Estimation robuste pour des modèles GLM et GLMM en population finie

Cyril Favre Martinoz, *Crest-Ensai/Irmar*
David Haziza, Université de Montréal
Nikos Tzavidis, Université de Southampton

mercredi 1er avril 2015

CREST Centre de recherche en économie et statistique

ENSAI

Sommaire

Sommaire

Estimation sur petits domaines

- Le petit domaine est souvent le résultat d'un découpage géographique fin.
- Les échantillons construits sont la plupart du temps prévus pour donner des estimations au niveau national, voire régional pour certaines enquêtes.
- C'est pourquoi on a très peu d'individus échantillonnés dans le petit domaine.
- L'approche sondage "classique" devient très périlleuse étant donnée le peu d'unités présentes dans chaque petit domaine.
- On a recours à une approche modèle qui diffère selon le niveau de disponibilité de l'information auxiliaire.
- Modélisation au niveau des petits domaines : Fay-Herriot (1979)
- Modélisation au niveau des unités : Battese, Harter et Fuller (1988)

Pourquoi faire de l'estimation robuste ?

- Les estimateurs classiques sous ces modèles sont asymptotiquement sans biais, mais ils sont très sensibles à la présence d'unités influentes.
- D'autant plus vrai dans un contexte d'estimation sur petit domaine, car la taille des échantillons est relativement faible.
- Cela permet aussi de se prémunir contre une mauvaise spécification du modèle.
- Objectif : produire un estimateur robuste aux valeurs influentes dans un contexte d'estimation sur petits domaines avec une modélisation au niveau individuel.

Sommaire

Les approches classiques

- La population U de taille N est découpée en k petits domaines.
- Un plan de sondage non informatif s est sélectionné dans la population U et il sera considéré comme fixe dans l'inférence.
- On dispose d'un vecteur de variables auxiliaires $\mathbf{x_{ij}}$ disponible pour toutes les unités j de la population dans le domaine i.
- La taille de chaque petit domaine N_i est supposée connue.
- On note n_i le nombre d'unités échantillonnées dans le petit domaine i.
- Pour chaque petit domaine i, on souhaite estimer le paramètre de population finie:

$$\theta_i = \frac{\sum_{j \in U_i} Y_{ij}}{N_i}$$

où Y_{ij} peut être une variable continue, binaire ou une variable de comptage.

Un exemple

- Par exemple, si on souhaite estimer un taux d'activité par iris i

$$\theta_i = \frac{\sum_{j \in U_i} Y_{ij}}{N_i},$$

où $Y_{ij} = 0$ si l'individu j dans l'iris i est actif et $Y_{ij} = 1$ sinon.
- On va utiliser une régression logistique pour expliquer la variable Y avec de l'information auxiliaire X et on va prédire les individus non observés.
- Si on souhaite prédire le nombre de visites dans les cabinets médicaux dans un département i

$$\theta_i = \sum_{j \in U_i} Y_{ij},$$

où $Y_{ij} =$ Nombre de visites dans le cabinet médical j du département i
- Dans ce cas, l'utilisation d'un modèle de régression Poissonien semble adaptée.
- Ces deux modèles sont des cas particuliers de Modèles Linéaires Généralisés.

Estimateur synthétique

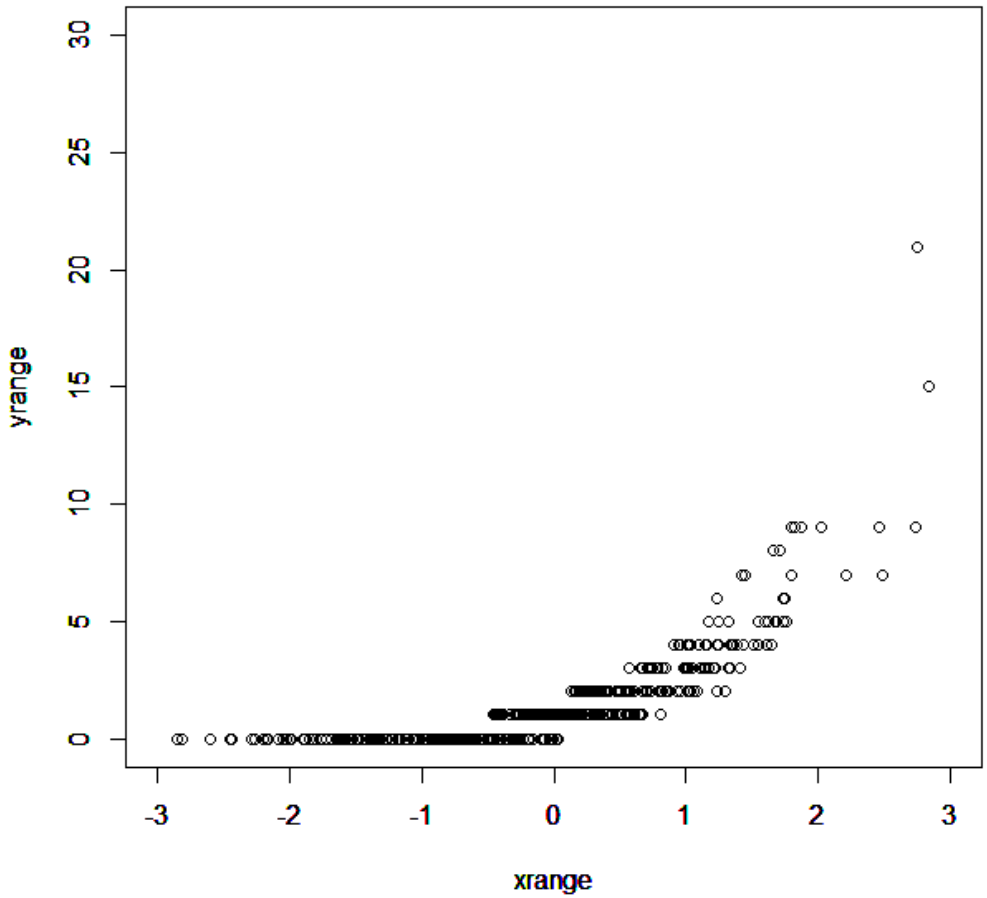
- Une première approche consiste à supposer que le modèle modélisant la variable Y est identique dans chacun des domaines

$$g(Y_{ij}) = x_{ij}^T \beta + e_{ij}$$

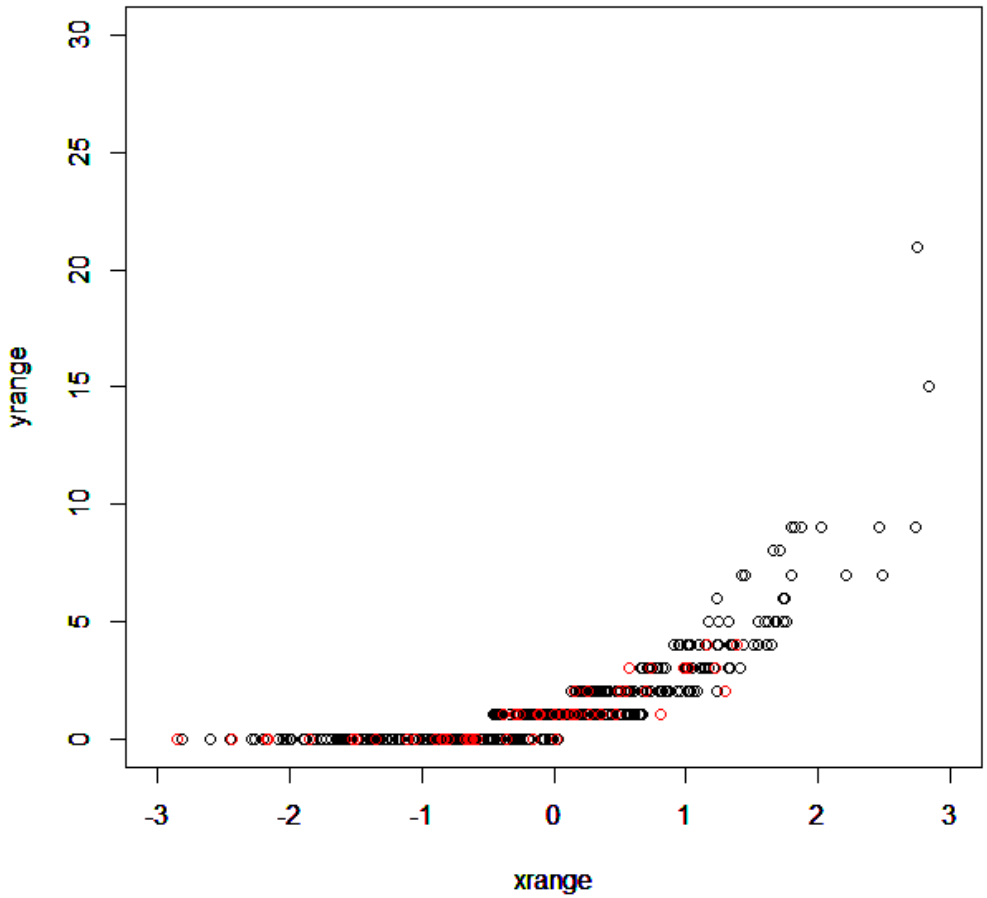
où $g(.)$ est la fonction de lien.

- On utilise une approche dite "synthétique" et on définit l'estimateur correspondant par : $\hat{\theta}_i^{SYN} = N_i^{-1} \left(\sum_{j \in S_i} Y_{ij} + \sum_{j \in U_i \setminus S_i} h(\mathbf{x_{ij}^T} \hat{\beta}) \right)$ où $h(.) = g(.)^{-1}$.
- Cet estimateur est efficace si la variabilité entre les petits domaines est entièrement expliquée par les variables auxiliaires x.
- En pratique, c'est très rarement le cas. On a plutôt recours à des modèles mixtes.

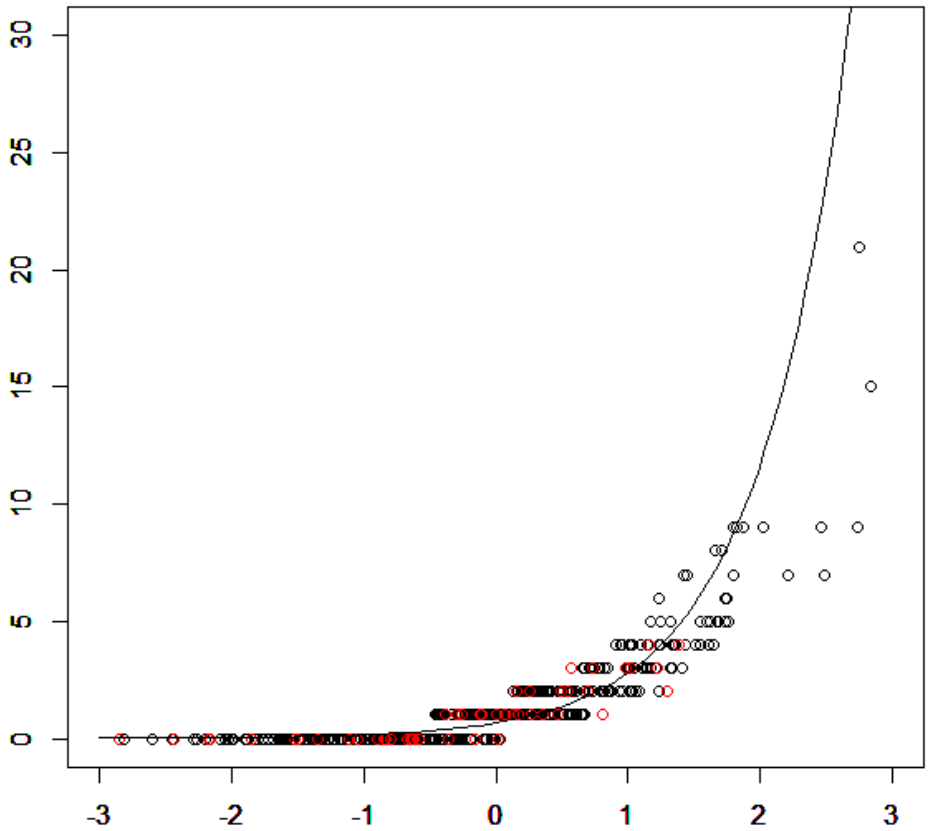
En pratique, cas favorable approche synthétique



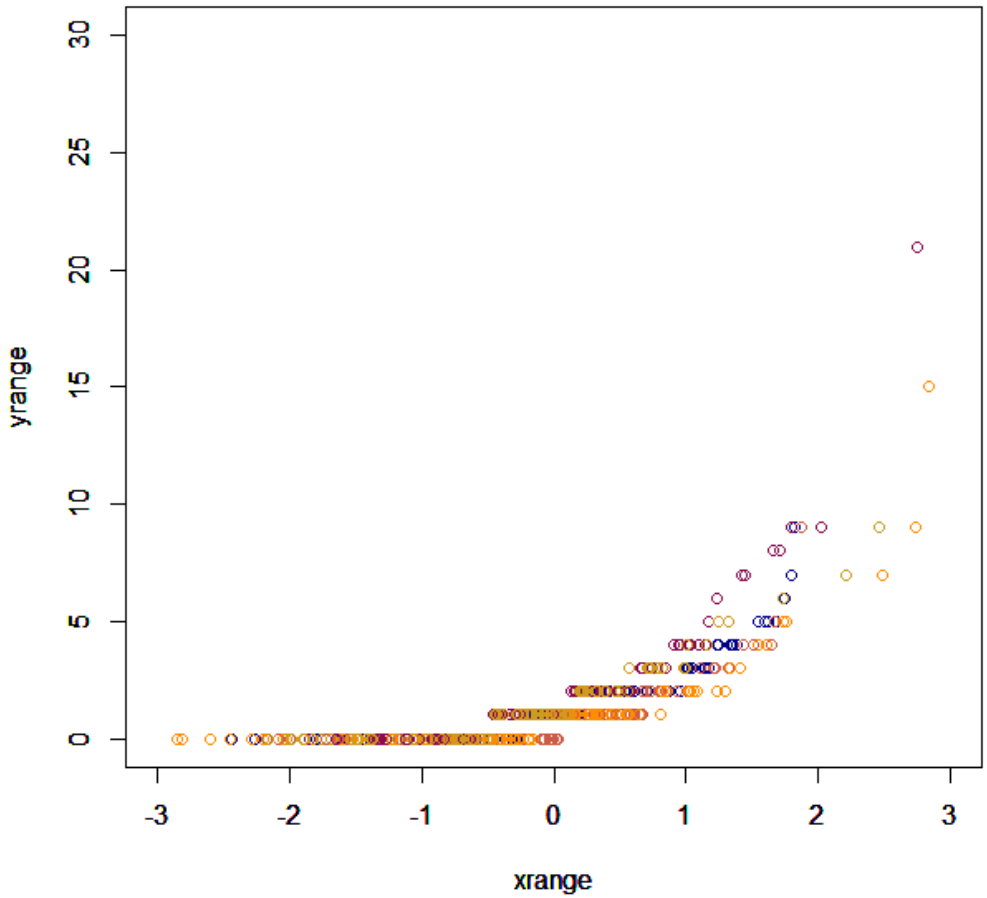
En pratique, cas favorable approche synthétique



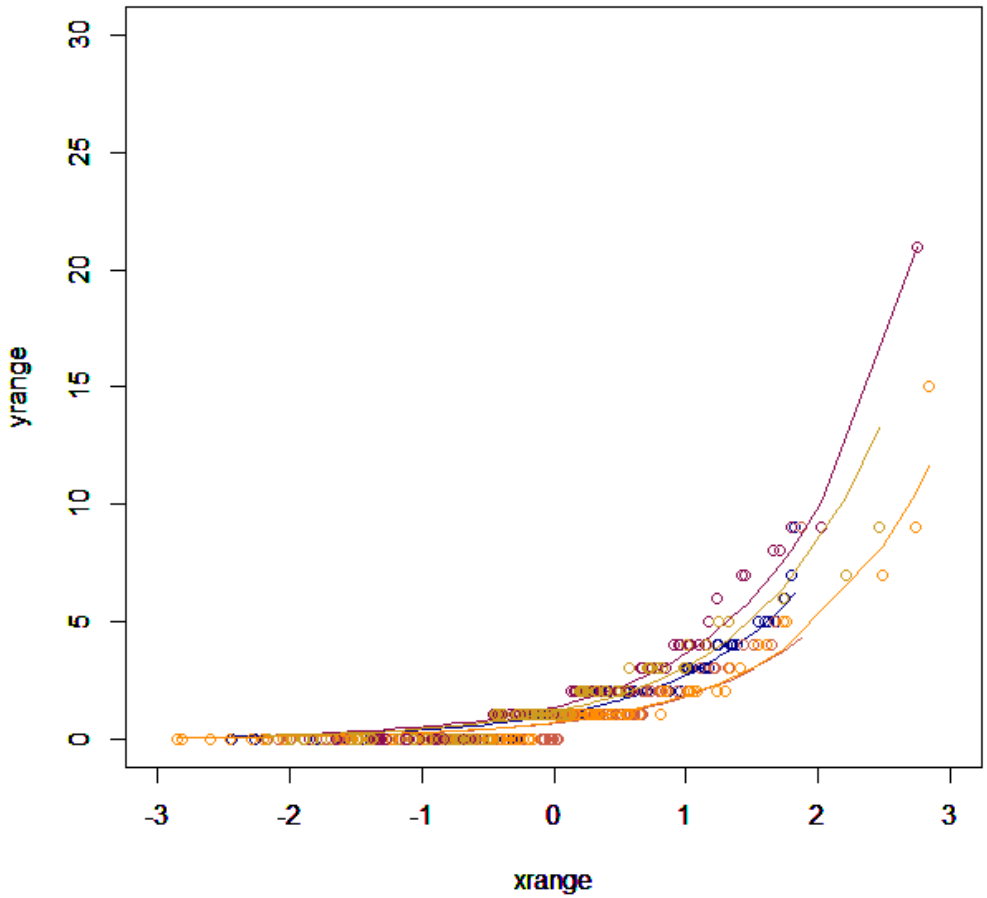
En pratique, cas favorable approche synthétique



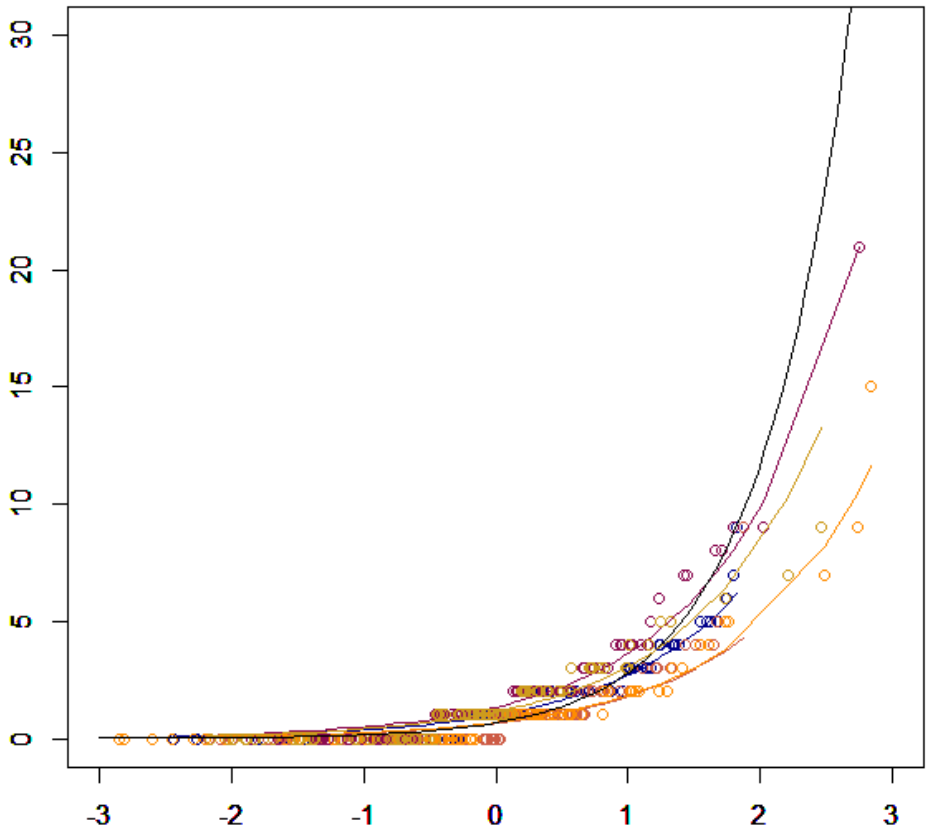
En pratique, cas favorable approche synthétique



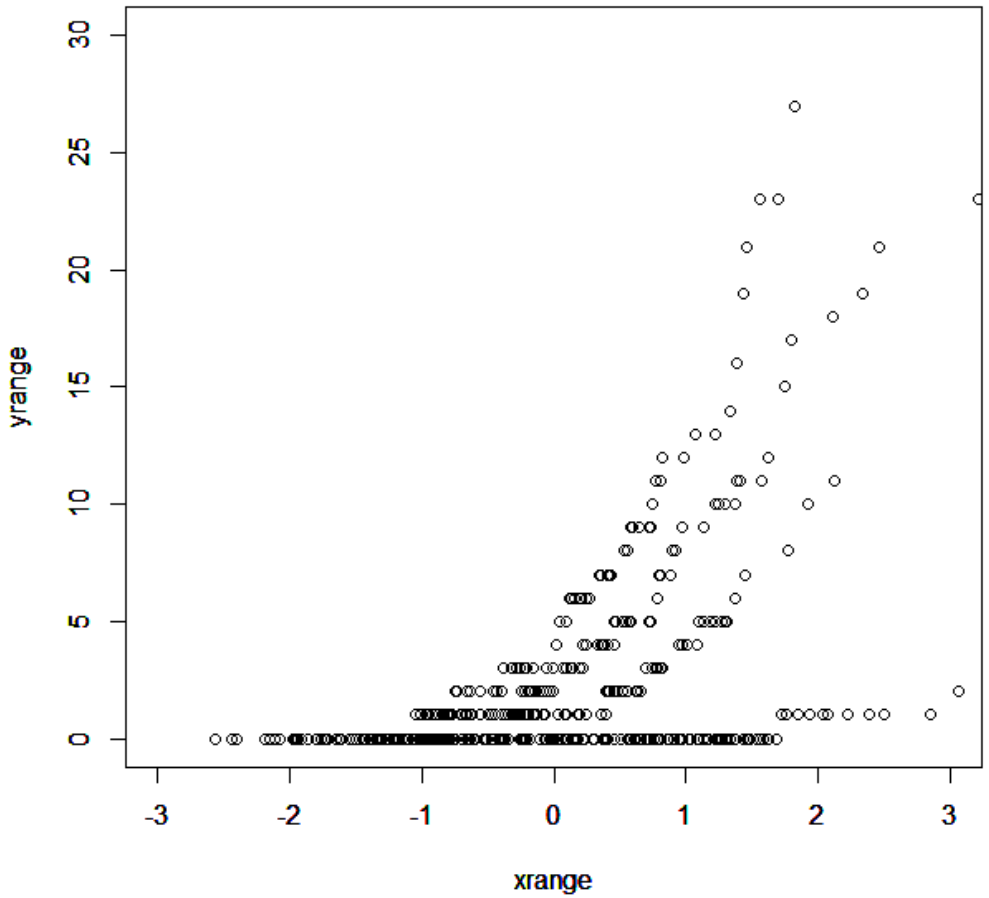
En pratique, cas favorable approche synthétique



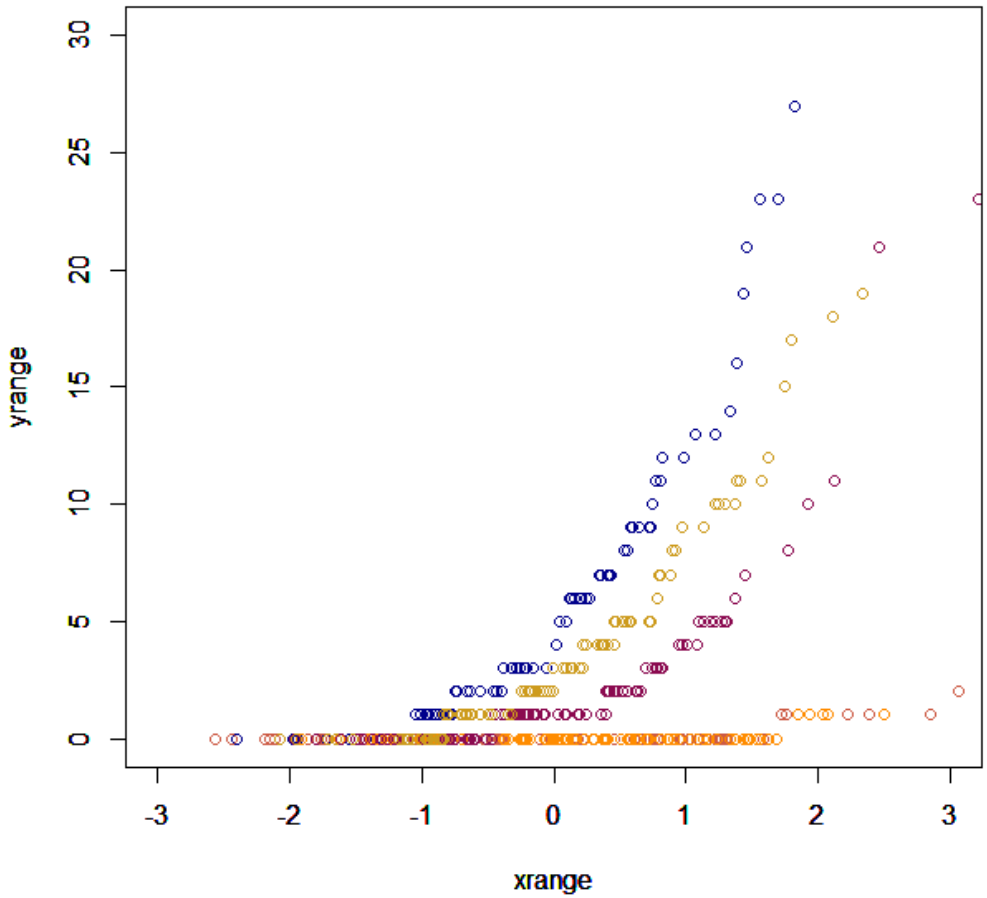
En pratique, cas favorable approche synthétique



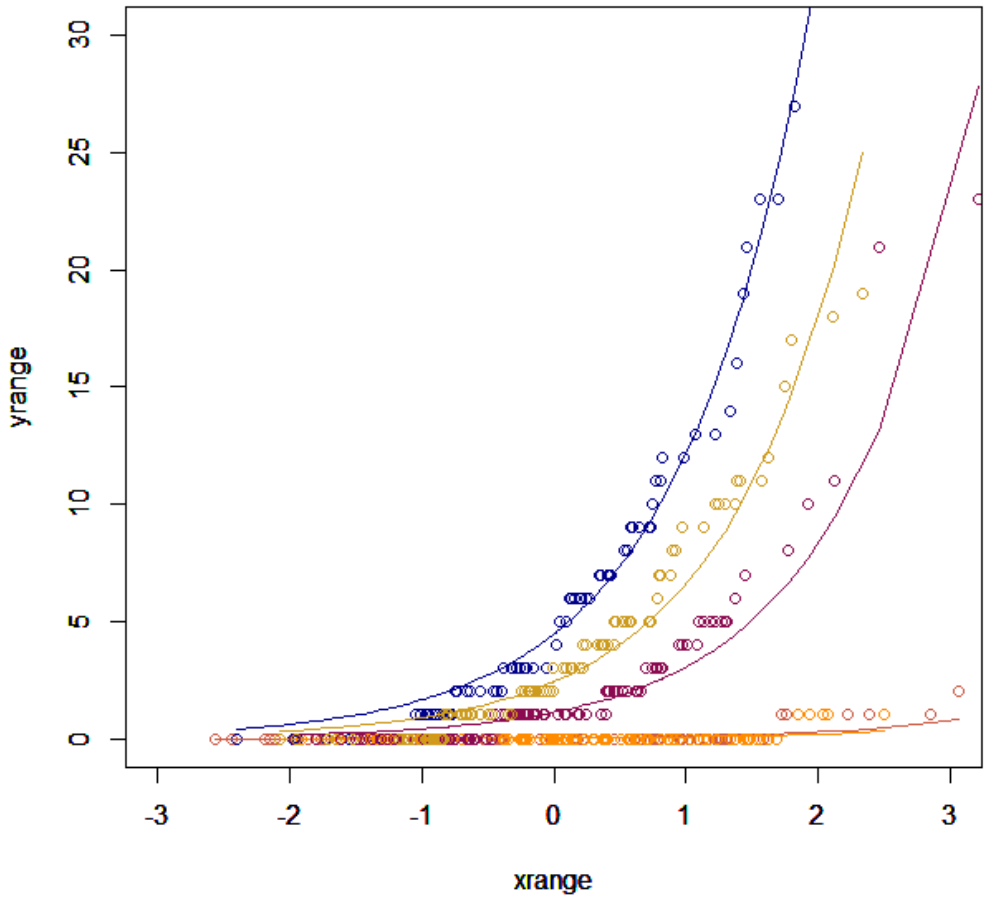
En pratique, cas défavorable approche synthétique



En pratique, cas défavorable approche synthétique



En pratique, cas défavorable approche synthétique



Approche avec effets aléatoires, estimateur Plug-In

- On utilise un modèle linéaire mixte généralisé $\mu_{ij} = E[y_{ij}|\boldsymbol{u}_i]$ de la forme

$$g(\mu_{ij}) = \eta_{ij} = x_{ij}^T\beta + u_i$$

où u_i est une variable aléatoire représentant l'effet aléatoire propre au petit domaine i.

- Le vecteur des effets mixtes u suit une loi normale de moyenne zéro et de variance constante ϕ.
- On définit l'estimateur Plug-in (Empirical Plug-in Predictor(EPP)) :

$$\hat{\theta}_i^{EPP} = N_i^{-1}\left(\sum_{j\in S_i} Y_{ij} + \sum_{j\in U_i\setminus S_i} h(\mathbf{x_{ij}^T}\hat{\beta} + \hat{u}_i)\right)$$

où $\hat{\beta}$ et $\hat{u}_i$ sont les estimateurs calculés par une procédure iterative qui combine Maximum de vraisemblance pénalisée (PQL) et Maximum de vraisemblance restreint (REML) voir Schall (1991).

Estimation robuste, objectifs

- But: Construire un estimateur robuste qui vérifie:
  - une efficacité "peu affectée" si le modèle m est mal spécifié;
  - une efficacité proche de l'estimateur optimal $\hat{\theta}^{EPP}$ si le modèle m est correct.
- Une mauvais spécification du modèle peut survenir lorsque :
  - Une petite proportion des données est générée selon un modèle différent (outliers).
  - La distribution des erreurs est très asymétrique (et non normal) $\rightarrow$ valeurs extrêmes dans les erreurs.

Les estimateurs robustes existants pour les petits domaines

- Dans le cas d'un modèle linéaire mixte : Sinha et Rao (2009), Chambers et al.(2013), Domgmo Jiongo et al. (2013)
- Dans le cas d'un modèle linéaire mixte généralisé, deux types d'estimateurs robustes ont été proposés :
  - Maiti (2001) développe un estimateur robuste à partir de modèles linéaires mixtes généralisés hiérarchiques ;
  - Chambers, Salvati & Tzavidis (2014) utilisent la régression quantile pour des variables dichotomiques ou des variables de comptage.
- Notre objectif est d'étendre la méthode développée par Domgmo Jiongo et al. (2013) basée sur le biais conditionnel pour des modèles linéaires mixtes à des modèles linéaires mixtes généralisés.

Sommaire

Définition

- Le biais conditionnel dans une approche sous le modèle est :

$$B_{ihj}\left(y_{hj}, u_h; \beta\right) = E_m\left\{\hat{\theta}_i^{EPP} - \theta_i | s,\, y_{hj},\, u\right\}.$$

- Il permet de quantifier l'influence de toutes les unités de la population.
- Dans le cas où $g \neq I$, $\hat{\theta}_i^{EPP}$ n'est plus une fonction linéaire des Y

$$\frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in U_i\setminus s_i} \hat{\mu}_{ij}\right)$$

où $\hat{\mu}_{ij} = h\left(x_{ij}^T\hat{\beta} + \hat{u}_i\right)$.

Approximation

- Le modèle initial de la forme :

$$g(\mu_{ij}) = \eta_{ij} = x_{ij}^T\beta + u_i; j = 1, ..., N_i,\ i = 1, ..., k,$$

est approché par un modèle linéaire mixte de la forme :

$$\xi_{ij} = x_{ij}^T\beta + u_i + e_{ij},\ j = 1, ..., N_i,\ i = 1, ..., k.$$

où

$$\xi_{ij} = \eta_{ij} + (y_{ij} - \mu_{ij})\, g^{'}(\mu_{ij}).$$

Plusieurs cas à distinguer

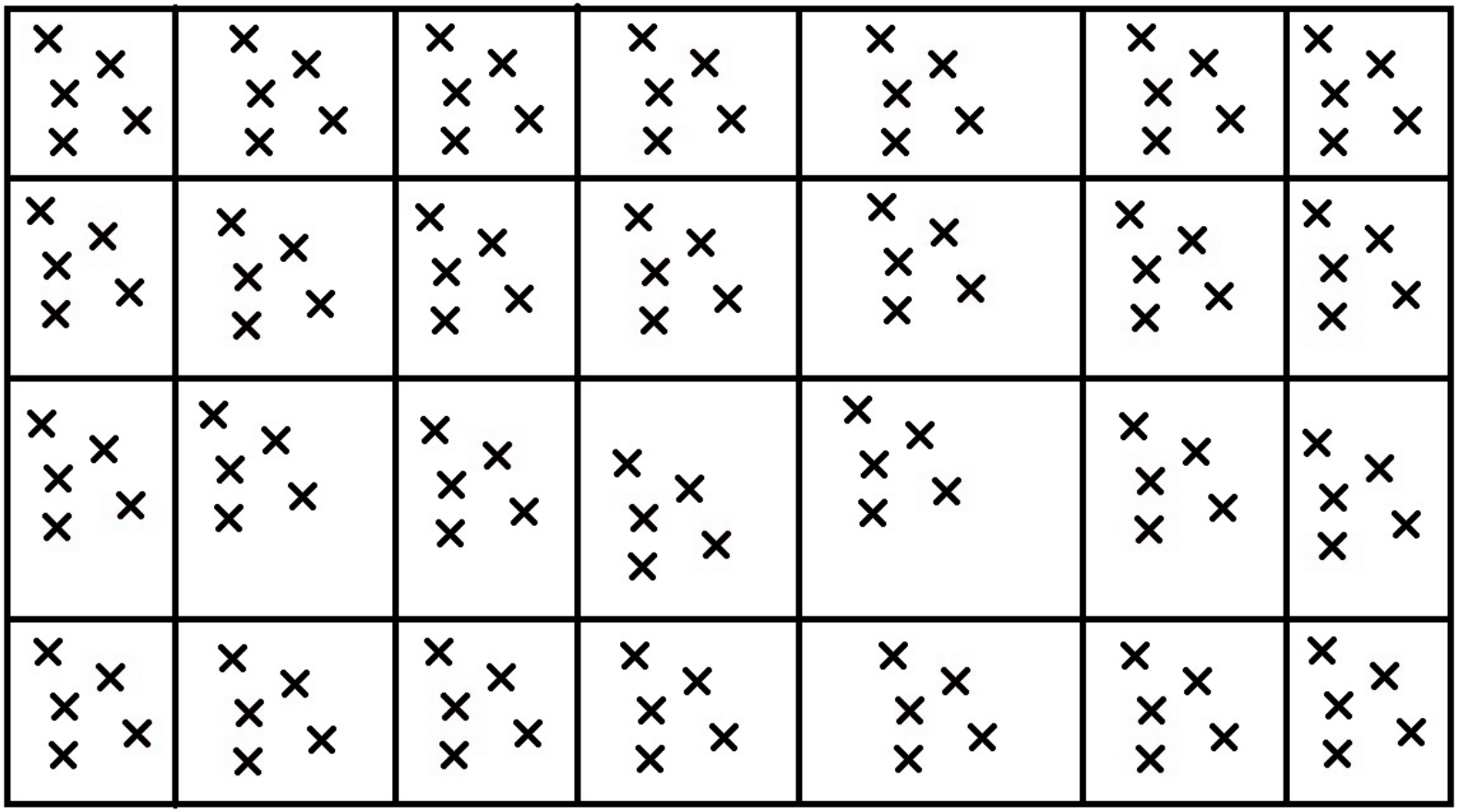


Figure 3:

Plusieurs cas à distinguer

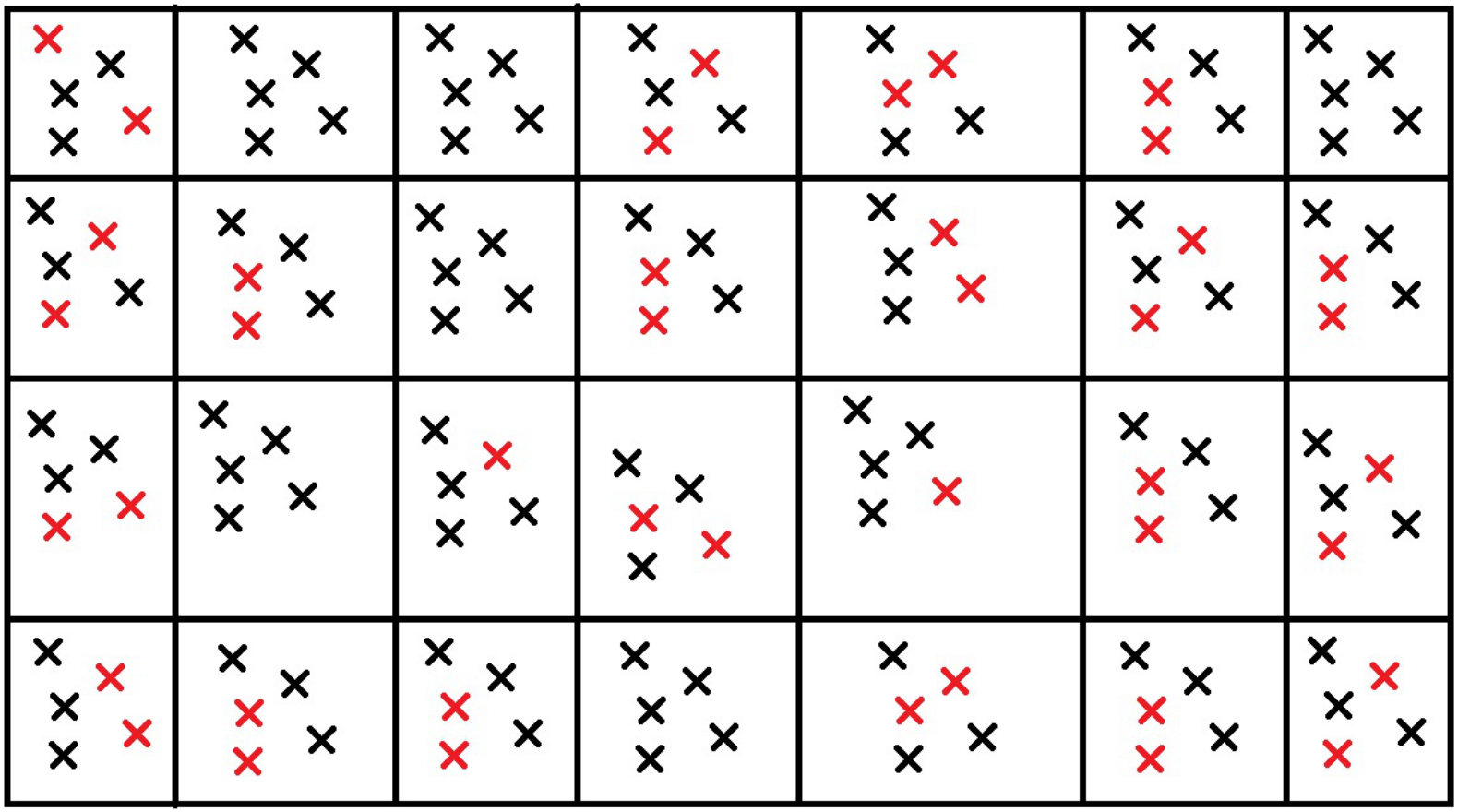


Figure 3:

Plusieurs cas à distinguer

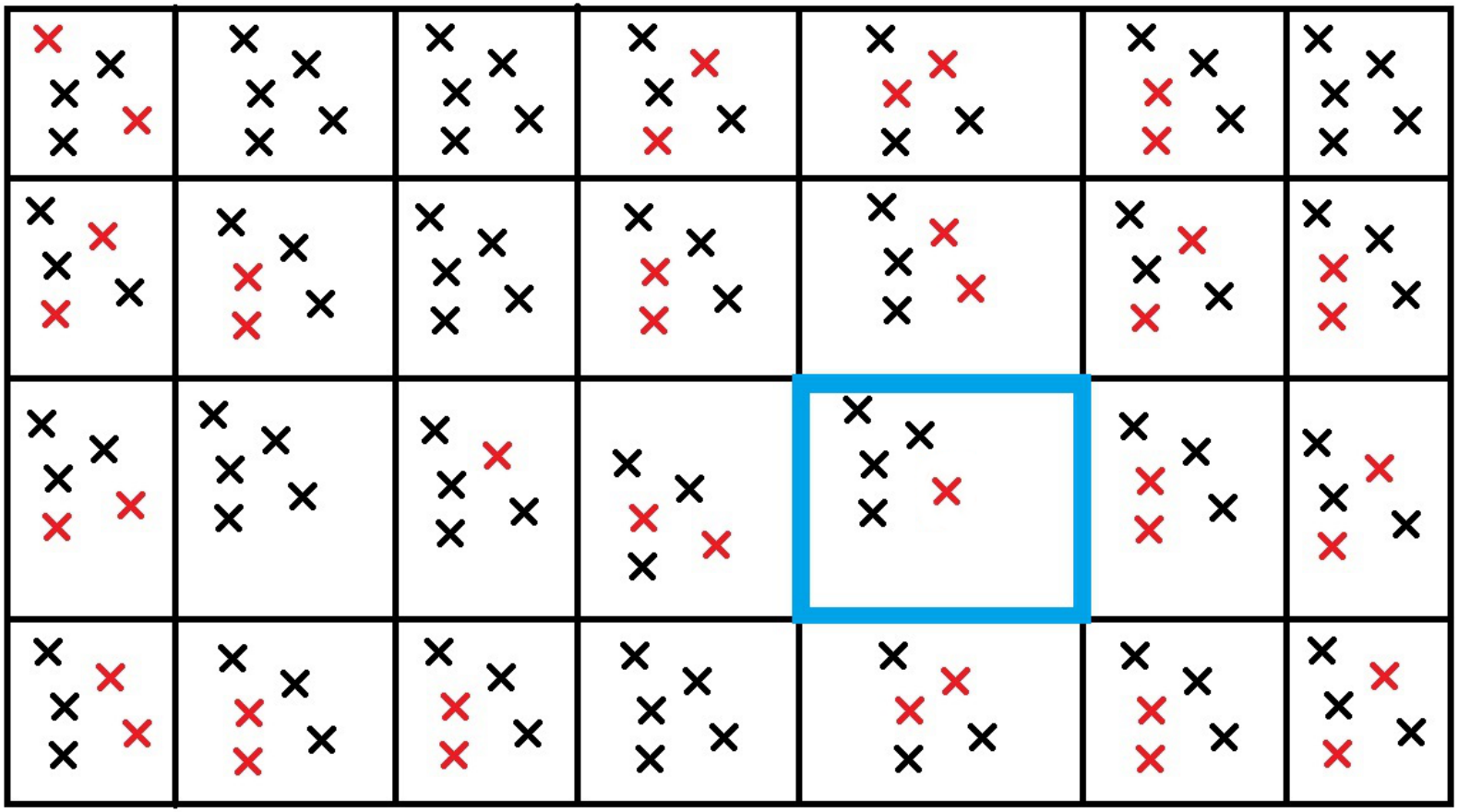


Figure 3:

Plusieurs cas à distinguer

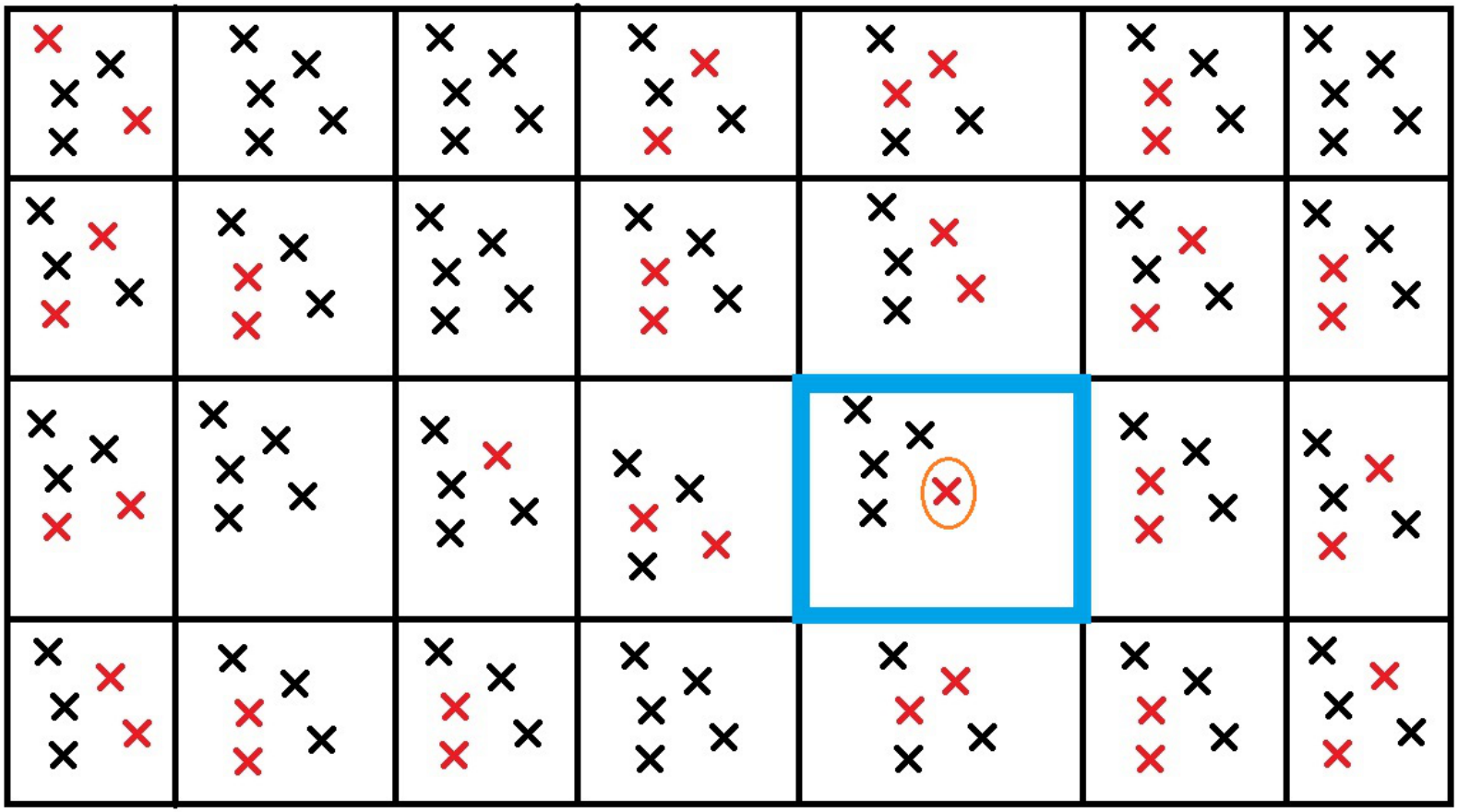


Figure 3:

Plusieurs cas à distinguer

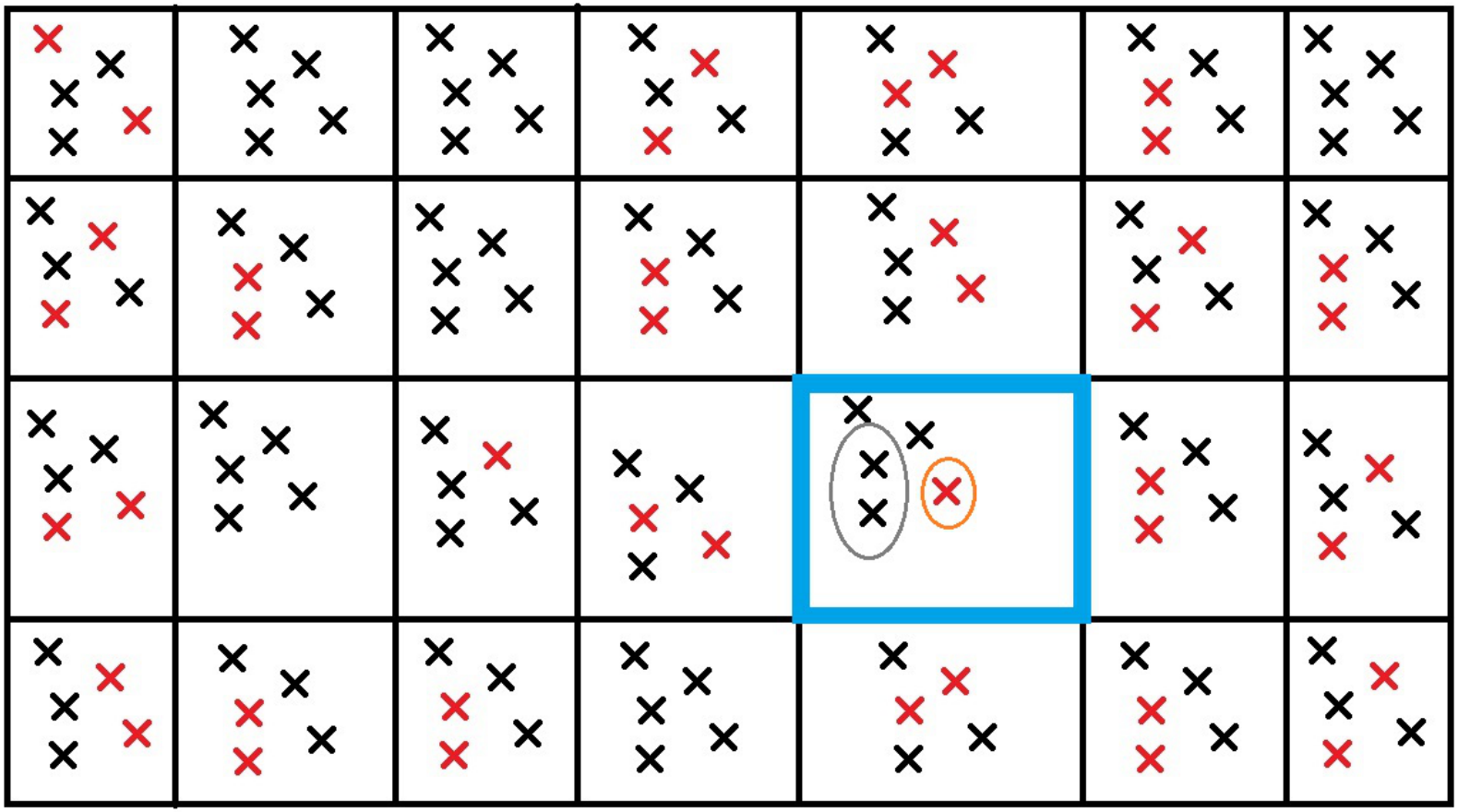


Figure 3:

Plusieurs cas à distinguer

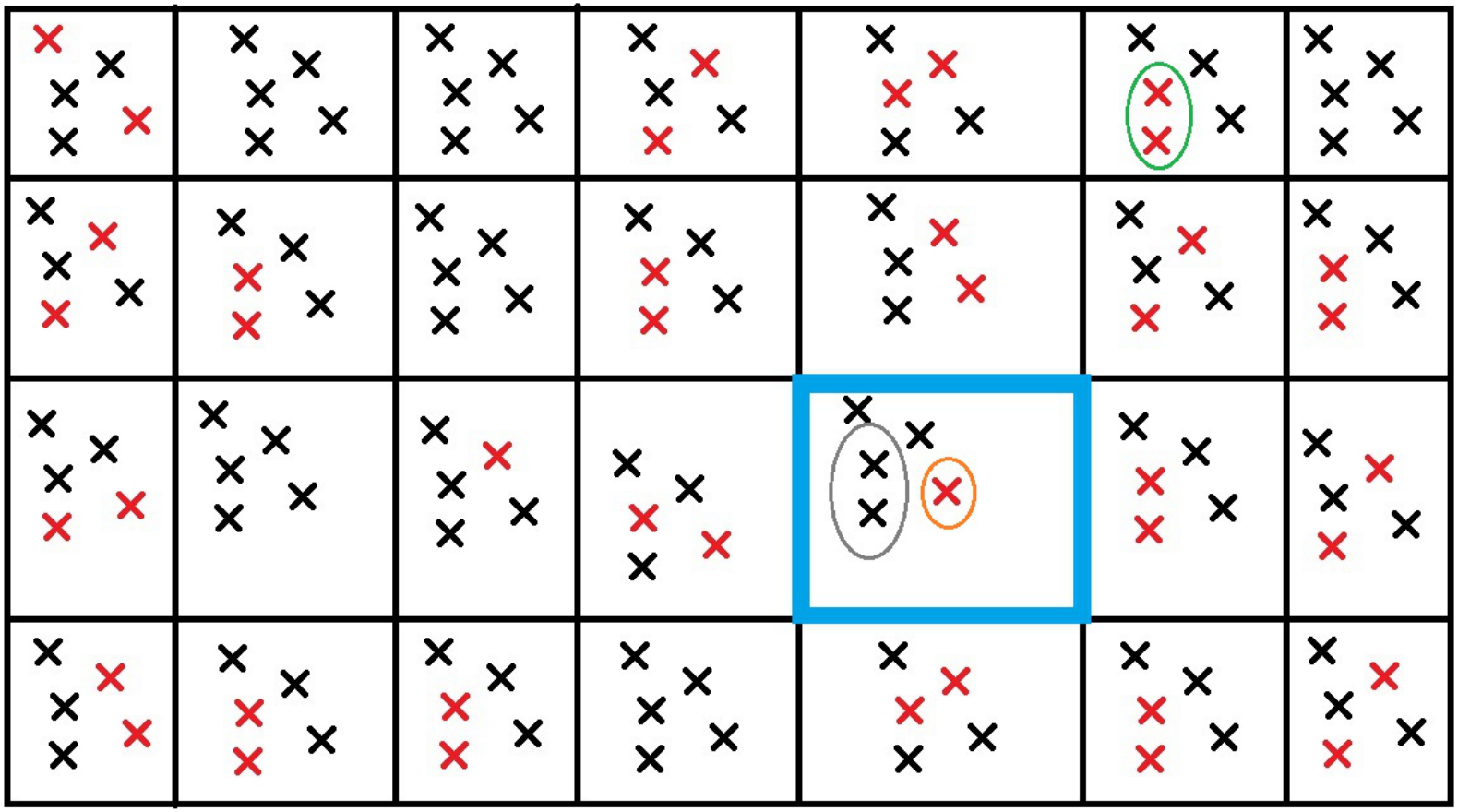


Figure 3:

Plusieurs cas à distinguer

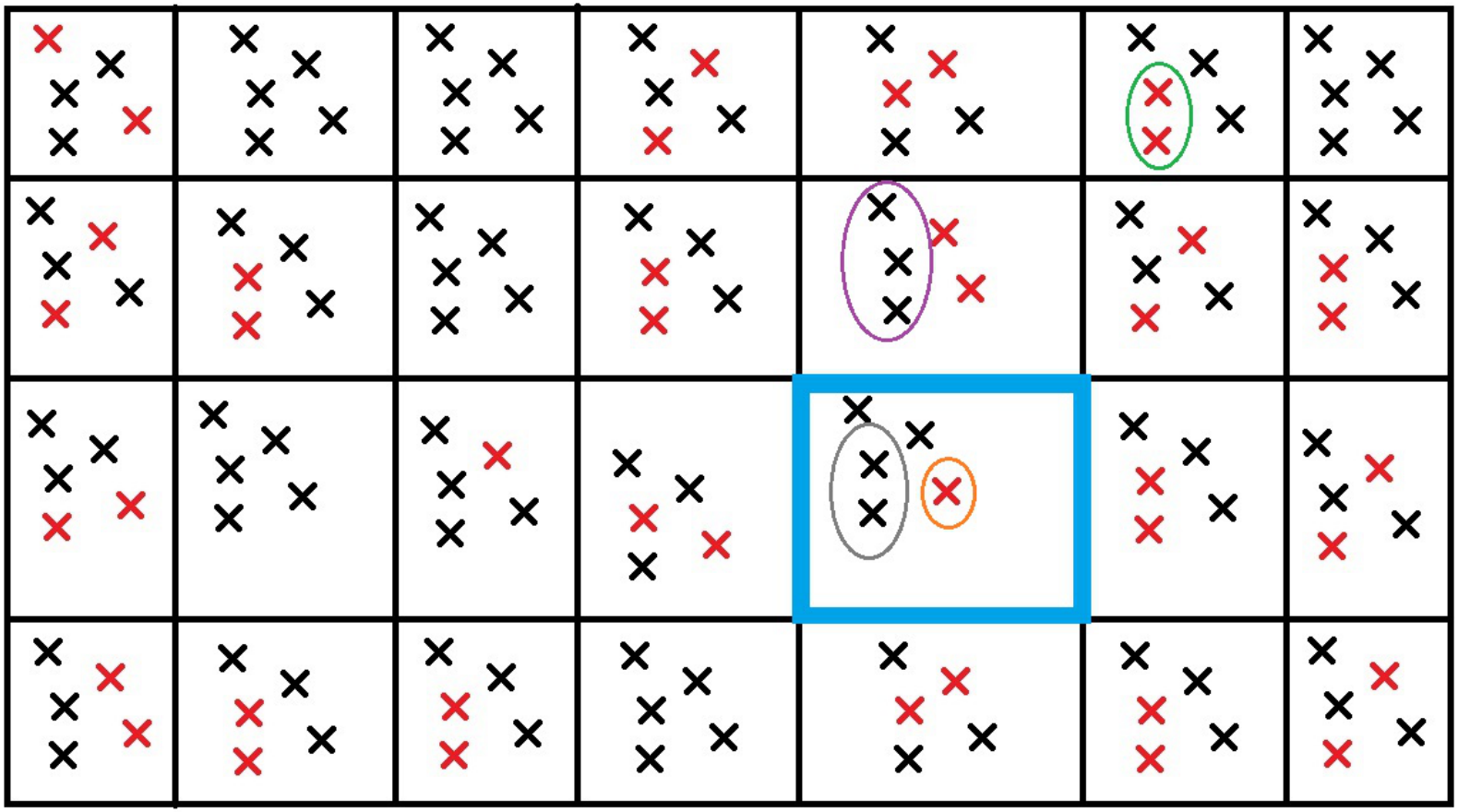


Figure 3:

Approximation du biais conditionnel pour l'estimateur EPP

$$B_{ihj}\left(y_{hj}, u_h; \beta\right) = \begin{cases} N_i^{-1}\left(\sum_{h=1}^{k}\sum_{j\in S_h} w_{ihj}u_h - \sum_{j\in U_i\setminus S_i}\frac{\partial h}{\partial \eta}\left(\eta_{ij}\right)u_i + w_{iij}e_{ij}\right) & j\in s_i \\ N_i^{-1}\left(\sum_{h=1}^{k}\sum_{j\in S_h} w_{ihj}u_h - \sum_{j\in U_i\setminus S_i}\frac{\partial h}{\partial \eta}\left(\eta_{ij}\right)u_i + w_{ihj}e_{hj}\right) & j\in s_h, h\neq i \\ N_i^{-1}\left(\sum_{h=1}^{k}\sum_{j\in S_h} w_{ihj}u_h - \sum_{j\in U_i\setminus S_i}\frac{\partial h}{\partial \eta}\left(\eta_{ij}\right)u_i - \frac{\partial h}{\partial \eta}\left(\eta_{ij}\right)e_{ij}\right) & j\in U_i\setminus s_i \\ N_i^{-1}\left(\sum_{h=1}^{k}\sum_{j\in S_h} w_{ihj}u_h - \sum_{j\in U_i\setminus S_i}\frac{\partial h}{\partial \eta}\left(\eta_{ij}\right)u_i\right) & j\in U_h\setminus s_h, h\neq i. \end{cases}$$

où

$$w_{ihj} = \begin{cases} k^{-1}a_i X_h^T C_h^{(j)} & j\in s_h \\ k^{-1}a_i X_i^T C_i^{(j)} + \left[\sum_{j'\in U_i\setminus S_i}\frac{\partial h}{\partial \eta}\left(\eta_{ij'}\right)\right]\sigma_u^2 1_{n_i}^T C_i^{(j)} & j\in s_i, \end{cases}$$

et

$$a_i = \left\{\sum_{j\in U_i\setminus S_i}\frac{\partial h}{\partial \eta}\left(\eta_{ij}\right)\left[x_{ij}^T - \sigma_u^2 1_{n_i}^T V_i^{-1}X_i\right]\right\}\left\{k^{-1}\sum_{i=1}^{k} X_i^T V_i^{-1}X_i\right\}^{-1}.$$

Sommaire

Erreur de prédiction et estimateur robuste

L'erreur de prédiction de l'estimateur EPP est approximativement

$$\hat{\theta}_i^{EPP} - \theta_i \approx \sum_{h=1}^{k} \sum_{j \in U_h} B_{ihj}\left(y_{hj}, u_h; \beta\right) - \frac{N-1}{N_i}\left[\sum_{h=1}^{k} \sum_{j \in S_h} w_{ihj} u_h - \sum_{j \in U_i \setminus S_i} \frac{\partial h}{\partial \eta}\left(\eta_{ij}\right) u_i\right].$$

En suivant Domgmo Jiongo et al. (2013), on propose un estimateur robuste de la forme

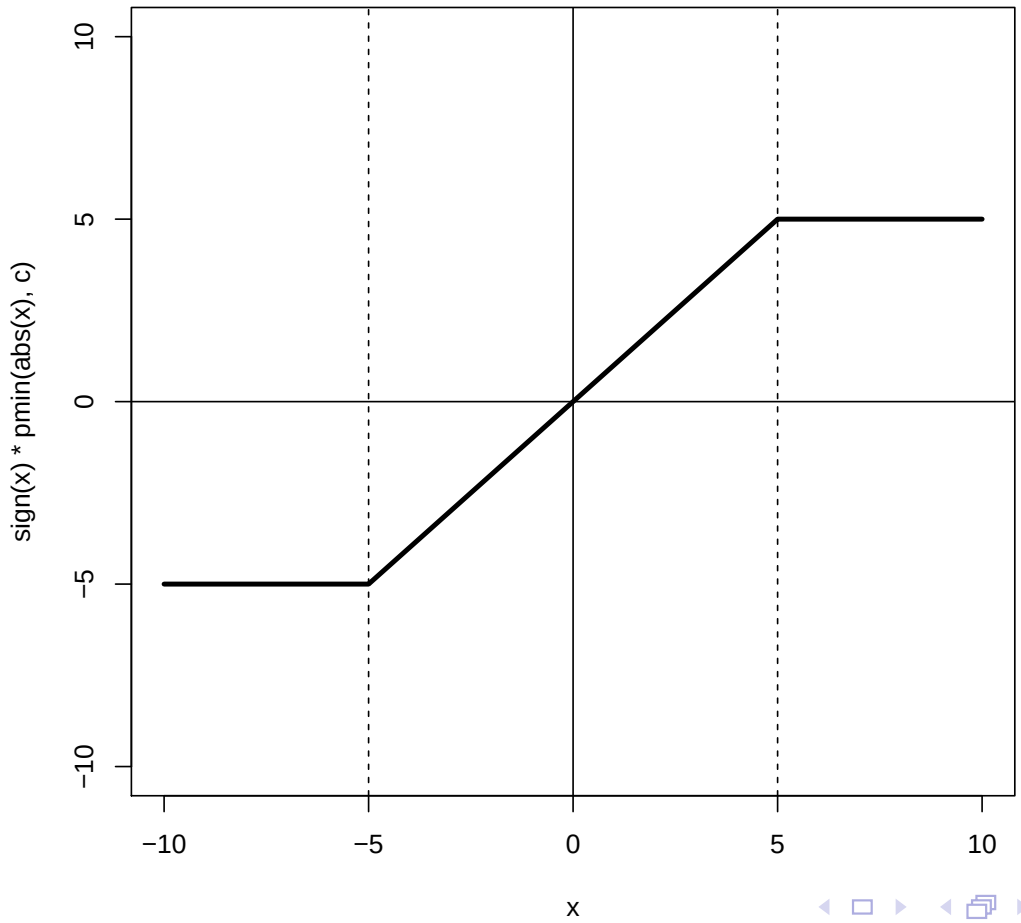
$$\hat{\theta}_i^{REBP} = \hat{\theta}_i^{EPP} - \sum_{h=1}^{k} \sum_{j \in S_h} B_{ihj}^{'}\left(y_{hj}, u_h; \beta\right) + \sum_{h=1}^{k} \sum_{j \in S_h} \psi_c\left\{B_{ihj}^{'}\left(y_{hj}, u_h; \beta\right)\right\},$$

où

$$B_{ihj}^{'}\left(y_{hj}, u_h; \beta\right) = \begin{cases} N_i^{-1} \psi_c\left\{w_{iij} e_{ij}\right\} & j \in s_i \\ N_i^{-1} \psi_c\left\{w_{ihj} e_{hj}\right\} & j \in s_h, h \neq i. \end{cases}$$

$\psi(.)$ est la fonction de Huber et c est la tuning constante qui effectue le compromis entre biais et variance.

Erreur de prédiction et estimateur robuste



Choix de la tuning constante

- On utilise un critère de type Min-Max.
- On cherche la constante c qui minimise le maximum des influences estimées sur l'estimateur robuste $\hat{\theta}^{REPP}$ pour toutes les unités j de l'échantillon:

$$\min_c \left(\max\{ \hat{B}_{ihj}^{REPP} | j \in S \} \right)$$

- En utilisant la tuning constante c solution du problème de minimisation, on peut montrer que l'estimateur robuste s'écrit

$$\hat{\theta}_i^{REPP} = \hat{\theta}_i^{EPP} - \frac{1}{2} \left(\hat{B}'_{max} + \hat{B}'_{min} \right)$$

où

$$\hat{B}'_{max} = \max_{j \in S} \left\{ \hat{B}'_{ihj} \left(y_{hj}, u_h; \beta \right) \right\}$$

et

$$\hat{B}'_{min} = \min_{j \in S} \left\{ \hat{B}'_{ihj} \left(y_{hj}, u_h; \beta \right) \right\}$$

Sommaire

Modèle linéaire et échantillonnage

- On a construit plusieurs modèles permettant de générer plusieurs populations de taille $N = 2000$ découpées en 40 petits domaines de taille égale $N_i = 125$. Les jeux de populations sont des mélanges entre un modèle linéaire mixte et un modèle qui produit des unités influentes.
- On a généré $P = 50000$ réalisations de modèles et on a sélectionné un sondage aléatoire simple stratifié de même taille dans chacun des petits domaines $n_i = 5$.
- On compare l'efficacité de l'estimateur robuste proposé en terme de biais relatif et d'efficacité avec les estimateurs robustes existants Sinha et Rao (2009), Chambers et al.(2013), Domgmo Jiongo et al. (2013).

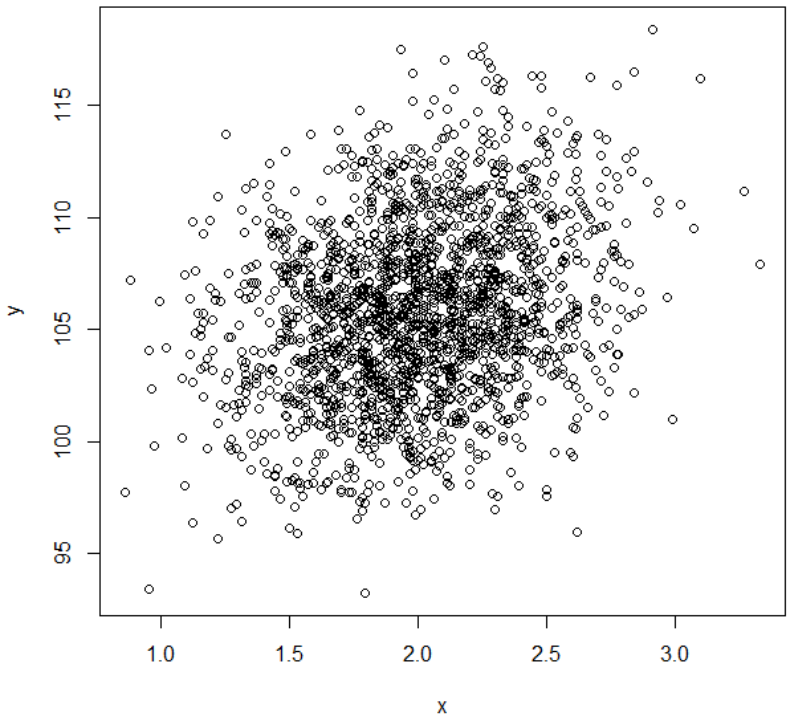
Résultats pour le cas linéaire

$$BR_{MC}\left(\hat{\theta}_p^R\right) = \frac{1}{P}\sum_{p=1}^{P}\frac{\left(\hat{\theta}_p^R - \theta_p\right)}{\theta_p} \times 100$$

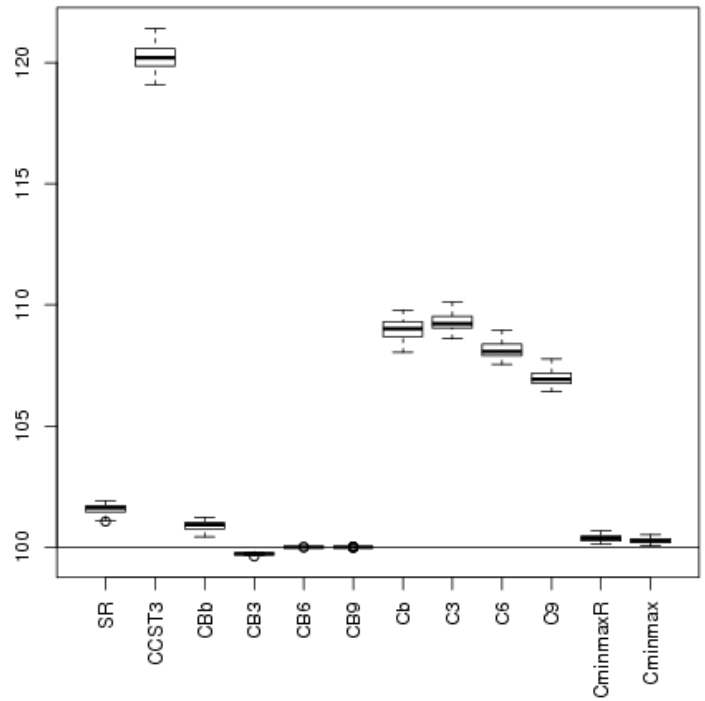
et

$$ER_{MC}\left(\hat{\theta}_p^R, \hat{\theta}\right) = \frac{\frac{1}{P}\sum_{p=1}^{P}\left(\hat{\theta}_p^R - \theta_p\right)^2}{\frac{1}{P}\sum_{p=1}^{P}\left(\hat{\theta}_p - \theta_p\right)^2} \times 100.$$

Résultats pour le cas linéaire



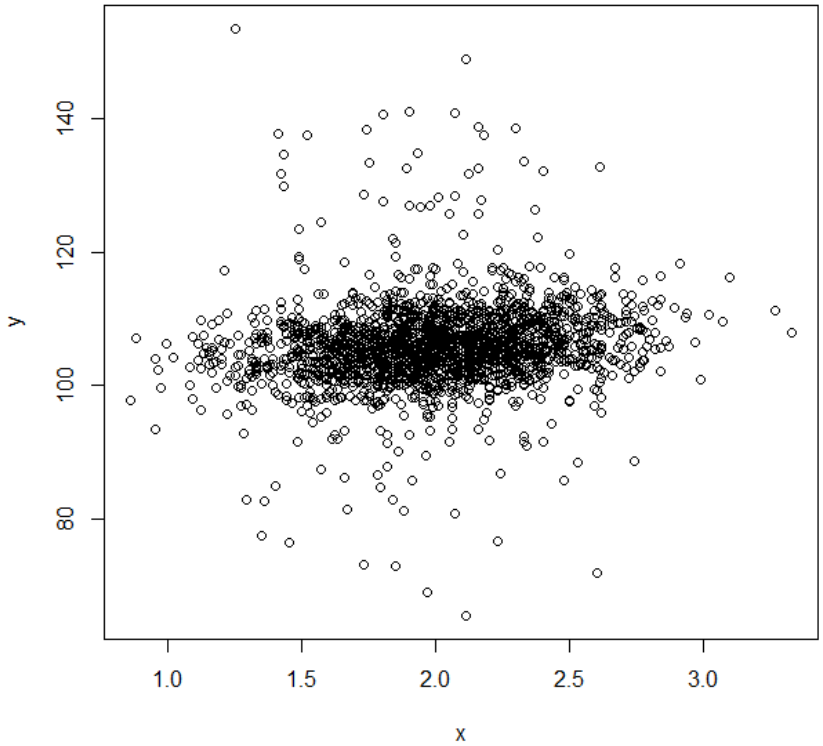
(a) Population 1



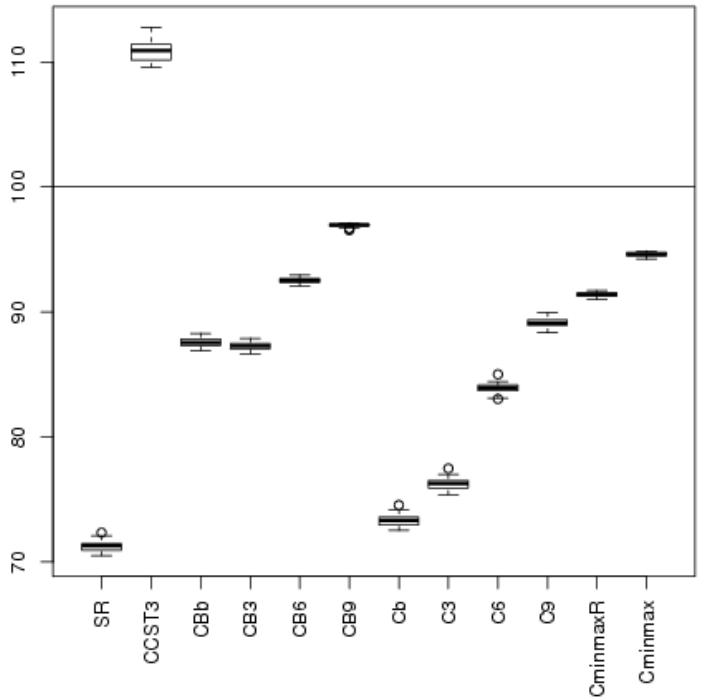
(b) Boxplot pour le cas sans outliers

Figure 5: Résultats cas linéaire

Résultats pour le cas linéaire



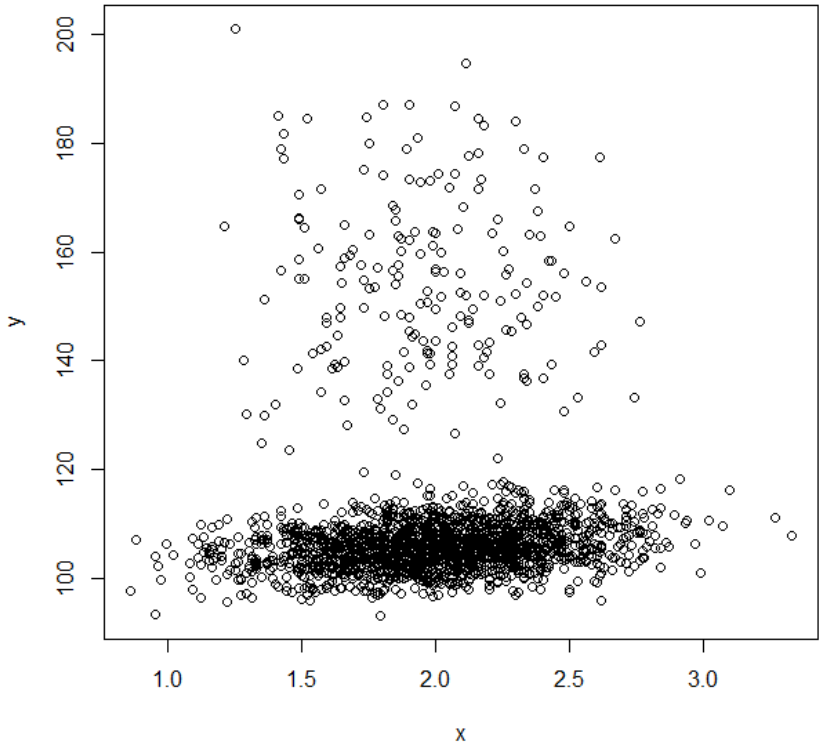
(a) Population 2



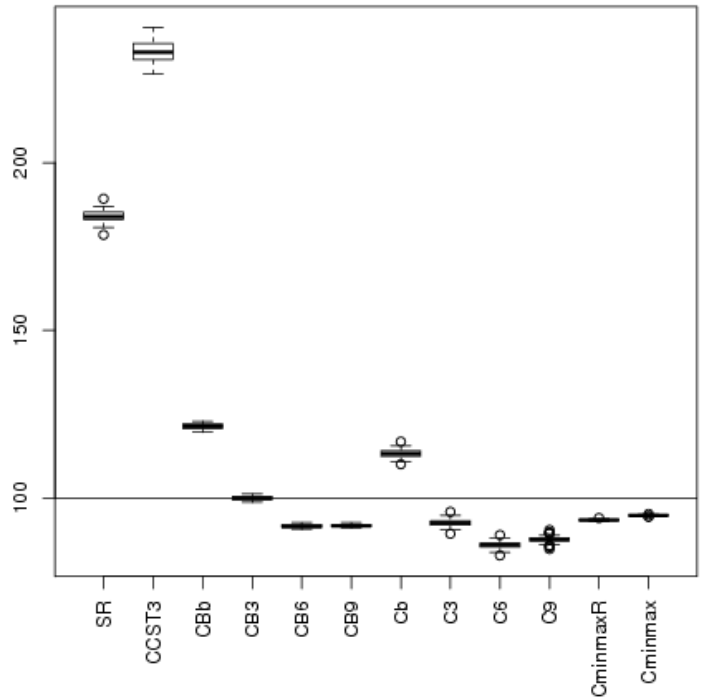
(b) Boxplot outliers effets mixtes

Figure 5: Résultats cas linéaire

Résultats pour le cas linéaire



(a) Population 3



(b) Boxplot outliers effets mixtes et pente

Figure 5: Résultats cas linéaire

Modèle Poissonien et échantillonnage

- On a construit plusieurs modèles permettant de générer plusieurs populations de taille $N = 5000$ découpées en 50 petits domaines de taille égale $N_i = 100$. Les jeux de populations sont des mélanges entre un modèle Poissonien et un modèle qui produit des unités influentes.
- On a généré $P = 5000$ réalisations de modèles et on a sélectionné un sondage aléatoire simple stratifié de même taille dans chacun des petits domaines $n_i = 10$.
- On compare l'efficacité de l'estimateur robuste proposé en terme de biais relatif et d'efficacité avec un autre estimateur utilisant la régression M-Quantile Tzavidis(2014).

Les populations pour le cas Poissonien

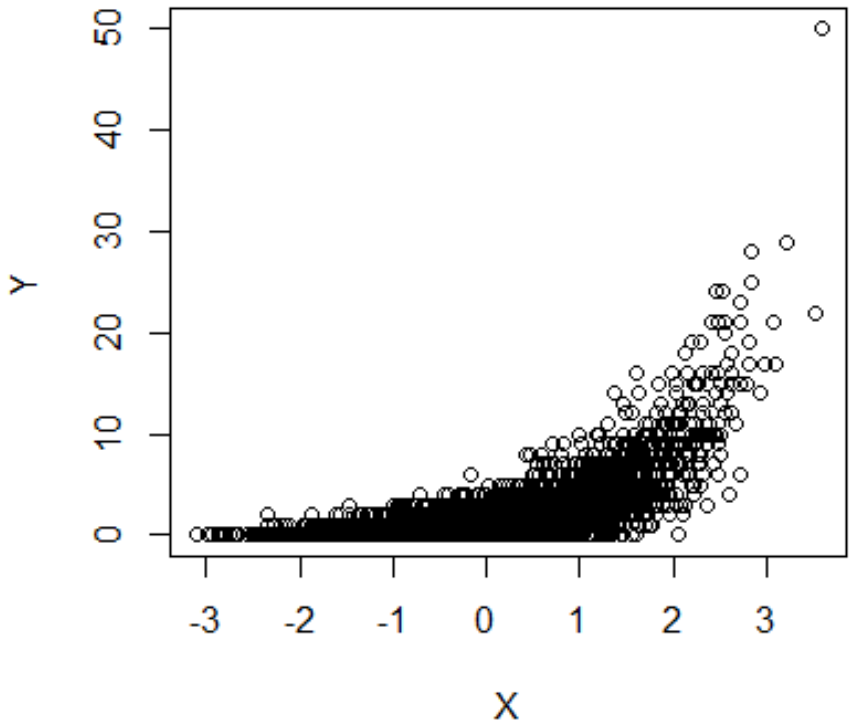


Figure 6: Cas sans outliers

Les populations pour le cas Poissonien

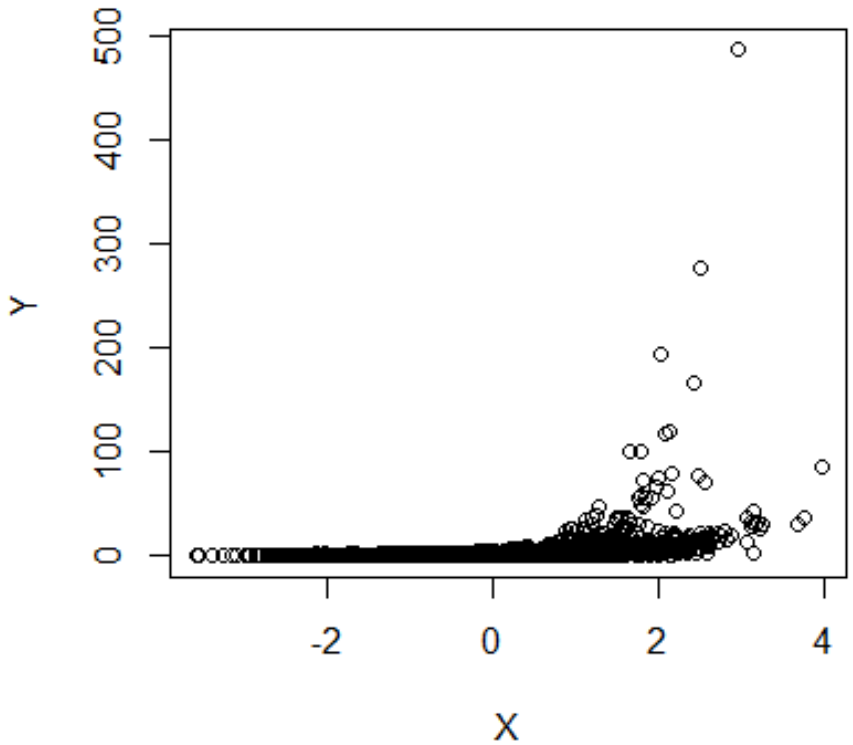


Figure 6: Cas modèle de Mélange

Les populations pour le cas Poissonien

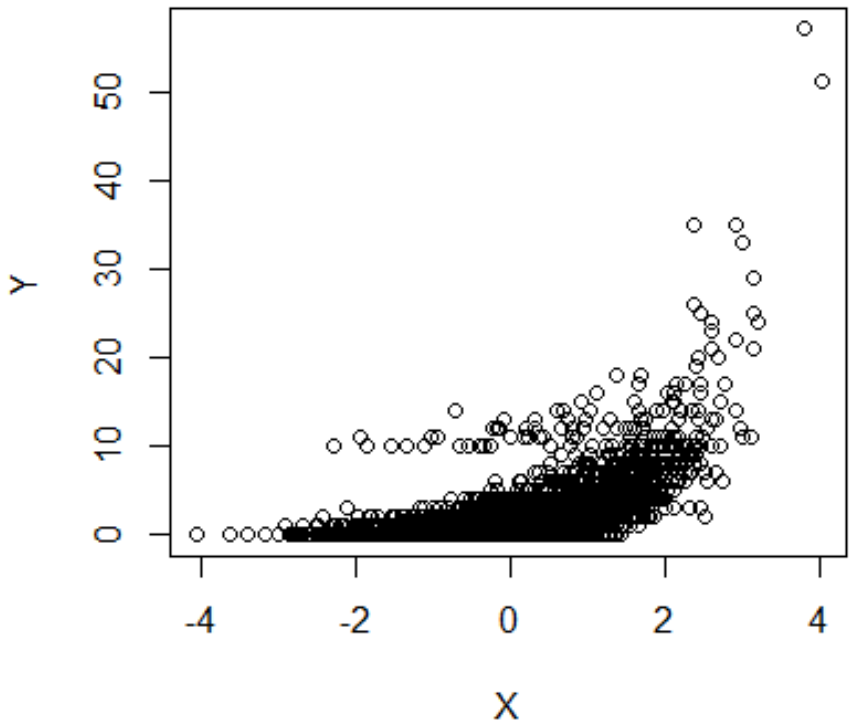


Figure 6: Cas erreur de mesure

Résultats pour le cas Poissonien

Scenarios	Taux outliers	$\hat{\theta}^{M-Quantile}$	$\hat{\theta}^{REPP}$	$\hat{\theta}^{Direct}$
Sans outliers	0	$-1.25(174)$	$-1.62(101)$	$-0.17(554)$
Modèle de mélange	0.1	$-4.55(105)$	$-2.13(90)$	$-0.17(400)$
	0.05	$-2.89(125)$	$-1.90(93)$	$-0.15(513)$
	0.01	$-1.48(160)$	$-1.58(99)$	$0.31(543)$
Outliers Erreur de mesure	0.1	$-22.97(84)$	$-4.14(91)$	$0.09(274)$
	0.05	$-15.09(77)$	$-3.65(87)$	$0.39(400)$
	0.01	$-4.86(125)$	$-2.47(89)$	$0.005(438)$

Table 1: Biais et efficacité médiane sur l'ensemble des domaines

Sommaire

Conclusion

- Le biais conditionnel est une mesure d'influence qui tient compte du modèle, du paramètre et de l'estimateur.
- On a proposé une version robuste de l'EPP adaptable pour tous les modèles de type GLMM.
- L'estimateur robuste proposé est aussi efficace que l'EPP si le modèle est bien spécifié.
- L'estimateur robuste proposé semble plus performant que l'EPP et l'estimateur M-Quantile dans le cas d'un modèle de mélange.
- Perspectives : Étendre ces techniques à une modélisation au niveau des domaines et tester des méthodes de Bootstrap pour estimer l'erreur quadratique moyenne de l'estimateur.

Bibliography I

Ray Chambers, Hukum Chandra, Nicola Salvati, and Nikos Tzavidis.

Outlier robust small area estimation.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1):47–69, 2014.

Wenceslao González-Manteiga, María José Lombardía, Isabel Molina, Domingo Morales, and L Santamaría.

Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model.
Computational statistics & data analysis, 51(5):2720–2733, 2007.

V Dongmo Jiongo, D Haziza, and P Duchesne.

Controlling the bias of robust small-area estimators.
Biometrika, page ast030, 2013.

Ayoub Saei and Ray Chambers.

Small area estimation under linear and generalized linear mixed models with time and area effects.
2003.

Sanjoy K Sinha and JNK Rao.

Robust small area estimation.
Canadian Journal of Statistics, 37(3):381–399, 2009.

Nikos Tzavidis, M Giovanna Ranalli, Nicola Salvati, Emanuela Dreassi, and Ray Chambers.

Robust small area prediction for counts.
Statistical methods in medical research, page 0962280214520731, 2014.