Estimateurs model-based non paramétriques de la fonction de répartition d'une variable censurée à droite sur petits domaines

Sandrine CASANOVA & Eve LECONTE
TSE (GREMAQ)
Université TOULOUSE 1 Capitole

7ème colloque francophone sur les sondages, Rennes
6 novembre 2012

Introduction

Soit U une population de taille N et s un échantillon de taille n

t_j valeur de la variable d'intérêt pour le $j$ème individu de la population U, éventuellement censurée par c_j.
Sur l'échantillon s, on observe $y_j = min(t_j, c_j)$ et $\delta_j = \mathbb{1}(t_j < c_j)$

Information auxiliaire au niveau individuel : covariable x_j continue mesurée pour le $j$ème individu de U (x_j connu sur U)

Objectif : estimer la fonction de répartition (fdr) sur U

$$F(t) = \frac{1}{N}\sum_{j\in U}\mathbb{1}(t_j \leq t) = \frac{1}{N}\left(\sum_{j\in s}\mathbb{1}(t_j \leq t) + \sum_{j\in U\setminus s}\mathbb{1}(t_j \leq t)\right)$$

On suppose que le plan d'échantillonnage est non informatif
$\hookrightarrow$ estimateurs *model-based*

Approche directe : on estime F par l'estimateur de Kaplan-Meier (généralise la fdr empirique au cas censuré)

$$\widehat{F}_{\mathrm{KM}}(t) = \begin{cases} 1 - \prod_{j=1}^{n} \left\{ 1 - \dfrac{1}{\sum_{r=1}^{n} \mathbb{1}(y_r \geq y_j)} \right\}^{\mathbb{1}(y_j \leq t, \delta_j = 1)} & \text{if} \quad t < y_{(n)} \\ 1 & \text{sinon.} \end{cases}$$

Amélioration de l'estimation de F
$\hookrightarrow$ nécessité de prédire $\mathbb{1}(t_j \leq t)$ pour $j \in U \setminus s$ à l'aide de l'information auxiliaire

Dans le cas non censuré :

- Chambers et Dunstan (1986) pour le cadre paramétrique
- Dorfman et Hall (1993) pour le cadre non paramétrique

Plan

- Estimateur de la fdr sur une population
- Estimateur de la fdr sur petits domaines
- Estimation bootstrap du biais et de la variance de l'erreur de prédiction
- Simulations *model-based*
- Perspectives

Estimation de la fdr sur une population

- Premier terme de la fdr :

$$\frac{1}{N}\sum_{j\in s}\mathbb{1}(t_j \leq t) = \frac{n}{N}\left(\frac{1}{n}\sum_{j\in s}\mathbb{1}(t_j \leq t)\right)$$

  Le terme entre parenthèses est la fdr sur l'échantillon s
  $\hookrightarrow$ estimation par Kaplan-Meier

- Second terme : prédictions des $\mathbb{1}(t_j \leq t)$ pour $j \in U \setminus s$

  Modèle non paramétrique de superpopulation ξ :

$$t_j = m(x_j) + \varepsilon_j,\ j = 1, \ldots, N$$

  avec ε_j variables i.i.d. de fdr G et $m(x_j)$ médiane conditionnelle de T sachant $X = x_j$

$$\mathbb{E}_\xi\left(\mathbb{1}(t_j \leq t)\right) = P(t_j \leq t) = G(t - m(x_j))$$

$\hookrightarrow$ estimer $G(t - m(x_j))$

1ère étape : estimation de $m(x_j)$

• Estimateur de Kaplan-Meier généralisé de $F(t|x)$ (Dabrowska, 1992)

$$\widehat{F}_{\text{GKM}}(t \mid x) = \begin{cases} 1 - \prod_{j=1}^{n} \left\{ 1 - \dfrac{B_j(x)}{\sum_{r=1}^{n} \mathbb{1}_{(y_r \geq y_j)} B_r(x)} \right\}^{\mathbb{1}(y_j \leq t, \delta_j = 1)} & \text{si} \quad t < y_{(n)} \\ 1 & \text{sinon} \end{cases}$$

$B_j(x) = \dfrac{K\left(\dfrac{x - x_j}{h_X}\right)}{\sum_{j=1}^{n} K\left(\dfrac{x - x_j}{h_X}\right)}$ poids de type Nadaraya-Watson, K noyau, h_X fenêtre

- Lissage en t de $\widehat{F}_{\text{GKM}}(t \mid x)$ (Leconte *et al.*, 2002)

$$\widehat{F}_{\text{SGKM}}(t \mid x) = \sum_{j=1}^{d} \left(\widehat{F}_{\text{GKM}}\left(y^{\dagger}_{(j)} \mid x\right) - \widehat{F}_{\text{GKM}}\left(y^{\dagger}_{(j-1)} \mid x\right)\right) H\left(\frac{t - y^{\dagger}_{(j)}}{h_T}\right)$$

$y^{\dagger}_j$ observations non censurées, $y^{\dagger}_{(d)} = y_{(n)}$, H noyau intégré, h_T fenêtre

$$\hookrightarrow \hat{m}(x_j) = \widehat{F}^{-1}_{\text{SGKM}}(0.5 \mid x_j)$$

2ème étape : estimation de G (fdr des erreurs)

Les n résidus $\hat{\epsilon}_j = y_j - \hat{m}(x_j)$ sont censurés à droite (comme les y_j)

$\hookrightarrow$ Estimation de G par Kaplan-Meier sur les résidus $\rightarrow \widehat{G}_{\mathrm{KM}}$

- Estimation de F

$$\widehat{F}_{\mathrm{M}}(t) = \frac{1}{N}\left(n\widehat{F}_{\mathrm{KM}}(t) + \sum_{j \in U \setminus s} \widehat{G}_{KM}(t - \hat{m}(x_j)) \right)$$

Estimation de la fdr sur petits domaines

Population U partitionnée en m domaines U_i de taille N_i
Soit s_i un échantillon de taille n_i du domaine U_i et $s = \cup_{i=1}^{m} s_i$

Si échantillon du domaine de petite taille
$\hookrightarrow$ nécessité d'"emprunter de la force aux voisins" en plus de l'utilisation d'une information auxiliaire pour améliorer la précision de l'estimation de la fdr sur U_i

Dans le cas non censuré

- Modèles mixtes : voir Rao (2003), Small Area Estimation
- Quantiles conditionnels
  - Cadre paramétrique : Chambers et Tzavidis (2006)
  - Cadre non paramétrique : Casanova (2012)

Les étapes de l'estimation

- Premier terme de la fdr : par Kaplan-Meier sur s_i
- Second terme : prédictions des $\mathbb{1}(t_{ij} \leq t)$ pour $j \in U_i \setminus s_i$

Modèle non paramétrique de superpopulation sur U :

$$t_{ij} = m(q_i, x_{ij}) + \varepsilon_{ij}, \;\; i = 1, \ldots, m, \;\; j = 1, \ldots, N_i$$

avec q_i coefficient dans $(0, 1)$ caractérisant la position du domaine U_i, ε_{ij} variables i.i.d. (à i fixé) de fdr G_i et $m(q_i, x_{ij})$ quantile conditionnel d'ordre q_i de T sachant $X = x_{ij}$

$\hookrightarrow$ estimer $G_i(t - m(q_i, x_{ij}))$

1. Estimation de q_i
   - Estimation des ordres des individus de s par $\hat{F}_{\text{SGKM}}(y_{ij} \mid x_{ij})$ calculé à l'aide de s
   - Ordre q_i du domaine U_i estimé par la moyenne des ordres des individus de s_i non censurés $\rightarrow \hat{q}_i$
2. Estimation des $m(\hat{q}_i, x_{ij})$ par inversion de $\hat{F}_{\text{SGKM}}$ pour $j \in U_i \setminus s_i$
3. Estimation de G_i par Kaplan-Meier sur les résidus $\hat{\epsilon}_{ij}$ de s_i $\rightarrow \widehat{G}^i_{\text{KM}}$

- Estimation de F^i

$$\widehat{F}^i_{\text{Q}}(t) = \frac{1}{N_i}\left(n_i \widehat{F}^i_{\text{KM}}(t) + \sum_{j \in U_i \setminus s_i} \widehat{G}^i_{\text{KM}}(t - \hat{m}(\hat{q}_i, x_{ij})) \right)$$

Les étapes du bootstrap

D'après Lombardia *et al.* (2004)
Echantillon $(y_j, \delta_j, x_j) \rightarrow \widehat{F}_{\mathrm{M}}$

- Génération de B populations P^* $(B = 50)$
  $t_k^* = \hat{m}(x_k) + \epsilon_k^*$ où les ϵ_k^* sont générés à partir de $\widehat{G}_\lambda$ (version lissée de $\widehat{G}$)
  c_k^* générés à partir de l'estimation par KM inverse de la fdr de la variable délai de censure
  $\hookrightarrow (y_k^*, \delta_k^*)$
- Tirage de R échantillons dans chaque population P^* $(R = 100)$

Estimation bootstrap du biais et de la variance de l'erreur de prédiction et de l'IC de $F(t)$

- $E(\widehat{F}(t) - F(t)|\mathcal{P})$ estimé par $E_*(E(\hat{F}^*(t) - F^*(t)|\mathcal{P}^*))$

  $\hookrightarrow$ approché par $\frac{1}{B}\frac{1}{R}\sum_{b=1}^{B}\sum_{r=1}^{R}[\widehat{F}^{*br}(t) - F^{*b}(t)]$

- $Var(\widehat{F}(t) - F(t)|\mathcal{P})$ estimé par $E_*(Var(\hat{F}^*(t) - F^*(t)|\mathcal{P}^*))$

  $\hookrightarrow$ approché par $\frac{1}{B}\frac{1}{R}\sum_{b=1}^{B}\sum_{r=1}^{R}[\widehat{F}^{*br}(t) - \widehat{F}^{*b}(t)]^2$

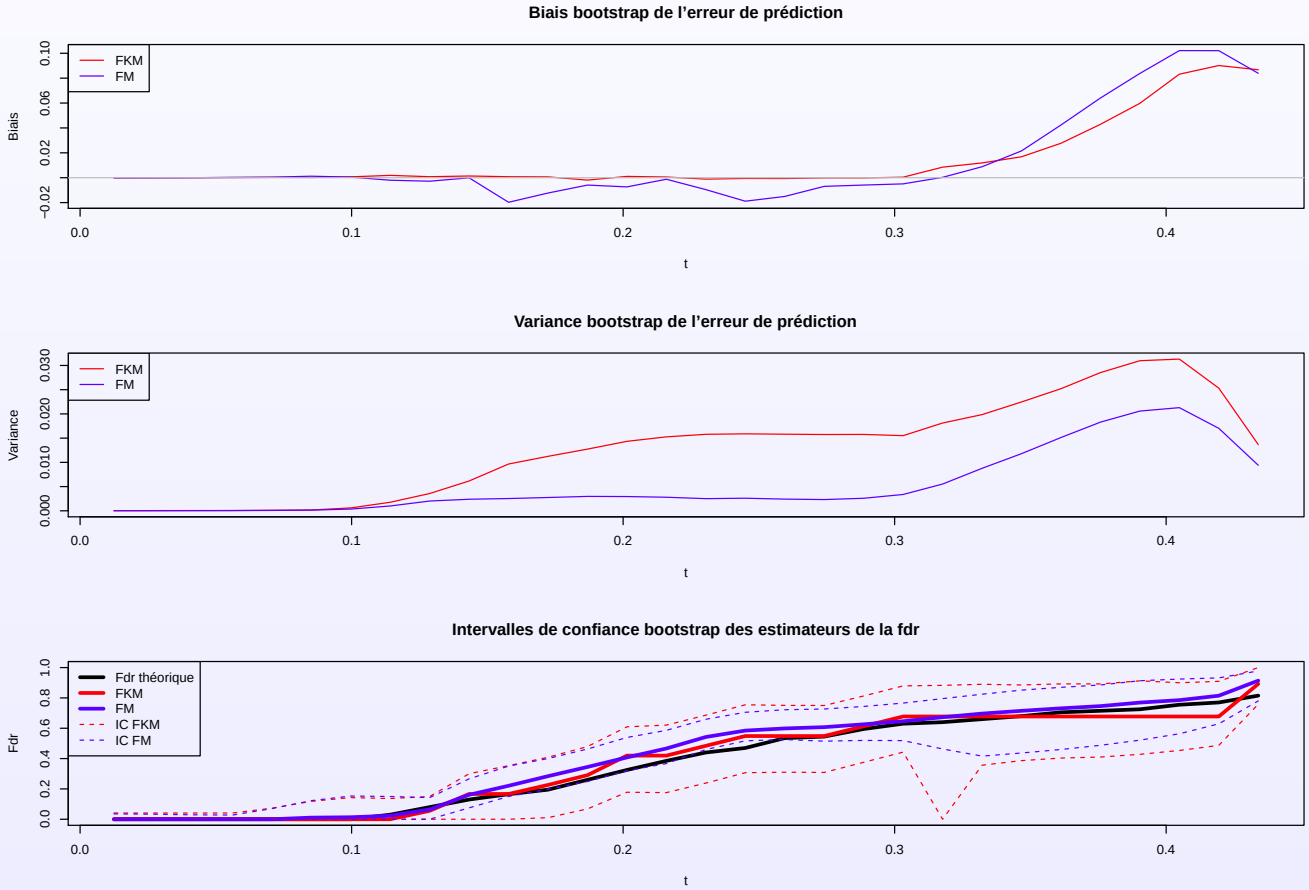
- IC de $F(t)$ au niveau de confiance $1-\alpha$:

  $$[\hat{F}(t) - q^*_{1-\frac{\alpha}{2}}, \hat{F}(t) + q^*_{\frac{\alpha}{2}}]$$

  avec q^* quantiles de l'estimation bootstrap de

  $$H(u) = P(\hat{F}(t) - F(t) \leq u \mid P)$$

Résultats pour $N = 200$ ($n = 20$), 25 % de censure



Simulations *model-based*

Description

- Pour chaque itération, génération d'une population partitionnée en 2 domaines suivant le modèle log-linéaire de régression :

$$log(t_{ij}) = \mu_i + 0.2 * x_{ij} + 0.1 * u_{ij}$$

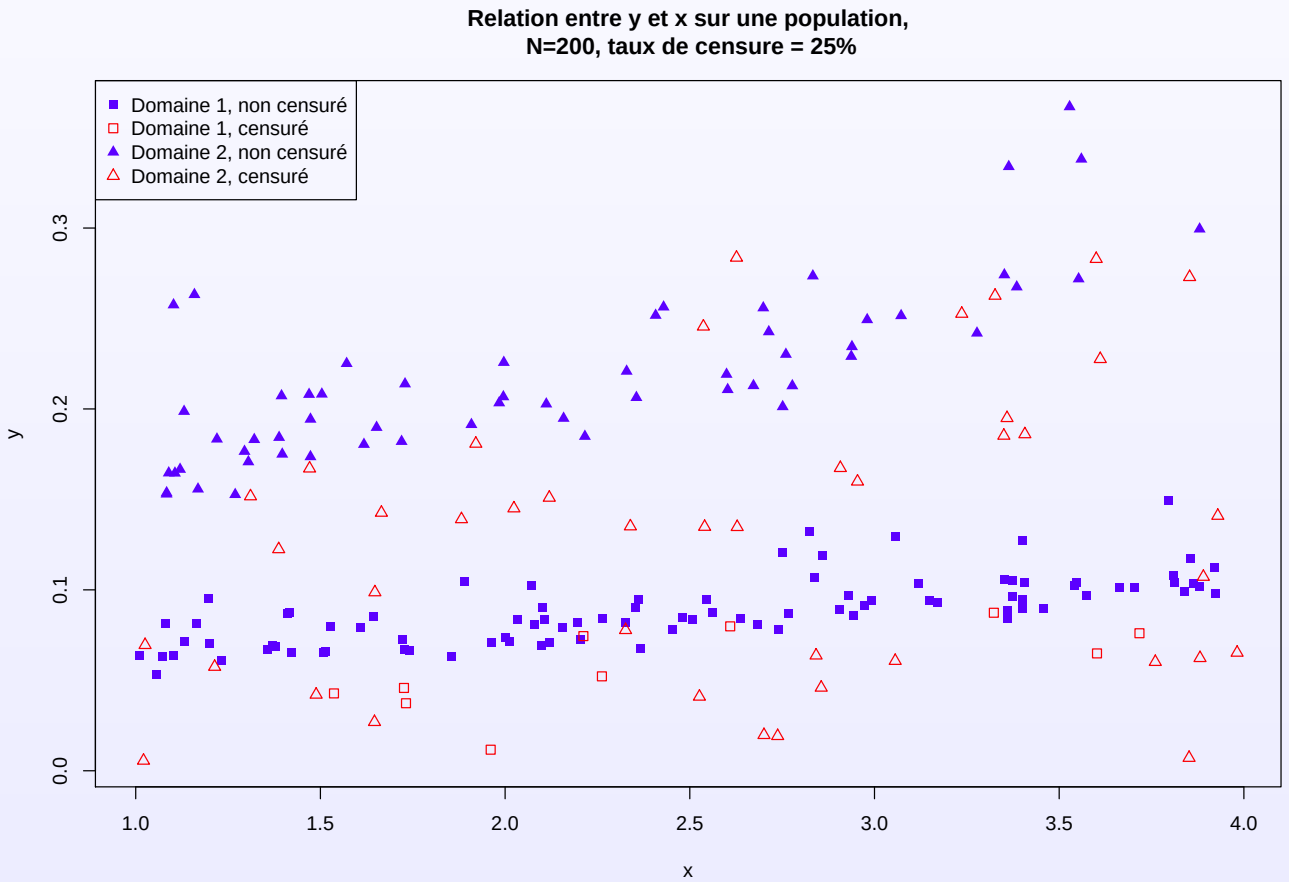
- $\mu_1 = -3, \mu_2 = -2$
- $x \sim U(1, 4)$
- u suit une distribution de valeur extrême $\rightarrow$ $t \sim$ Weibull
- $c \sim U(0, dp)$, dp choisi de façon à obtenir 10%, 25% ou 50% de censure
- $y_{ij} = min(t_{ij}, c_{ij})$ et $\delta_{ij} = \mathbb{1}(t_{ij} < c_{ij})$

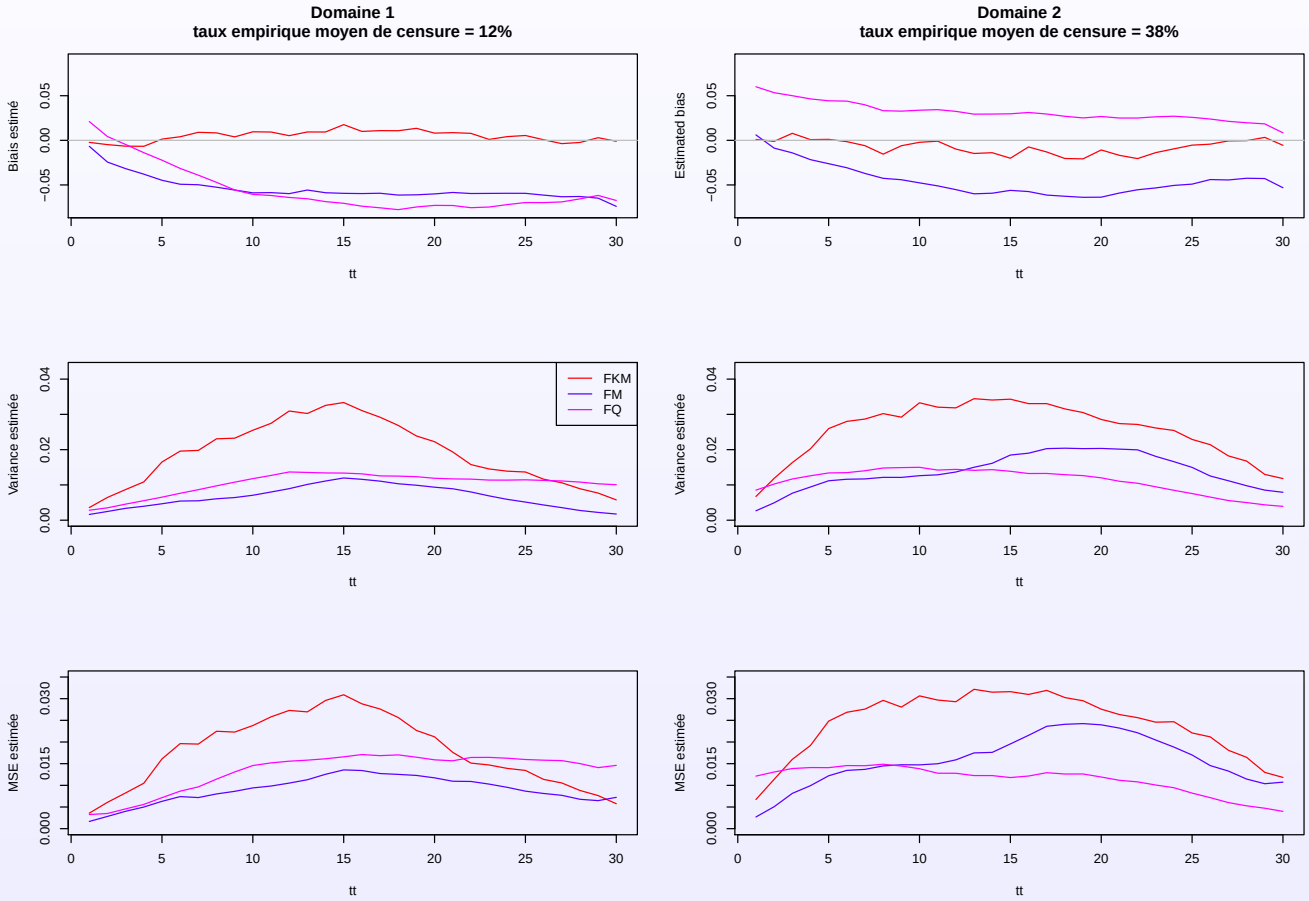
- 2 domaines de tailles égales : $N/2$ ($N = 100, 200, 400$)
- Dans chaque domaine, tirage de l'échantillon au $1/10$
- $S = 200$ itérations
- Choix des paramètres de lissage de $\widehat{F}^i_{\mathrm{M}}$
  Pour chaque itération, pour chaque domaine, le couple de fenêtres (h_X, h_T) est choisi de façon à minimiser le critère

$$\mathrm{ASE}(\widehat{F}^i_{\mathrm{M}}) = \frac{1}{K}\sum_{k=1}^{K}\left(\widehat{F}^i_{\mathrm{M}}(tt_k) - F^i(tt_k)\right)^2$$

  (évalué sur une grille de $K = 30$ points)
- Choix des paramètres de lissage de $\widehat{F}^i_{\mathrm{Q}}$: (h_X, h_T) tel que $\mathrm{ASE}(\hat{F}^1_{\mathrm{Q}}) + \mathrm{ASE}(\hat{F}^2_{\mathrm{Q}})$ minimum

Résultats


Relation entre y et x sur une population,
N=200, taux de censure = 25%
Domaine 1, non censuré
Domaine 1, censuré
Domaine 2, non censuré
Domaine 2, censuré
y
0.0
0.1
0.2
0.3
x
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Domaine 1
taux empirique moyen de censure = 12%
Domaine 2
taux empirique moyen de censure = 38%
Biais estimé
Estimated bias
Variance estimée
MSE estimée
tt
FKM
FM
FQ

Résultats

Tableau des MASE (moyenne des ASE sur les 200 itérations) dans chaque domaine ($\times 10^{-5}$)

		Domaine 1				Domaine 2			
τ	N	$\tau_1(\%)$	KM	M	Q	$\tau_2(\%)$	KM	M	Q
10 %	100	5.3	3398	2235	2833	14.8	3519	3174	2188
	200	6.0	1548	860	1299	14.2	1826	1194	1097
	400	5.5	744	532	488	14.8	825	452	453
25 %	100	14.7	3809	2509	3410	37.3	5963	5092	2664
	200	12.2	1792	892	1334	37.7	2431	1580	1135
	400	13.2	841	646	586	36.2	1207	585	626
50 %	100	25.6	4587	3215	4277	71.2	9107	6645	4256
	200	26.6	2103	954	1462	72.2	5095	3206	2224
	400	28.8	1007	593	800	73.1	2865	1222	1092

Perspectives

- Simulations avec d'autres types de relation entre t et x et plus de domaines
- Utilisation du bootstrap pour le choix de fenêtres optimales sur les données réelles (au préalable, validation croisée pour le choix des fenêtres pilotes)
- Trouver un bon exemple d'application