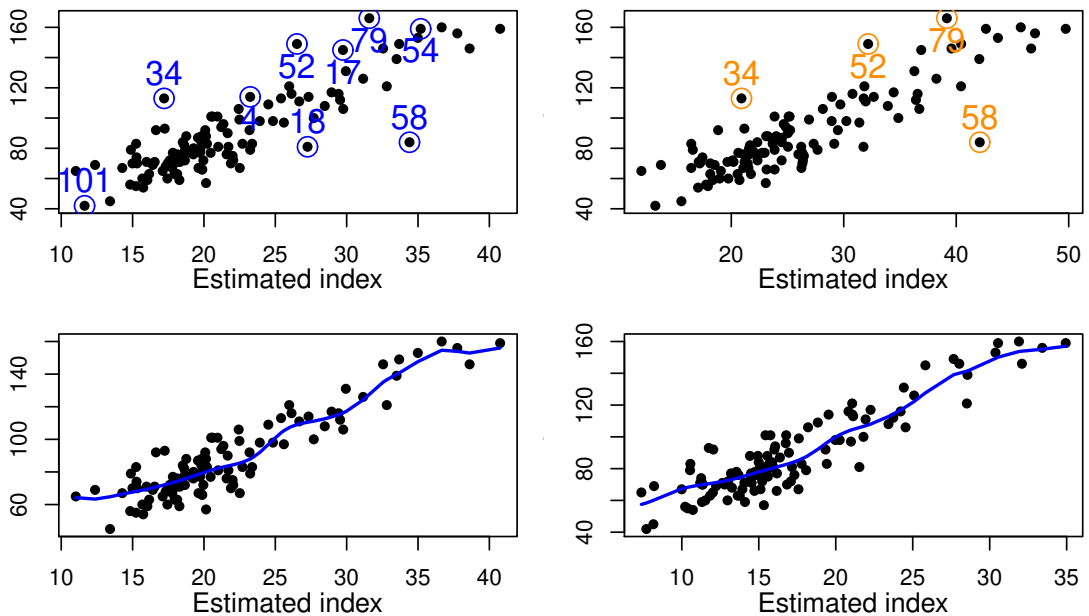


Figure – A gauche les résultats avec **TTR** et à droite avec **BOOT** (**borderline** et **outlier**)

Les **4 borderlines** correspondent à un chassé-croisé de départ en vacances.

Conclusion

Intérêt de la distinction **outlier**/**borderline**.

Réflexion sur la notion de « dynamique d'apprentissage ».

Réflexion sur les fonctions utilisées pour discriminer (`log` de l'erreur).

Applicabilité directe à des données réelles.

Article à paraître comme chapitre d'un livre, simulations à l'appui.

Code en libre accès :
`https://github.com/hlorenzo/outlierSIR`

References I

R. Dennis Cook et Frank Critchley. « Identifying Regression Outliers and Mixtures Graphically ». In : *Journal of the American Statistical Association* 95.451 (2000), p. 781-794.

Pierre-Andre Cornillon et al. *R for Statistics*. CRC press, 2012.

N. Duan et K.-C. Li. « Slicing regression : a link-free regression method ». In : *The Annals of Statistics* 19 (1991), p. 505-530.

Ursula Gather, Torsten Hilker et Claudia Becker. « A note On outlier sensitivity of Sliced Inverse Regression ». In : *Statistics* 36.4 (2002), p. 271-281.