Prédiction de la fonction de survie par sélection de modéle

Ion Grama, Jean-François Petiot

PREDICTION DE LA FONCTION DE SURVIE PAR SELECTION DE MODELE

Ion GRAMA & Jean-François PETIOT

Laboratoire LMAM, Université de Bretagne Sud, Centre Yves Coppens, Campus de Tohannic BP 573, 56017 VANNES, FRANCE

Résumé

Nous proposons un estimateur semi-paramétrique d'une fonction de survie $S(t) = P(T \geq t)$ à partir des observations censurées $Z_i = \min\{T_i, C_i\}$ et des indicateurs respectifs $\Delta_i = 1(X_i \leq C_i)$, où les temps de censure indépendants C_i sont indépendants des durées de vie indépendantes T_i. Notre but est d'obtenir une prédiction des probabilités de survie $S(t)$ au-delà des durées observées, c'est-à-dire pour $t > \max_i\{Z_i\}$. L'idée principale de l'approche proposée est de choisir de façon automatique un seuil u à partir duquel les prévisions pour les durées de vie sont encore fiables. En dessous de ce seuil $S(t)$ est estimée par une méthode complètement non paramétrique, comme celle de Kaplan Meier. Au dessus de ce seuil un modèle paramétrique est choisi, nous utiliserons ici la loi exponentielle. Le choix du seuil u sera assuré par une suite de tests d'ajustement. La méthode est appliquée à des données de ré-hospitalisation, la durée de vie étant ici le délai écoulé entre une sortie d'un hôpital et une ré-admission pour la même cause médicale.

Mots clés: Données de survie et données censurées, modèles semi et non paramétriques.

Abstract

We propose a semiparametric estimator of the survival function $S(t) = P(T \geq t)$ from the censored observation $Z_i = \min\{T_i, C_i\}$ and the corresponding indicators $\Delta_i = 1(X_i \leq C_i)$, where the independent cencored times C_i are independent of the independent survival times T_i. Our goal is to obtain predictions of the survival probabilities $S(t)$ outside the range of the observed times, that is for $t > \max_i\{Z_i\}$. The main idea of the proposed approch is to to choose adaptively a threshold u starting from which the predictions of the survival times are still reliable. Below the threshold $S(t)$ is estimated by a completely non-parametric method, such as the Kaplan-Meyer one. Above the the threshold a parametric model is proposed - here we use an exponential law. The choice of the threshold u is performed by a sequence of goodness-of-fit tests. This method is applied to rehospitalization data, where survival times are time lengths to readmission for the same medical reason.

Keywords: Survival analysis and censored data, semi and non parametric models.

1 Introduction

Soit $T_1, ..., T_n$, des durées de vie i.i.d. de fonction de survie $S(t) = P(T \geq t)$ supposée continue et strictement décroissante sur l'intervalle $[0, \infty)$. Ici T est la notation générique de T_i. Soit $C_1, ..., C_n$ une suite des variables aléatoires i.i.d. de même support $[0, \infty)$. On observe les réalisations $z_1, ..., z_n$ des variables censurées $Z_i = \min\{X_i, C_i\}$ et les réalisatons respectives $\delta_1, ..., \delta_n$ des indicateurs $\Delta_i = 1(X_i \leq C_i)$.

Le but est d'obtenir des prédictions non triviales des probabilités de survie $S(t) = P(T \geq t)$ pour les valeurs t au-delà des durées observées z_i, i.e. pour $t > \max\{z_1, ..., z_n\}$. Pour cela on choisira le meilleur modèle dans une famille dont les propriétes des prédiction sont bien adaptées au problème considéré et qui, d'autre part, est suffisament flexible pour ajuster convenablement les données. Nous proposons une famille des modèles qui pré-suppose qu'au-delà d'un seuil u les données peuvent être bien approchées par un modèle paramétrique. Pour les valeurs en dessous du seuil u on adopte un modèle non-paramétrique. Le seuil u sera choisi par une procedure de sélection du modèle.

Pour décrire formellement la famille des modèles on note d'abord que pour $t \geq u$ la probabilité conditionnelle de survie $S_u(t) = P(T \geq t | T \geq u)$ satisfait

$$S_u(t) = \exp\left(-\int_u^t h(x)\, dx\right),$$

où $h(x)$, $x > 0$ est la fonction de risque instantané. On supposera que sur l'intervalle $[u, \infty)$ la fonction $h(\cdot)$ peut être convenablement ajustée par la famille $h_\mu(x - u)$, $\mu \in \Theta$. Pour les valeurs $t < u$ la survie est une fonction arbitraire $q(t)$, $t \in [0, u[$. Cela signifie qu'on suppose que la probabilité de survie de T a une structure semi-paramétrique comme suit:

$$S(t) = P(T \geq t) = \begin{cases} q(t), & t \in [0, u[, \\ q(u) \exp\left(-\int_u^t h_\mu(x - u)\, dx\right), & t \geq u. \end{cases} \tag{1}$$

Notons par $\mathcal{S}_u$ l'ensemble des fonctions de survie satisfaisant (1), pour un $u \geq 0$ donné. Le modèle ajusté sera choisi dans la famille des modèles $\mathcal{S}_u$, $u \geq 0$.

D'abord nous construisons un estimateur de la fonction de survie S appartenant à la famille $\mathcal{S}_u$, pour une valeur de temps u donnée. Pour cela on construit un estimateur du paramètre μ et de la fonction inconnue $q(t)$, $t \in [0, u[$, qui donnera une famille d'estimateurs $\widehat{S}_u$, $u \geq 0$.

Ensuite nous proposons une procédure du choix du modèle dans la famille $\mathcal{S}_u$, $u \geq 0$. Ceci sera effectué au moyen d'une procédure séquentielle de tests d'ajustement qui consistent à tester le modèle sous une hypothèse nulle contre un modèle plus large, comme décrit aux paragraphes 2 et 3.

2 Estimateur du maximum de vraisemblance sous l'hypothèse nulle

Nous construisons l'estimateur du maximum de vraisemblance sous l'hypothèse nulle

$$h_\mu(x) = \frac{1}{\mu}, \quad x \geq u,$$

pour $\mu > 0$. Cette hypothèse implique que $S_u(t)$ suit une loi exponentielle avec le paramètre inconnu $\lambda = 1/\mu$. Sans perte de généralité nous pouvons supposer que $z_1 < z_2 < ... < z_n$. Dans la suite u est choisi dans $\{z_1, ..., z_n\}$, i.e. $u = z_k$, où k est un entier avec $1 \leq k \leq n$. On note $\gamma_i = 1(z_i \leq u) = 1(i \leq k)$.

Nous construisons une famille d'estimateurs avec l'approche non paramétrique de Kiefer et Folfowitz (voir Bickel et all. (1992)). En laissant de côté le formalisme mathématique, nous supposons que les survies conditionnelles sont des paramètres inconnus $q_i \in [0, 1]$:

$$S_{z_{i-1}}(z_i) = P(T \geq z_i | T \geq z_{i-1}) = q_i,$$

pour $i = 1, ..., k$. Cela implique, pour $z_i \leq u = z_k$:

$$S(z_i) = S_0(z_1) S_{z_1}(z_2) \cdot ... \cdot S_{z_{i-1}}(z_i) = \prod_{z_j \leq z_i} q_j,$$

avec $q_0 = 1$, et pour $t > u = z_k$:

$$S(t) = \left(\prod_{z_j \leq u = z_k} q_j \right) P(T \geq t | T \geq u) = \left(\prod_{z_j \leq u = z_k} q_j \right) e^{-\frac{1}{\mu}(t-u)}.$$

D'où la fonction de survie :

$$S(t) = \begin{cases} \prod_{z_i \leq t} q_i, & t \leq u, \\ e^{-\mu^{-1}(t-u)} \prod_{i: z_i \leq u} q_i, & t > u. \end{cases}$$

La log vraisemblance partielle semi paramétrique est :

$$\begin{aligned} L_u(q, \mu) &= \sum_{i=1}^{k} (n-i) \ln q_i + \sum_{i=1}^{k} (1 - \delta_i) \ln q_i + \sum_{i=1}^{k} \delta_i \ln(1 - q_i) \\ &\quad - \ln \mu \sum_{i=k+1}^{n} \delta_i + \mu^{-1} \sum_{i=k+1}^{n} (z_i - u). \end{aligned}$$

La maximisation de $L_u(q, \mu)$ par rapport à q_j et μ donne les estimateurs MV :

$$\widehat{q}_j = \frac{n - j + 1 - \delta_j}{n - j + 1}, \; j \leq k,$$

et

$$\widehat{\mu}_u \equiv \frac{\sum_{i=k+1}^{n}(z_i - u)}{\sum_{i=k+1}^{n}\delta_i}.$$

La fonction de survie $S(t)$ est estimée par :

$$\widehat{S}_n(t) = \begin{cases} \prod_{i:z_i \le t} \frac{n-i+1-\delta_i}{n-i+1}, & t \le u, \\ e^{-\frac{1}{\widehat{\mu}_u}(t-u)} \prod_{i:z_i \le u} \frac{n-i+1-\delta_i}{n-i+1}, & t > u. \end{cases}$$

En notant $\widehat{n}_u = \sum_{z_i > u} \delta_i$. et après des calculs élémentaires, on obtient le logarithme du rapport de vraisemblance :

$$L_u(q, \widehat{\mu}_u) - L_u(q, \mu) = \widehat{n}_u \mathcal{K}(\widehat{\mu}_u, \mu),$$

où $\mathcal{K}(x, y) = G\left(\frac{x}{y} - 1\right)$ pour $x, y > 0$ et $G(x) = x - \ln(x+1)$ pour $x > -1$. Il est facile de vérifier que $\mathcal{K}(\mu_1, \mu_2)$ est la divergence de Kullback-Leibler entre deux lois exponentielles de paramètres μ_1 et μ_2.

3 Estimateur du maximum de vraisemblance sous l'hypothèse alternative

Nous construisons ici une famille d'estimateurs sous l'hypothèse alternative avec un risque instantané comportant un point de rupture v:

$$h_{\mu_1,\mu_2}(x) = \begin{cases} \frac{1}{\mu_1}, & u < x \le v, \\ \frac{1}{\mu_2}, & x > v. \end{cases}$$

où $\mu_1 > 0, \mu_2 > 0$ et $\alpha > 0$. Cette hypothèse implique que la queue de la distribution de T suit une loi exponentielle avec un point de rupture. On obtient la fonction de survie suivante :

$$S(t) = \begin{cases} \prod_{z_i \le t} q_i, & t \le u = z_k, \\ \prod_{z_i \le t} q_i \exp\left(-\frac{t-u}{\mu_1}\right), & u < t \le v, \\ \prod_{z_i \le t} q_i \exp\left(-\frac{t-u}{\mu_1}\right) \exp\left(-\frac{t-v}{\mu_2}\right), & t > v. \end{cases}$$

La log vraisemblance peut être facilement calculée Soit $\widehat{n}_v = \sum_{z_i > v} \delta_i$, $\widehat{n}_{u,v} = \sum_{u < z_i \le v} \delta_i$ et $\overline{n}_v = card\{Z_i > v\}$, $\overline{n}_{u,v} = card\{u \le Z_i < v\}$. Alors la log vraisemblance a l'expression suivante :

$$\begin{aligned} L_{u,v}(q, \mu_1, \mu_2) &= \sum_{z_i \le u} (n-i) \ln q_i + \sum_{z_i \le u} (1 - \delta_i) \ln q_i + \sum_{z_i \le u} \delta_i \ln(1 - q_i) \\ &\quad -\frac{1}{\mu_1} \sum_{u < z_i \le v} (z_i - u) - \frac{1}{\mu_2} \sum_{z_i > v} (z_i - v) \\ &\quad -\overline{n}_v \frac{1}{\mu_1}(v - u) - \widehat{n}_{u,v} \ln \mu_1 - \widehat{n}_v \ln \mu_2. \end{aligned}$$

En maximisant $L_{u,v}(q, \mu_1, \mu_2)$ par rapport à q_j, μ_1 et μ_2 on obtient les estimateurs :

$$\widehat{q}_j = \frac{n - j + 1 - \delta_j}{n - j + 1},\ j = 1, ..., k$$

et

$$\widehat{\mu}_{u,v} = \frac{\sum_{u<z_i\leq v}(z_i - u) + \sum_{v<z_i}(v - u)}{\sum_{u<z_i\leq v}\delta_i} = \frac{\sum_{u<z_i}(\min\{z_i, v\} - u)}{\sum_{u<z_i\leq v}\delta_i},$$

$$\widehat{\mu}_v = \frac{\sum_{z_i>v}(z_i - v)}{\sum_{z_i>v}\delta_i}.$$

Après quelques calculs, le log du rapport des vraisemblances s'écrit :

$$L_{u,v}(q, \widehat{\mu}_{u,v}, \widehat{\mu}_v) - L_{u,v}(q, \mu) = \widehat{n}_v \mathcal{K}(\widehat{\mu}_v, \mu) + \widehat{n}_{u,v}\mathcal{K}(\widehat{\mu}_{u,v}, \mu), \tag{2}$$

où $\mathcal{K}(\widehat{\mu}_v, \mu)$ est la divergence de Kullback-Leibler définie précédemment.

4 Procédure de sélection du seuil u

Nous proposons ici la procédure de sélection pour u. On suppose que le seuil u et le point de rupture v sont tous les deux choisis dans $\{z_1, ..., z_n\}$, i.e. $u = z_k$ et $v = z_l$, où k et l sont des entiers tels que $k_0 \leq l \leq k \leq n$. Cela signifie que c'est l'entier k qu'il faut choisir. Soit δ', δ'' deux constantes telles que $0 < \delta', \delta'' < \frac{1}{3}$. On considère la constante $\tau > 0$, qui joue le rôle de valeur critique dans la procédure de test ci dessous. Les valeurs k_0, δ', δ'' et τ sont les paramètres qui doivent être calibrés empiriquement. Soit

$$\begin{aligned} T_{u,v} &= \widehat{n}_v \mathcal{K}(\widehat{\mu}_v, \widehat{\mu}_u) + \widehat{n}_{u,v}\mathcal{K}(\widehat{\mu}_{u,v}, \widehat{\mu}_u), \\ \widetilde{T}_{u,v} &= \widehat{n}_v \mathcal{K}(\widehat{\mu}_v, \widehat{\mu}_u) \end{aligned} \tag{3}$$

On effectue ensuite des tests consécutifs de l'hypothèse nulle contre l'alternative, introduites précédemment.

La procédure de choix de k est la suivante:

Etape 1. Poser $k = k_0$.

Etape 2. Calculer la statistique de test

$$T_{z_k} = \max_{\delta' k \leq l \leq (1-\delta'')k} T_{z_k, z_l}.$$

Etape 3. Si $T_{z_k} \leq \tau$ et $k \leq n$ incrémenter k de 1 et retourner à l'étape 2. Si $T_{z_k} > \tau$ et $k \leq n$ on définit la valeur adaptative

$$\widehat{k} = \arg \max_{\delta' k \leq l \leq (1-\delta'')k} \widetilde{T}_{z_k, z_l}$$

et on termine la procédure. Si $k > n$ on définit $\widehat{k} = n$ et on sort également de la procédure.

Le seuil adaptatif est défini par $\widehat{u} = z_{\widehat{k}}$.

Remarque. Dans le cas de durées de vie suivant la loi exponentielle, la statistique de test dépend peu du paramètre de cette loi. Ceci suggère de calculer la valeur critique τ du test par des simulations de Monte Carlo selon un modèle exponentiel de durée de vie. Le choix des autres paramètres sera discuté.

5 Résultats asymptotiques et simulations

Nous montrons que si la fonction de survie $S(u)$ est bien approximée par une loi exponentielle avec un paramètre μ_n pour les valeurs de temps u au dessus du seuil u_n, alors sous certaines conditions

$$\widehat{\mu}_{u_n} - \mu_n \to 0, \quad n \to \infty, \tag{4}$$

en probabilité. Nous donnons également une borne pour la vitesse de convergence dans (4). Bien sûr la valeur de u_n n'est pas connue dans les situations pratiques. On montre qu'avec le choix adaptatif $\widehat{u}_n = z_{\widehat{k}}$ défini auparavant, l'estimateur adaptatif $\widehat{\mu}_{\widehat{u}_n}$ estime le paramètre inconnu μ_n aussi bien que si u_n était connu.

Nous présenterons des simulations qui confirment nos résultats théoriques et une application pour des durées de ré-hospitalisation.

Bibliographie

[1] Bickel, P.J., Klaassen, C.A.J., Rytov, Y. and Wellner, J.A. (1992). Efficient and Adaptive Estimation in Semiparametric Models. John Hopkins Univ. Press.

[2] Grama, I. and Spokoiny, V. (2008) Statistics of Extremes by Oracle Estimation. *Ann. Statist.* Vol. 36. No. 4. 1619-1648.