Estimation de la fonction de répartition conditionnelle à partir de données censurées par intervalle, cas 1, par sélection de modèles

Sandra Plancade

Estimation de la fonction de distribution conditionnelle en présence de donnees censurées par intervalle, cas I, par sélection de modèles

Sandra Plancade

Laboratoire MAP5 (UMR CNRS 8145)
Université Paris Descartes
45 rue des Saints Pères
75270 Paris Cedex 6

Résumé Considérons une variable aléatoire positive Y appelée temps de survie, dépendant d'une covariable X. Dans le cadre de la censure par intervalle, cas I (ou "current status data"), Y n'est pas directement observée. La seule information dont on dispose est la donnée du triplet (X, U, δ) ou U est un temps de mesure indépendant de Y conditionnellement à X, et $\delta = 1\{Y \leq U\}$. Le but de cet exposé est de construire un estimateur de la fonction de distribution conditionnelle de Y sachant X à partir d'un échantillon de (X, U, δ), par sélection de modèles. Après avoir construit une collection de modèles pour des fonctions de deux variables par produits tensoriels de fonctions d'une variable, on calcule dans chacun de ces modèles l'estimateur des moindres carrés. On obtient ainsi une collection d'estimateurs. Enfin, une procédure de sélection de modèles pénalisée fournit un estimateur adaptatif.

Abstract Let Y be a positive random variable of interest, depending on a covariable X. The interval censoring (also called "current status data") arises when Y is unobserved and the only information we have is the triplet (X, U, δ) where U is a positive random variable independent of Y given X, and $\delta = 1\{Y \leq U\}$. In this presentation we propose an adaptive estimator of the distribution function of Y given X, constructed by a model selection procedure. A collection of models of $L^2(\mathbb{R}^2)$ is built as tense products of functions of $L^2(\mathbb{R})$, and a least square estimator is calculated in each model. We get a collection of estimators from which one is choosen by a model selection procedure.

Mots-clés Donnes de survie et donnes censures. Modles semi et non paramtriques.

1 Introduction

Dans certaines études de données de survie, le temps de survie Y n'est pas directement observé. La seule donnée que l'on observe est si Y est antérieur ou postérieur à un temps de mesure U indépendant de T, éventuellement conditionnellement à une variable explicative X. Par exemple, si l'on s'intéresse à la date de contamination Y d'un patient par un virus, un test effectué à la date U permet de déterminer si le patient a été contaminé ou non

avant la date U. Ce type de censure est appelé censure par intervalle, cas I ou current status data.

Plus précisément, soient $(X_i, Y_i)_{i=1,\dots,n}$ des couples indépendants identiquement distribués (i.i.d.), où les (X_i) sont des variables aléatoires i.i.d. de densité f_X et les (Y_i) sont des variables aléatoires positives appelées temps de survie tels que pour tout i, Y_i dépend de la covariable X_i. Soit $F(x, y)$ la fonction de répartition conditionnelle de Y_i sachant X_i,

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

où $P[E_1|E_2]$ désigne la probabilité conditionnelle de E_1 sachant E_2. Soit $(U_i)_{i=1,\dots,n}$ un échantillon de variables aléatoires i.i.d. telles que pour tout i, U_i et Y_i sont indépendantes conditionnellement à X_i. L'échantillon observé est

$$(X_i, U_i, \delta_i)_{i=1,\dots,n} \tag{1}$$

où $U_i = 1_{\{Y_i \leq U_i\}}$. Le but de cet exposé est de construire un estimateur de $F(x, y)$ à partir de l'échantillon (1).

Les données censurées par intervalle ont été largement étudiées ces vingt dernières années. La majorité des résultats concernant l'estimation non paramétrique de fonctions de survie est basée sur l'estimateur du maximum de vraissemblance (EMV). Ainsi, Groeneboom et Wellner (1992) montrent que l'EMV converge ponctuellement à la vitesse $n^{-1/3}$, et Van der Geer (1993) établit un résultat similaire pour le risque L^2. Plus récemment, des méthodes d'estimation dérivées de l'EMV permettent de mieux prendre en compte la régularité de la fonction de survie. Ainsi, Hudgens et al (2007) construisent trois estimateurs à partir de l'EMV et comparent leurs performances sur des données simulées et réelles. Van der Laan et Van der Vaart (2006) appliquent des méthodes de régularisation à l'EMV pour estimer la fonctions de survie en présence de covariables de grandes dimensions. Par ailleurs, Birgé (1999) propose un estimateur en histogramme, plus simplement implémentable que les estimateurs de type EMV, et qui atteint le taux de convergence optimal pour le risque L^1. Mais les procédures proposées dans ces articles ne sont pas adaptatives. Très peu de résultats existent à ce sujet, et ils n'incluent pas de covariables. Ma et Kosorov (2006) calculent l'EMV et l'estimateur des moindres carrés sur les classes de Sobolev, puis effectuent une sélection du paramètre de régularité de la classe, alors que Brunel et Comte (2009) calculent l'estimateur des moindres carrés dans des bases d'approximation classiques puis réalisent une sélection de modèles.

La procédure présentée ici est inspirée de Brunel et Comte (2009) : l'estimateur est construit par minimisation d'un contraste des moindres carrés, puis par sélection de modèle. Néanmoins, la démonstration est différente et permet de s'affranchir de certaines hypothèses sur la collection de modèles, ainsi que d'améliorer les constantes intervenant dans le résulat principal. De plus, notre modèle inclut des covariables, ce qui n'est pas le cas des deux articles précédemment cités.

2 Hypothèses et définition de l'estimateur de F

2.1 Notations

La fonction $F(x,y)$ est estimée sur un compact $A = A_1 \times A_2$ où A_1 est un intervalle compact de $\mathbb{R}$, et $A_2 = [0, a_2]$.

Soit $f_{(X,U)}$ la densité du couple (X_i, U_i) pour tout $i \in \{1, \dots, n\}$.

Pour tout $s, t \in L^2(A)$, on note

$$\langle s,t\rangle_{f_{(X,U)}} = \int_{x\in A_1}\int_{u\in A_2} s(x,u)t(x,u)f_{(X,U)}(x,u)dxdu$$

le produit scalaire associé à $f_{(X,U)}$ et $\|t\|_{f_{(X,U)}} = \sqrt{\langle t,t\rangle_{f_{(X,U)}}}$. Leurs équivalents empiriques sont

$$\langle s,t\rangle_n = \frac{1}{n}\sum_{i=1}^{n} t(X_i,U_i)s(X_i,U_i)$$

et $\|t\|_n = \sqrt{\langle t,t\rangle_n}$.

2.2 Collection de modèles

Pour estimer F fonction de deux variables, on définit une collection de sous espaces vectoriels (s.e.v) de $L^2(A)$ de dimension finie appelés "modèles", et construits comme produits tensoriels de modèles de $L^2(A_1)$ et $L^2(A_2)$. Plus précisément, soit $\mathcal{M}_n^1 = \{S_m^1, m \in I_n^1\}$ une collection de s.e.v. de $L^2(A_1)$ telle que pour tout $m_1 \in I_n^1$ $dim(S_m^1) = D_{m_1} < +\infty$. Soit $(\phi_k^{m_1})_{k=1,\dots,D_{m_1}}$ une base orthonormale de $S_{m_1}^1$. Soit $\mathcal{M}_n^2 = \{S_m^2, m \in I_n^2\}$ une collection de s.e.v. de $L^2(A_2)$ telle que pour tout $m_2 \in I_n^2$ $dim(S_m^2) = D_{m_2} < +\infty$. Soit $(\psi_k^{m_2})_{k=1,\dots,D_{m_2}}$ est une base orthonormale de $S_{m_2}^2$. Pour tout $m = (m_1, m_2) \in I_n := I_n^1 \times I_n^2$, soit S_m le modèle suivant,

$$S_m = \{t : A \to \mathbb{R}, t(x,y) = \sum_{(k,l)\in J_m} a_{k,l}\phi_k^{m_1}(x)\psi_l^{m_2}(y)\},$$

avec

$$J_m = ((1,1),\dots,(1,D_{m_2}),(2,1),\dots,(2,D_{m_2}),\dots,(D_{m_1},1),\dots,(D_{m_1},D_{m_2}))\,.$$

On considère alors la collection $\mathcal{M}_n = \{S_m, m = (m_1, m_2) \in I_n\}$.

Supposons que les collections $\mathcal{M}_n^1$ et $\mathcal{M}_n^2$ vérifient l'hypothèse suivante.

(**H**) : Pour $i = 1$ et 2, il existe un modèle $S_n^i \in \mathcal{M}_n^i$, tel que pour tout $m_i \in I_n^1$, $S_{m_i}^i \subset S_n^i$. Soit $N_n^{(i)} = dim(S_n^i)$. De plus, pour tout $a > 0$, il existe une constante A telle que

$$\sum_{m_i \in I_n^i} \exp(-a\sqrt{D_{m_i}}) \leq A, \quad \forall n \in \mathbb{N}^*.$$

Enfin, il existe un polynôme P tel que $Card(\mathcal{M}_n) \leq P(n)$.

La Proposition suivante établit des propriétés similaires pour la collection $\mathcal{M}_n$.

Proposition 2.1 *Supposons que* ***(H)*** *soit vérifiée.*

1. *Pour tout $m \in I_n$, la famille $\{\phi_k^{m_1}\phi_l^{m_2}, k = 1, \ldots, D_{m_1},\ l = 1, \ldots, D_{m_2}\}$ est une base orthonormale de S_m, et S_m a pour dimension $D_m = D_{m_1}D_{m_2}$.*

2. *Pour tout $a > 0$, il existe une constante A telle que*

$$\sum_{m \in I_n} \exp(-a\sqrt{D_m}) \leq A, \quad \forall n \in \mathbb{N}^*.$$

3. *Pour tout $m \in I_n$, $S_m \subset S_n$ où*

$$S_n = \{t : A \to \mathbb{R}, t(x,y) = \sum_{k=1,\ldots,N_n^1,\ l=1,\ldots,N_n^2} a_{k,l}\phi_k^n(x)\psi_l^n(y)\}.$$

2.3 Construction du contraste

Notons $\mathbb{E}[B_1|B_2]$ l'espérance de B_1 sachant B_2. On remarque que pour tout $(x,u) \in \mathbb{R}^2$,

$$\mathbb{E}[\delta_1|(X_1,U_1) = (x,u)] = \mathbb{E}[1_{\{Y_1 \leq u\}}|(X_1,U_1) = (x,u)].$$

Or Y_1 et U_1 sont indépendants conditionnellement à X_1 donc

$$\mathbb{E}[\delta_1|(X_1,U_1) = (x,u)] = \mathbb{E}[1_{\{Y_1 \leq u\}}|X_1 = x] = P[Y_1 \leq u|X_1 = x] = F(x,u).$$

Ainsi, soit $\gamma_n(t)$ le contraste suivant,

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n} \left(t(X_i,U_i) - \delta_i\right)^2,$$

$\mathbb{E}[\gamma_n(F)] = 0$, donc F minimise $\mathbb{E}[\gamma_n(t)]$ sur $L^2(A)$.

2.4 Estimateurs non adaptatifs

Le contraste γ_n construit dans la section précédente permet de définir un estimateur $\widehat{F}_m$ pour chaque modèle $m \in \mathcal{M}_n$:

$$\widehat{F}_m = \arg\min_{t \in S_m} \gamma_n(t). \tag{2}$$

Notons $\widehat{F}_m(x,u) = \sum_{(k,l)\in J_m} \widehat{a}_{k,l}\phi_k^{m_1}(x)\psi_l^{m_2}(u)$, alors (2) équivaut à

$$G_m A_m = V_m$$

où $A_m = [a_{k,l}]_{(k,l)\in J_m}$ est un vecteur colonne,

$$G_m = \left[\frac{1}{n}\sum_{i=1}^{n} \phi_k^{m_1}(X_i)\psi_l^{m_2}(U_i)\phi_{k'}^{m_1}(X_i)\psi_{l'}^{m_2}(U_i)\right]_{((k,l),(k',l'))\in J_m\times J_m}$$

est une matrice carrée à D_m lignes, et

$$V_m = \left[\frac{1}{n}\sum_{i=1}^{n} \phi_k^{m_1}(X_i)\psi_k^{m_2}(U_i)\delta_i\right]_{(k,l)\in J_m}$$

est un vecteur colonne. La matrice G_m est la matrice de Gram de la base $(\phi_k^{m_1}\psi_l^{m_2})$ pour le produit scalaire $\langle .,. \rangle_n$. On remarque que $\widehat{F}_m$ est définie de manière unique ssi la matrice G_m est inversible. Examinons l'existence et l'unicité de $\widehat{F}_m$.

Soit $\widehat{S}_m$ le sev de $\mathbb{R}^n$ défini par

$$\widehat{S}_m = \{(t(X_1,U_1),\ldots,t(X_n,U_n)), t \in S_m\},$$

et $\widehat{Z}_m = \arg\min_{Z\in\widehat{S}_n} \frac{1}{n}\sum_{i=1}^{n}(Z_i - \delta_i)$.[2] Le vecteur $\widehat{Z}_m$ est la projection orthogonale du vecteur $(\delta_1,\ldots,\delta_n)$ sur $\widehat{S}_m$ pour le produit scalaire canonique de $\mathbb{R}^n$, il est donc défini de manière unique. De plus, par définition de $\widehat{S}_m$, il existe au moins une fonction $G \in S_m$ telle que $\widehat{Z}_m = (G(X_1,U_1),\ldots,G(X_n,U_n))$, alors G minimise $\gamma_n(t)$ sur S_m, ce qui prouve l'existence de $\widehat{F}_m$. De plus, si deux telles fonctions G existent, elles sont égales sur $(X_i,U_i)_{i=1,\ldots,n}$ donc la quantité $\mathbb{E}[\|G-F\|_n^2]$ est identique. La définition de $\arg\min_{t\in S_m}\gamma_n(t)$ est donc pertinente.

2.5 Procédure de selection de modèles

On dispose d'une collection d'estimateurs $\{\widehat{F}_m, m \in I_n\}$ parmi lesquel on choisit l'estimateur suivant.

$$\widehat{m} = \arg\min_{m\in I_n}\left[\gamma_n(\widehat{F}_m) + pen(m)\right],$$

où $pen(m) = \theta \frac{D_m}{n}$ avec $\theta > 1$. En quelques mots, le principe de la sélection de modèle pénalisée, développé par Birgé et Massart s'appuie sur le raisonnement suivant. Le meilleur modèle dans la collection $\{\widehat{F}_m, m \in I_n\}$ est celui pour lequel l'erreur $\|\widehat{F}_m - F\|_n^2$ est la plus faible. Or $\|\widehat{F}_m - F\|_n^2$ se décompose en un terme de biais, de l'ordre de $\gamma_n(\widehat{F}_m)$ et un terme de variance de l'ordre de $pen(m)$. Le modèle choisi est donc celui qui minimise la somme $\gamma_n(\widehat{F}_m) + pen(m)$. (cf Massart (2007) pour plus de détails.)

L'estimateur de F considéré est $\widehat{F}_{\widehat{m}}$.

3 Résultat

L'estimateur $\widehat{F}_{\widehat{m}}$ vérifie le résultat suivant.

Theorem 3.1 *Supposons que l'hypothèse* (**H**) *soit vérifiée et que* $N_n \leq n$, *alors*

$$\mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F\|_n^2\right] \leq C \inf_{m \in I_n} \{ \inf_{t \in S_m} \|F - t\|_{f_{(X,U)}}^2 + pen(m)\} + \frac{C'}{n}$$

où C *et* C' *sont des constantes numériques.*

Bibliographie

[1] Birgé, L. (1999) Interval censoring: a nonasymptotic point of view *Math. Methods Statist.* 8, no. 3, 285–298.

[2] Brunel E. et Comte, F. (2009) Cumulative distribution function estimation under censoring case 1 *Electron. J. Stat.*, 3, 1–24.

[3] Groeneboom, P. et Wellner, J.A. (1992) *Information bounds and nonparametric maximum likelihood estimation* DMV seminar, 19, Birkhauser Verlag, Basel.

[4] Hudgens, M.G., Maathuis, M. H. et Gilbert, P.B. (2007) Nonparametric estimation of the joint distribution of a survival time subject to interval censoring and a continuous mark variable *Biometrics*, 63, no. 2, 372–380.

[5] Massart, P. (2007) *Concentration inequalities and model selection* Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, Lecture Notes in Mathematics, 1896. Springer, Berlin.

[6] Van de Geer, S. (1993) Hellinger-consistency of certain nonparametric loglikelihood estimators *Ann. Stat.*, 21, 14–44.

[7] Van der Laan, M.J. and Van der Vaart, A. (2006) Estimating a survival distribution with current status data and high-dimensional covariates *Int. J. Biostat.* 2, 9–42.