Estimation de la fonction de distribution conditionnelle à partir de données censurées par intervalle, cas I

Sandra Plancade

MAP5 Université Paris 5

5 mai 2010

Plan de l'exposé

I) Introduction
II) Définition de l'estimateur
III) Résultats

I) Introduction

- Données censurées par intervalle, cas I (ou current status data)

$$Y \in \mathbb{R}^+ = \text{ temps de survie}$$
$$X \in \mathbb{R} = \text{ covariable}$$
$$T \in \mathbb{R}^+ = \text{ temps d'observation}$$
$$\delta = \mathbb{1}_{Y \leq T}$$

avec $Y \perp T | X$.

I) Introduction

- Données censurées par intervalle, cas I (ou current status data)

$$Y \in \mathbb{R}^+ = \text{ temps de survie}$$
$$X \in \mathbb{R} = \text{ covariable}$$
$$T \in \mathbb{R}^+ = \text{ temps d'observation}$$
$$\delta = \mathbb{1}_{Y \leq T}$$

  avec $Y \perp T | X$.
- On observe un échantillon i.i.d.

$$(X_i, T_i, \delta_i = \mathbb{1}_{Y_i \leq T_i})_{i=1,\dots,n} \quad \text{avec} \quad Y_i \perp T_i | X_i, \ \forall i = 1, \dots, n.$$

- But : estimer la fonction de distribution conditionnelle

$$F(x, y) = P[Y \leq y | X = x]$$

sur un compact $A_1 \times A_2 \subset \mathbb{R} \times \mathbb{R}^+$ par sélection de modèle.

- But : estimer la fonction de distribution conditionnelle

$$F(x, y) = P[Y \leq y | X = x]$$

sur un compact $A_1 \times A_2 \subset \mathbb{R} \times \mathbb{R}^+$ par sélection de modèle.

- On considère deux risques pour l'estimateur $\widehat{F}$ de F,

$$\mathbb{E}[\|\widehat{F} - F\|^2_{f_{(\boldsymbol{X},\boldsymbol{T})}}] \quad \text{où} \quad \|g\|^2_{f_{(\boldsymbol{X},\boldsymbol{T})}} = \int g^2(x) f_{(X,T)}(x) dx,$$

$$\mathbb{E}[\|\widehat{F} - F\|^2_n | (X_i, T_i)_{i=1,\ldots,n}] \quad \text{où} \quad \|g\|^2_n = \frac{1}{n} \sum_{i=1}^{n} g^2(X_i, T_i).$$

- But : estimer la fonction de distribution conditionnelle

$$F(x,y) = P[Y \leq y | X = x]$$

sur un compact $A_1 \times A_2 \subset \mathbb{R} \times \mathbb{R}^+$ par sélection de modèle.

- On considère deux risques pour l'estimateur $\widehat{F}$ de F,

$$\mathbb{E}[\|\widehat{F} - F\|^2_{f_{(\boldsymbol{X},\boldsymbol{T})}}] \quad \text{où} \quad \|g\|^2_{f_{(\boldsymbol{X},\boldsymbol{T})}} = \int g^2(x) f_{(X,T)}(x) dx,$$

$$\mathbb{E}[\|\widehat{F} - F\|^2_n | (X_i, T_i)_{i=1,\ldots,n}] \quad \text{où} \quad \|g\|^2_n = \frac{1}{n} \sum_{i=1}^{n} g^2(X_i, T_i).$$

- Notre estimateur converge à la vitesse $n^{-\overline{\beta}/(\overline{\beta}+1)}$ sur les espaces de Besov anisotropes $\mathcal{B}^{(\beta_1,\beta_2)}_{2,\infty}$.

.

II) Définition de l'estimateur

- **Construction d'un contraste empirique**

$$\begin{array}{rccc} \gamma_n : & L^2(A_1 \times A_2) & \to & \mathbb{R} \\ & g & \to & \gamma_n(g) \end{array}$$

déterminé par les observations.

$$\widehat{F} = \arg\min_{g \in ?} \gamma_n(g).$$

- **Collection de modèles** (= sev de $L^2(A_1 \times A_2)$).
- **Sélection de modèles.**

Construction du contraste

- $$\begin{aligned}\mathbb{E}[\delta|(X,T)=(x,t)] &= \mathbb{E}[\mathbb{1}_{Y\leq T}|(X,T)=(x,t)]\\ &= \mathbb{E}[\mathbb{1}_{Y\leq t}|X=x]\\ &= P[Y\leq t|X=x]\\ &= F(x,t)\end{aligned}$$

Construction du contraste

- $$\begin{aligned}\mathbb{E}[\delta|(X,T)=(x,t)] &= \mathbb{E}[\mathbb{1}_{Y\leq T}|(X,T)=(x,t)]\\ &= \mathbb{E}[\mathbb{1}_{Y\leq t}|X=x]\\ &= P[Y\leq t|X=x]\\ &= F(x,t)\end{aligned}$$

$$\delta = F(X,T)+\epsilon \quad \text{avec} \quad \mathbb{E}[\epsilon|X,T]=0.$$

Construction du contraste

- $$\begin{aligned}\mathbb{E}[\delta|(X,T)=(x,t)] &= \mathbb{E}[\mathbb{1}_{Y\leq T}|(X,T)=(x,t)]\\ &= \mathbb{E}[\mathbb{1}_{Y\leq t}|X=x]\\ &= P[Y\leq t|X=x]\\ &= F(x,t)\end{aligned}$$

$$\delta = F(X,T)+\epsilon \quad \text{avec} \quad \mathbb{E}[\epsilon|X,T]=0.$$

- Contraste des moindres carrés

$$\gamma_n(g) = \frac{1}{n}\sum_{i=1}^{n}\Big(\underbrace{\delta_i - g(X_i,T_i)}_{\mathbb{E}[\quad |X,T]=F(X,T)-g(X,T)}\Big)^2$$

Estimateurs non adaptatifs

- La collection de modèles : les modèles sur $A_1 \times A_2$ sont construits comme produits tensoriels de modèles sur A_1 et A_2.

Estimateurs non adaptatifs

- La collection de modèles : les modèles sur $A_1 \times A_2$ sont construits comme produits tensoriels de modèles sur A_1 et A_2.
- On considère deux collections de modèles $\mathcal{M}_n^{(j)} = \{S_m^{(j)}, m \in J_n^{(j)}\}$ où

$$S_m^{(j)} = Vect\left(\phi_{k,m}^{(j)}, k = 1, \ldots, D_m^{(j)}\right), \quad m \in J_n^{(j)}$$

est un sev de $L^2(A_j)$, pour $j = 1, 2$.

Estimateurs non adaptatifs

- La collection de modèles : les modèles sur $A_1 \times A_2$ sont construits comme produits tensoriels de modèles sur A_1 et A_2.
- On considère deux collections de modèles $\mathcal{M}_n^{(j)} = \{S_m^{(j)}, m \in J_n^{(j)}\}$ où

$$S_m^{(j)} = Vect\left(\phi_{k,m}^{(j)}, k = 1, \ldots, D_m^{(j)}\right), \quad m \in J_n^{(j)}$$

  est un sev de $L^2(A_j)$, pour $j = 1, 2$.
- Pour tout $m = (m_1, m_2) \in J_n = J_n^{(1)} \times J_n^{(2)}$, soit

$$S_m = Vect\left((x,y) \to \phi_{k_1,m_1}^{(1)}(x)\phi_{k_2,m_2}^{(2)}(y), k_1 = 1, \ldots, D_{m_1}^{(1)}, k_2 = 1, \ldots, D_{m_2}^{(2)}\right)$$

$$\text{et} \quad \mathcal{M}_n = \{S_m, m = (m_1, m_2) \in J_n\}.$$

Estimateurs non adaptatifs

- La collection de modèles : les modèles sur $A_1 \times A_2$ sont construits comme produits tensoriels de modèles sur A_1 et A_2.
- On considère deux collections de modèles $\mathcal{M}_n^{(j)} = \{S_m^{(j)}, m \in J_n^{(j)}\}$ où

$$S_m^{(j)} = Vect\left(\phi_{k,m}^{(j)}, k = 1, \ldots, D_m^{(j)}\right), \quad m \in J_n^{(j)}$$

est un sev de $L^2(A_j)$, pour $j = 1, 2$.

- Pour tout $m = (m_1, m_2) \in J_n = J_n^{(1)} \times J_n^{(2)}$, soit

$$S_m = Vect\left((x,y) \to \phi_{k_1,m_1}^{(1)}(x)\phi_{k_2,m_2}^{(2)}(y), k_1 = 1, \ldots, D_{m_1}^{(1)}, k_2 = 1, \ldots, D_{m_2}^{(2)}\right)$$

$$\text{et} \quad \mathcal{M}_n = \{S_m, m = (m_1, m_2) \in J_n\}.$$

- Pour tout $m \in J_n$,

$$\widehat{F}_m = \arg \min_{g \in S_m} \gamma_n(g)$$

Sélection de modèle

- Pour tout $m \in J_n$, presque surement,

$$\mathbb{E}\left[\|\widehat{F}_m - F\|_n^2 | (X_i, T_i)_{i=1,\dots,n}\right] \leq \underbrace{\inf_{g \in S_m} \|F - g\|_n^2}_{\text{estimé par } \gamma_n(\widehat{F}_m)} + \frac{D_m}{n}$$

Sélection de modèle

- Pour tout $m \in J_n$, presque surement,

$$\mathbb{E}\left[\|\widehat{F}_m - F\|_n^2 | (X_i, T_i)_{i=1,\dots,n}\right] \leq \underbrace{\inf_{g \in S_m} \|F - g\|_n^2}_{\text{estimé par } \gamma_n(\widehat{F}_m)} + \frac{D_m}{n}$$

- $\widehat{m} = \arg\min_{m \in J_n} \left\{ \gamma_n(\widehat{F}_m) + A\frac{D_m}{n} \right\}$

Récapitulatif

- Nous avons contruit un contraste empirique γ_n.
- Nous avons défini une collection $\{S_m, m \in J_n\}$, et

$$\widehat{F}_m = \arg \min_{g \in S_m} \gamma_n(g).$$

- Nous avons sélectionné un modèle $\widehat{m}$.

Récapitulatif

- Nous avons contruit un contraste empirique γ_n.
- Nous avons défini une collection $\{S_m, m \in J_n\}$, et

$$\widehat{F}_m = \arg \min_{g \in S_m} \gamma_n(g).$$

- Nous avons sélectionné un modèle $\widehat{m}$.

  $\Rightarrow$ Notre estimateur est $\widehat{F}_{\widehat{m}}$.

Inégalités oracle

Théorème

$$\mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F\|_n^2 | (X_i, T_i)_{i=1,\dots,n}\right] \leq C \inf_{m \in J_n} \underbrace{\left\{ \inf_{g \in S_m} \|F - g\|_n^2 + A\frac{D_m}{n} \right\}} + \frac{C'}{n}$$

$$\simeq \mathbb{E}\left[\|\widehat{F}_m - F\|_n^2 | (X_i, T_i)_{i=1,\dots,n}\right]$$

Inégalités oracle

Théorème

$$\mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F\|_n^2 | (X_i, T_i)_{i=1,\dots,n}\right] \leq C \inf_{m \in J_n} \underbrace{\left\{ \inf_{g \in S_m} \|F - g\|_n^2 + A\frac{D_m}{n} \right\}}_{\simeq \mathbb{E}\left[\|\widehat{F}_m - F\|_n^2 | (X_i, T_i)_{i=1,\dots,n}\right]} + \frac{C'}{n}$$

Théorème

Supposons que pour tout $m \in J_n$, $D_m \leq \sqrt{n}/\log n$,

$$\mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F\|_{f_{(\boldsymbol{X},\boldsymbol{T})}}^2\right] \leq C \inf_{m \in J_n} \underbrace{\left\{ \inf_{g \in S_m} \|F - g\|_{f_{(\boldsymbol{X},\boldsymbol{T})}}^2 + A\frac{D_m}{n} \right\}}_{\simeq \mathbb{E}\left[\|\widehat{F}_m - F\|_{f_{(\boldsymbol{X},\boldsymbol{T})}}^2\right]} + \frac{C'}{n}$$

Vitesse de convergence sur les espaces de Besov

- Soit $\mathcal{B}_{2,\infty}^{(\beta_1,\beta_2)}(L)$ la boule de rayon L d'un espace de Besov anisotrope.

Vitesse de convergence sur les espaces de Besov

- Soit $\mathcal{B}_{2,\infty}^{(\beta_1,\beta_2)}(L)$ la boule de rayon L d'un espace de Besov anisotrope.

Proposition

Si $F \in \mathcal{B}_{2,\infty}^{(\beta_1,\beta_2)}(L)$, *avec* $\overline{\beta} \geq 1$,

$$\mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F\|^2_{f(\boldsymbol{x},\boldsymbol{T})}\right] \leq Cn^{-2\overline{\beta}/(2\overline{\beta}+2)} \quad \text{où} \quad \frac{2}{\overline{\beta}} = \frac{1}{\beta_1} + \frac{1}{\beta_2}.$$

De plus, l'estimateur $\widehat{F}_{\widehat{m}}$ *atteint la vitesse minimax.*

Bibliographie

- P. Groeneboom and J.A. Wellner, *Information bounds and nonparametric maximum likelihood estimation*, DMV seminar (1992)
- S. Van de Geer, *Hellinger consistency of certain nonparametric maximum likelihood estimators*, Ann. Statist. (1993)
- M. Hudgens et al, *Nonparametric estimation of the joint distribution of a survival time subject to interval censoring and a continuous times variable*, Harmon. Anal., (2007)
- M. J. Van der Laan and A. Van der Vaart, *Estimating a survival distribution with current status data and high dimensionnal covariates*, Int. J. Biostat. (2006)
- E. Brunel and F. Comte, *Cumulative distribution function under interval censoring case 1*, Electro. J. Stat., (2009)
- S. Ma and M. R. Kosorok, *Adaptive penalised M-estimation with current status data*, Ann. Instit. Statist. Math. (2006)

Conclusion

- Etude minimax sur des espaces de régularité
- Généralisable à $X \in \mathbb{R}^2$, $\mathbb{R}^3$,...
- Problème pour généraliser à des covariables de grande dimension
- Autre méthode d'estimation de fonction de régression