Estimation du risque instantané à partir de données censurées à droite

Sandra Plancade

MAP5 Université Paris 5

16 août 2009

Plan de l'exposé

1) Données censurées à droite : cadre général.
2) Estimation du risque instantané : un bref aperçu des méthodes existantes.
3) Estimation du risque instantané par minimisation d'un contraste.
4) Selection de modèle et résultat.

1) Données censurées à droite

- Variable d'intéret (temps de survie) : X
  Temps de censure : C

  On observe $\begin{cases} Y = \min(X, C) \\ \delta = 1_{X \leq C} \end{cases}$ avec X et C indépendants.

  Exemple :

  - ⋆ X= temps de survie d'un patient après la pose d'un pacemaker.
  - ⋆ C= temps auquel l'individu abandonne le programme d'étude, meurt d'une autre cause...

1) Données censurées à droite

- Variable d'intéret (temps de survie) : X
  Temps de censure : C

  On observe $\begin{cases} Y = \min(X, C) \\ \delta = 1_{X \leq C} \end{cases}$ avec X et C indépendants.

  Exemple :
  - ⋆ X= temps de survie d'un patient après la pose d'un pacemaker.
  - ⋆ C= temps auquel l'individu abandonne le programme d'étude, meurt d'une autre cause...
- Echantillon i.i.d. $\{Y_i = \min(X_i, C_i), \delta_i = 1_{X_i \leq C_i}\}_{i=1,\ldots,n}$ avec $X_i \perp C_i$.

1) Données censurées à droite

- Variable d'intéret (temps de survie) : X
  Temps de censure : C

  On observe $\begin{cases} Y = \min(X, C) \\ \delta = 1_{X \leq C} \end{cases}$ avec X et C indépendants.

  Exemple :
  - ⋆ X= temps de survie d'un patient après la pose d'un pacemaker.
  - ⋆ C= temps auquel l'individu abandonne le programme d'étude, meurt d'une autre cause...
- Echantillon i.i.d. $\{Y_i = \min(X_i, C_i), \delta_i = 1_{X_i \leq C_i}\}_{i=1,\dots,n}$ avec $X_i \perp C_i$.
- Idée "naive" : éliminer de l'étude les variables censurées $\longrightarrow$ résultats biaisés

1) Données censurées à droite

- Variable d'intéret (temps de survie) : X
  Temps de censure : C

  On observe $\begin{cases} Y = \min(X, C) \\ \delta = 1_{X \leq C} \end{cases}$ avec X et C indépendants.

  Exemple :
  - ⋆ X= temps de survie d'un patient après la pose d'un pacemaker.
  - ⋆ C= temps auquel l'individu abandonne le programme d'étude, meurt d'une autre cause...
- Echantillon i.i.d. $\{Y_i = \min(X_i, C_i), \delta_i = 1_{X_i \leq C_i}\}_{i=1,\ldots,n}$ avec $X_i \perp C_i$.
- Idée "naive" : éliminer de l'étude les variables censurées $\longrightarrow$ résultats biaisés

  $\Rightarrow$ méthodes d'estimation qui intègrent les données censurées et non censurées.

Estimation du risque instantané : un bref aperçu des méthodes existantes

- **Notations**

  f_X : densité de X_i

  $\overline{F}_X(x) = P(X \geq x)$

  $\overline{F}_C(x) = P(C \geq x)$

  $\overline{F}_Y(x) = P(Y \geq x) = P(\{X \geq x\} \cap \{C \geq x\}) = \overline{F}_X(x) \times \overline{F}_C(x)$

Estimation du risque instantané : un bref aperçu des méthodes existantes

- **Notations**

  f_X : densité de X_i

  $\overline{F}_X(x) = P(X \geq x)$

  $\overline{F}_C(x) = P(C \geq x)$

  $\overline{F}_Y(x) = P(Y \geq x) = P(\{X \geq x\} \cap \{C \geq x\}) = \overline{F}_X(x) \times \overline{F}_C(x)$

- Définition du risque instantané $h(x)$:

$$\begin{aligned} h(x) &= \frac{P(X \in [x, x+dx] | X \geq x)}{dx} \\ &= \frac{P(\{X \in [x, x+dx]\} \cap \{X \geq x\})}{P(X \geq x)dx}) = \frac{f_X(x)}{\overline{F}_X(x)} \end{aligned}$$

Estimation du risque instantané : un bref aperçu des méthodes existantes

- **Notations**
  f_X : densité de X_i
  $\overline{F}_X(x) = P(X \geq x)$
  $\overline{F}_C(x) = P(C \geq x)$
  $\overline{F}_Y(x) = P(Y \geq x) = P(\{X \geq x\} \cap \{C \geq x\}) = \overline{F}_X(x) \times \overline{F}_C(x)$
- Définition du risque instantané $h(x)$:

$$\begin{aligned} h(x) &= \frac{P(X \in [x, x+dx] | X \geq x)}{dx} \\ &= \frac{P(\{X \in [x, x+dx]\} \cap \{X \geq x\})}{P(X \geq x)dx}) = \frac{f_X(x)}{\overline{F}_X(x)} \end{aligned}$$

- Par ailleurs :

$$h(x) = -\frac{1}{\overline{F}_X(x)} \frac{d\overline{F}_X}{dx}(x) = -\frac{d(\log(\overline{F}_X))}{dx}(x)$$

Deux méthodes générales d'estimation de h

- A partir de l'expression $h(x) = \frac{f_X(x)}{\overline{F}_X(x)}$: quotient de deux estimateurs.

Deux méthodes générales d'estimation de h

- A partir de l'expression $h(x) = \frac{f_X(x)}{\overline{F}_X(x)}$: quotient de deux estimateurs.
- A partir de l'expression $h(x) = -[\log(\overline{F}_X)]'(x)$: estimation de $\log(\overline{F}_X(x))$, puis dérivation discrète.

Deux méthodes générales d'estimation de h

- A partir de l'expression $h(x) = \frac{f_X(x)}{\overline{F}_X(x)}$: quotient de deux estimateurs.
- A partir de l'expression $h(x) = -[\log(\overline{F}_X)]'(x)$: estimation de $\log(\overline{F}_X(x))$, puis dérivation discrète.

$\Rightarrow$ Procédures en deux étapes (puis sélection de modèle, sélection de fenêtre...)

Estimation de h par minimisation d'un contraste

- **Hypothèses :** on estime h sur un intervalle compact A tel que :

$$\begin{cases} \overline{F}_Y(x) \geq \overline{F}_0 > 0, \quad \forall x \in A \\ \|h\|_\infty := \sup_{x \in A} h(x) < \infty \end{cases}$$

Estimation de h par minimisation d'un contraste

- **Hypothèses :** on estime h sur un intervalle compact A tel que :

$$\begin{cases} \overline{F}_Y(x) \geq \overline{F}_0 > 0, \quad \forall x \in A \\ \|h\|_\infty := \sup_{x \in A} h(x) < \infty \end{cases}$$

- **Notations :** soient $t,\, t' \in L^2(A)$

$$\|t\|^2_{\overline{F}_Y} = \int_A t^2(x)\overline{F}_Y(x)dx, \quad \langle t, t' \rangle_{\overline{F}_Y} = \int_A t(x)t'(x)\overline{F}_Y(x)dx$$

$$\|t\|^2_n = \int_A t^2(x)(\frac{1}{n}\sum_{i=1}^n 1_{Y_i \geq x})dx, \quad \langle t, t' \rangle_n = \int_A t(x)t'(x)(\frac{1}{n}\sum_{i=1}^n 1_{Y_i \geq x})dx$$

- Estimation par minimisation d'un contraste

$$
\begin{aligned}
h &= \arg \min_{t \in L^2(A)} \|h - t\|^2_{\overline{F}_Y} = \arg \min_{t \in L^2(A)} [\|t\|^2_{\overline{F}_Y} - 2\langle t, h \rangle_{\overline{F}_Y} + \|h\|^2_{\overline{F}_Y}] \\
&= \arg \min_{t \in L^2(A)} [\underbrace{\|t\|^2_{\overline{F}_Y} - 2\langle t, h \rangle_{\overline{F}_Y}}_{\text{on estime cette quantité}}]
\end{aligned}
$$

- Estimation par minimisation d'un contraste

$$
\begin{aligned}
h &= \arg \min_{t \in L^2(A)} \|h - t\|^2_{F_Y} = \arg \min_{t \in L^2(A)} [\|t\|^2_{F_Y} - 2\langle t, h \rangle_{F_Y} + \|h\|^2_{F_Y}] \\
&= \arg \min_{t \in L^2(A)} [\underbrace{\|t\|^2_{F_Y} - 2\langle t, h \rangle_{F_Y}}_{\text{on estime cette quantité}}]
\end{aligned}
$$

- $\|t\|^2_{F_Y}$ est estimé par $\|t\|^2_n$.

- Estimation par minimisation d'un contraste

$$
\begin{aligned}
h &= \arg \min_{t \in L^2(A)} \|h - t\|^2_{\overline{F}_Y} = \arg \min_{t \in L^2(A)} [\|t\|^2_{\overline{F}_Y} - 2\langle t, h \rangle_{\overline{F}_Y} + \|h\|^2_{\overline{F}_Y}] \\
&= \arg \min_{t \in L^2(A)} [\underbrace{\|t\|^2_{\overline{F}_Y} - 2\langle t, h \rangle_{\overline{F}_Y}}_{\text{on estime cette quantité}}]
\end{aligned}
$$

- $\|t\|^2_{\overline{F}_Y}$ est estimé par $\|t\|^2_n$.
- $\langle t, h \rangle_{\overline{F}_Y}$ est estimé par $\frac{1}{n} \sum_{i=1}^n \delta_i t(Y_i)$.

- Estimation par minimisation d'un contraste

$$\begin{aligned} h &= \arg\min_{t\in L^2(A)} \|h-t\|^2_{\overline{F}_Y} = \arg\min_{t\in L^2(A)} [\|t\|^2_{\overline{F}_Y} - 2\langle t,h\rangle_{\overline{F}_Y} + \|h\|^2_{\overline{F}_Y}] \\ &= \arg\min_{t\in L^2(A)} [\underbrace{\|t\|^2_{\overline{F}_Y} - 2\langle t,h\rangle_{\overline{F}_Y}}_{\text{on estime cette quantité}}] \end{aligned}$$

- $\|t\|^2_{\overline{F}_Y}$ est estimé par $\|t\|^2_n$.
- $\langle t,h\rangle_{\overline{F}_Y}$ est estimé par $\frac{1}{n}\sum_{i=1}^n \delta_i t(Y_i)$.

En effet : $\mathbb{E}[\delta_i t(Y_i)|X_i] = \mathbb{E}[1_{X_i\le C_i} t(X_i)|X_i]$
$= t(X_i)P[X_i \le C_i|X_i] = t(X_i)\overline{F}_C(X_i)$

Donc :

$$\begin{aligned} \mathbb{E}[\delta_i t(Y_i)] &= \mathbb{E}[t(X_i)\overline{F}_C(X_i)] = \int_A t(x)\overline{F}_C(x) f_X(x)dx \\ &= \int_A t(x)\overline{F}_C(x)\overline{F}_X(x)\frac{f_X(x)}{\overline{F}_X(x)}dx = \int_A t(x)\overline{F}_Y(x)h(x)dx \\ &= \langle t,h\rangle_{\overline{F}_Y} \end{aligned}$$

- Soit $S_m = Vect(\phi_1, \ldots, \phi_{D_m}$ un s.e.v. de $L^2(A)$ de dimension finie D_m :

$$\widehat{h}_m = \arg\min_{t \in S_m} \gamma_n(t), \quad \text{où } \gamma_n(t) = \|t\|_n^2 - \frac{2}{n} \sum_{i=1}^{n} \delta_i t(Y_i)$$

- Soit $S_m = Vect(\phi_1, \ldots, \phi_{D_m}$ un s.e.v. de $L^2(A)$ de dimension finie D_m :

$$\widehat{h}_m = \arg\min_{t \in S_m} \gamma_n(t), \quad \text{où } \gamma_n(t) = \|t\|_n^2 - \frac{2}{n}\sum_{i=1}^{n} \delta_i t(Y_i)$$

- **Décomposition biais-variance** : soit $h_m = \arg\min_{t \in S_m} \|t - h\|_{F_Y}^2$, alors pour tout $x \in A$, $\mathbb{E}[\widehat{h}_m(x)] = h_m(x)$ et :

$$\begin{aligned} \mathbb{E}[\|\widehat{h}_m - h\|_{F_Y}^2] &= \|h - h_m\|_{F_Y}^2 + \mathbb{E}[\|\widehat{h}_m - h_m\|_{F_Y}^2] \\ &\leq \|h - h_m\|_{F_Y}^2 + \|h\|_\infty \frac{D_m}{n} \end{aligned}$$

Précisions sur le calcul de l'estimateur $\widehat{h}_m$

La fonction $\tilde{h}_m = \sum_{k=1}^{D_m} \tilde{a}_k \phi_k$ minimise le contraste γ_n sur S_m ssi $\tilde{A}_m = (\tilde{a}_1, \ldots, \tilde{a}_{D_m})^t$ vérifie :

$$\widehat{G}_m \tilde{A}_m = \widehat{V}_m \text{ où } \begin{cases} \widehat{G}_m = \left(\langle \phi_k, \phi_{k'} \rangle_n\right)_{k,k'=1,\ldots,D_m} \\ \widehat{V}_m = \left(\frac{1}{n} \sum_{i=1}^n \delta_i \phi_k(Y_i)\right)_{k=1,\ldots,D_m} \end{cases}$$

On définit un ensemble Δ_0 tel que $\widehat{G}_m$ est inversible sur Δ_0 et $P[\Delta_0^c]$ décroit exponentiellement en n. On considère alors l'estimateur $\widehat{h}_m = \sum_{k=1}^{D_m} \widehat{a}_k \phi_k$ l'estimateur suivant :

$$(\widehat{a}_1, \ldots, \widehat{a}_{D_m})^t = \begin{cases} \widehat{G}_m^{-1} \widehat{V}_m & \text{sur } \Delta_0 \\ 0 & \text{sur } \Delta_0^c \end{cases}$$

4) Sélection de modèle et résultat.

- Soit $\mathcal{M}_n = \{S_m, m = 1, \ldots, N_n\}$ une collection de s.e.v. de $L^2(A)$, qui vérifie :
  - $S_m \subset S_{N_n}$, pour tout $m \leq N_n$.
  - $dim(S_{N_n}) \leq \frac{n}{(\ln n)^2}$.

4) Sélection de modèle et résultat.

- Soit $\mathcal{M}_n = \{S_m, m = 1, \ldots, N_n\}$ une collection de s.e.v. de $L^2(A)$, qui vérifie :
  - $S_m \subset S_{N_n}$, pour tout $m \leq N_n$.
  - $dim(S_{N_n}) \leq \frac{n}{(\ln n)^2}$.
- Soit $\{\widehat{h}_m, m = 1, \ldots, N_n\}$ la collection d'estimateurs associée.

4) Sélection de modèle et résultat.

- Soit $\mathcal{M}_n = \{S_m, m = 1, \ldots, N_n\}$ une collection de s.e.v. de $L^2(A)$, qui vérifie :
  - $S_m \subset S_{N_n}$, pour tout $m \leq N_n$.
  - $dim(S_{N_n}) \leq \frac{n}{(\ln n)^2}$.
- Soit $\{\widehat{h}_m, m = 1, \ldots, N_n\}$ la collection d'estimateurs associée.

Theorem

Soit $A > 0$ une constante, et : $\widehat{m} = \arg\min[\gamma_n(\widehat{h}_m) + A\|h\|_\infty \frac{dim(S_m)}{n}]$, alors :

$$\mathbb{E}[\|\widehat{h}_{\widehat{m}} - h\|^2_{\overline{F}_Y}] \leq C \inf_{m \in \mathcal{M}_n} \{\|h - h_m\|^2_{\overline{F}_Y} + A\|h\|_\infty \frac{dim(S_m)}{n}\} + \frac{C'}{n}$$

où C est une constante numérique, et C' dépend de $\overline{F}_0$ et de $\|h\|_\infty$.

4) Sélection de modèle et résultat.

- Soit $\mathcal{M}_n = \{S_m, m = 1, \ldots, N_n\}$ une collection de s.e.v. de $L^2(A)$, qui vérifie :
  - $S_m \subset S_{N_n}$, pour tout $m \leq N_n$.
  - $dim(S_{N_n}) \leq \frac{n}{(\ln n)^2}$.
- Soit $\{\widehat{h}_m, m = 1, \ldots, N_n\}$ la collection d'estimateurs associée.

Theorem

Soit $A > 0$ une constante, et : $\widehat{m} = \arg\min[\gamma_n(\widehat{h}_m) + A\|h\|_\infty \frac{dim(S_m)}{n}]$, alors :

$$\mathbb{E}[\|\widehat{h}_{\widehat{m}} - h\|^2_{\overline{F}_Y}] \leq C \inf_{m \in \mathcal{M}_n} \{\|h - h_m\|^2_{\overline{F}_Y} + A\|h\|_\infty \frac{dim(S_m)}{n}\} + \frac{C'}{n}$$

où C est une constante numérique, et C' dépend de $\overline{F}_0$ et de $\|h\|_\infty$.

- **Remarque :** $\|h\|_\infty$ peut être remplacé par un estimateur.