Analyse de durées de vie :

Analyse séquentielle du modèle des risques proportionnels et tests d'homogénéité.

CHRISTELLE BREUILS

15 décembre 2003

Université de Technologie de Compiègne

Laboratoire de Mathématiques Appliquées de Compiègne

Plan de l'exposé

- Intervalles séquentiels d'arrêt pour le modèle de Cox
  - Introduction (règles séquentielles, modèle de Cox)
  - Propriétés des estimations séquentielles
  - Simulations

Plan de l'exposé

- Intervalles séquentiels d'arrêt pour le modèle de Cox
  - Introduction (règles séquentielles, modèle de Cox)
  - Propriétés des estimations séquentielles
  - Simulations
- Tests d'homogénéité
  - Modèle adopté et notations
  - Statistique de test et propriétés

Première partie : Intervalles séquentiels d'arrêt pour le modèle de Cox

Construction des règles d'arrêt

Cadre

Soit β_0, le paramètre réel inconnu d'une loi ℓ.

Soit $\hat{\beta}_n$ un estimateur de β_0 pour un n-échantillon issu de ℓ qui est :

- convergent : $\hat{\beta}_n \xrightarrow{p.s.} \beta_0$,

Construction des règles d'arrêt

Cadre

Soit β_0, le paramètre réel inconnu d'une loi ℓ.

Soit $\hat{\beta}_n$ un estimateur de β_0 pour un n-échantillon issu de ℓ qui est :

- convergent : $\hat{\beta}_n \xrightarrow{p.s.} \beta_0$,
- asymptotiquement normal de variance asymptotique σ^2 : $\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$, avec $\sigma > 0$,

Construction des règles d'arrêt

Cadre

Soit β_0, le paramètre réel inconnu d'une loi ℓ.

Soit $\hat{\beta}_n$ un estimateur de β_0 pour un n-échantillon issu de ℓ qui est :

- convergent : $\hat{\beta}_n \xrightarrow{p.s.} \beta_0$,
- asymptotiquement normal de variance asymptotique σ^2 : $\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$, avec $\sigma > 0$,
- et tel que l'on dispose d'un estimateur $\hat{\sigma}_n^2$ de σ^2 convergent.

Construction des règles d'arrêt

Cadre

Soit β_0, le paramètre réel inconnu d'une loi ℓ.

Soit $\hat{\beta}_n$ un estimateur de β_0 pour un n-échantillon issu de ℓ qui est :

- convergent : $\hat{\beta}_n \xrightarrow{p.s.} \beta_0$,
- asymptotiquement normal de variance asymptotique σ^2 : $\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$, avec $\sigma > 0$,
- et tel que l'on dispose d'un estimateur $\hat{\sigma}_n^2$ de σ^2 convergent.

Objectif

On fixe $d > 0$. On cherche à déterminer la taille minimale de l'échantillon pour que l'amplitude de l'intervalle de confiance (au niveau 95%) soit inférieure à $2d$.

Une définition

Soit u_α tel que $2\Phi(u_\alpha) - 1 = 1 - \alpha$,
(*avec* Φ *la fonction de répartition d'une loi normale centrée réduite*).
Ainsi $P\left(|\hat{\beta}_n - \beta_0| \leq \frac{u_\alpha \hat{\sigma}_n}{\sqrt{n}}\right) \to 1 - \alpha$ et l'amplitude de l'I.C. est $\frac{2\hat{\sigma}_n u_\alpha}{\sqrt{n}}$.
On pose alors

$$N_d = N_{d,n_0,\alpha} = \min\left\{n \geq n_0; n \geq \frac{\hat{\sigma}_n^{*2} u_\alpha^2}{d^2}\right\}.$$

Remarques

- n_0 est la taille minimale de l'échantillon considéré.

Une définition

Soit u_α tel que $2\Phi(u_\alpha) - 1 = 1 - \alpha$,
(*avec* Φ *la fonction de répartition d'une loi normale centrée réduite*).
Ainsi $P\left(|\hat{\beta}_n - \beta_0| \leq \frac{u_\alpha \hat{\sigma}_n}{\sqrt{n}}\right) \to 1 - \alpha$ et l'amplitude de l'I.C. est $\frac{2\hat{\sigma}_n u_\alpha}{\sqrt{n}}$.
On pose alors

$$N_d = N_{d,n_0,\alpha} = \min\left\{n \geq n_0; n \geq \frac{\hat{\sigma}_n^{*2} u_\alpha^2}{d^2}\right\}.$$

Remarques

- n_0 est la taille minimale de l'échantillon considéré.
- $\hat{\sigma}_n^{*2}$ est choisi strictement positif et convergent vers σ^2.

Propriétés des règles d'arrêt

Proposition 1

- *Existence* N_d existe presque sûrement.
- *Consistance* $N_d \xrightarrow[d \to 0^+]{p.s.} +\infty$.
- *Régularité* $N_d / n_d \xrightarrow[d \to 0^+]{P} 1$, avec

$$n_d = \min \left\{ n \geq n_0 ; n \geq \frac{u_\alpha^2 \sigma^2}{d^2} \right\}.$$

Remarque

Ces résultats sont obtenus pour la variable d'arrêt que nous avons choisie. (voir par exemple Ghosh *et al.* 1997)

Type de données

- Objectif : établir une relation entre la distribution du temps de fonctionnement (temps de survenue de $\mathcal{E}$) et des variables explicatives, pour un événement $\mathcal{E}$ donné.

Type de données

- Objectif : établir une relation entre la distribution du temps de fonctionnement (temps de survenue de $\mathcal{E}$) et des variables explicatives, pour un événement $\mathcal{E}$ donné.
- Considérons n matériels et $\mathcal{E}$, un événement donné. Pour chaque matériel i, on répertorie

Type de données

- Objectif : établir une relation entre la distribution du temps de fonctionnement (temps de survenue de $\mathcal{E}$) et des variables explicatives, pour un événement $\mathcal{E}$ donné.
- Considérons n matériels et $\mathcal{E}$, un événement donné. Pour chaque matériel i, on répertorie
  - sa date d'entrée dans l'étude,

Type de données

- Objectif : établir une relation entre la distribution du temps de fonctionnement (temps de survenue de $\mathcal{E}$) et des variables explicatives, pour un événement $\mathcal{E}$ donné.
- Considérons n matériels et $\mathcal{E}$, un événement donné. Pour chaque matériel i, on répertorie
  - sa date d'entrée dans l'étude,
  - le temps écoulé jusqu'à l'événement $\mathcal{E}$ étudié,

Type de données

- Objectif : établir une relation entre la distribution du temps de fonctionnement (temps de survenue de $\mathcal{E}$) et des variables explicatives, pour un événement $\mathcal{E}$ donné.
- Considérons n matériels et $\mathcal{E}$, un événement donné. Pour chaque matériel i, on répertorie
  - sa date d'entrée dans l'étude,
  - le temps écoulé jusqu'à l'événement $\mathcal{E}$ étudié,
  - les valeurs de p variables explicatives $Z_i = (Z_i^{(1)}, \ldots, Z_i^{(p)})$ (peuvent dépendre du temps).

Modèles à risques proportionnels

- il lie les variables explicatives au taux de défaillance (Cox, 1972) :

$$\lambda(t; Z) = \lambda_0(t) e^{\beta_0^T Z(t)},$$

avec

$$\begin{cases} \beta_0 \in \mathbb{R}^p, & \text{un paramètre inconnu,} \\ \lambda_0, & \text{taux de défaillance inconnu.} \end{cases}$$

Modèles à risques proportionnels

- il lie les variables explicatives au taux de défaillance (Cox, 1972) :

$$\lambda(t; Z) = \lambda_0(t) e^{\beta_0^T Z(t)},$$

avec

$$\begin{cases} \beta_0 \in \mathbb{R}^p, & \text{un paramètre inconnu,} \\ \lambda_0, & \text{taux de défaillance inconnu.} \end{cases}$$

- Censure : on observe effectivement $\mathcal{E}$ pour le $i^{ème}$ matériel seulement s'il n'a pas été exclu de l'étude avant.
  ➪ Données censurées à droite.

Données

$(X_i, Z_i, \delta_i)_{1 \leq i \leq n}$ indépendantes avec

$$\begin{cases} X_i &= T_i \wedge C_i & \text{durée de fonctionnement observée,} \\ \delta_i &= \mathbb{1}_{\{T_i \leq C_i\}} & \text{fonction indicatrice d'une panne,} \end{cases}$$

où pour le matériel i,

$$\begin{cases} C_i & \text{variable de censure,} \\ T_i & \text{temps écoulé jusqu'à } \mathcal{E}, \\ Z_i & \text{vecteur des variables explicatives.} \end{cases}$$

Remarque : Z_i peut dépendre du temps, on la notera toutefois Z_i pour simplifier les notations.

Estimation et propriétés

- Vraisemblance partielle de Cox (1972): on note $\mathcal{R}_i$, l'ensemble de risque (*ensemble des matériels fonctionnant à l'instant* T_i). La log-vraisemblance s'écrit :

$$C_n(\beta) = \sum_{i=1}^{n} \left(\delta_i \beta Z_i - \log \left(\sum_{l \in \mathcal{R}_i} \exp(\beta Z_l) \right) \right) \mathbb{1}_{\{X_i \leq \tau\}},$$

elle est maximisée par l'estimateur $\hat{\beta}_n = \hat{\beta}_n(\tau)$ de β_0 (τ est la borne supérieure de l'intervalle d'étude).

Estimation et propriétés

- Vraisemblance partielle de Cox (1972): on note $\mathcal{R}_i$, l'ensemble de risque (*ensemble des matériels fonctionnant à l'instant* T_i). La log-vraisemblance s'écrit :

$$C_n(\beta) = \sum_{i=1}^{n} \left(\delta_i \beta Z_i - \log \left(\sum_{l \in \mathcal{R}_i} \exp(\beta Z_l) \right) \right) \mathbb{1}_{\{X_i \leq \tau\}},$$

  elle est maximisée par l'estimateur $\hat{\beta}_n = \hat{\beta}_n(\tau)$ de β_0 (τ est la borne supérieure de l'intervalle d'étude).

- Sous les hypothèses d'Andersen et Gill (1982), on a :
  - $\hat{\beta}_n \xrightarrow{P} \beta_0$,
  - $\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$ (où σ^2 est définie positive),
  - on dispose d'un estimateur convergent $\hat{\sigma}_n^2$ de σ^2 .

Modèle semi-paramétrique d'étude

- Caractéristiques du modèle
  - $X_1, \ldots, X_n$ issues d'une distribution $\mathbb{P}$ de probabilité sur $(\mathbb{X}, \mathscr{B})$, espace euclidien muni de sa tribu borélienne,
  - β et θ suffisent à déterminer $\mathbb{P}$ entièrement et de manière unique,
  - l'estimation de $\beta \in \mathbb{R}^p$ ne requiert pas celle de $\theta \in \Theta$ (où $\dim(\Theta) \leq +\infty$).

Modèle semi-paramétrique d'étude

- Caractéristiques du modèle
  - $X_1, \ldots, X_n$ issues d'une distribution $\mathbb{P}$ de probabilité sur $(\mathbb{X}, \mathscr{B})$, espace euclidien muni de sa tribu borélienne,
  - β et θ suffisent à déterminer $\mathbb{P}$ entièrement et de manière unique,
  - l'estimation de $\beta \in \mathbb{R}^p$ ne requiert pas celle de $\theta \in \Theta$ (où $\dim(\Theta) \leq +\infty$).
- Hypothèses sur la fonction d'estimation
  - $\beta \mapsto C_n(\beta)$ fonction d'estimation de β_0, vraie valeur de β,
  - $C_n \in \mathscr{C}^2(\mathbb{R}^p)$,
  - C_n concave sur un voisinage de β_0.

Modèle semi-paramétrique d'étude

- Caractéristiques du modèle
  - $X_1, \ldots, X_n$ issues d'une distribution $\mathbb{P}$ de probabilité sur $(\mathbb{X}, \mathscr{B})$, espace euclidien muni de sa tribu borélienne,
  - β et θ suffisent à déterminer $\mathbb{P}$ entièrement et de manière unique,
  - l'estimation de $\beta \in \mathbb{R}^p$ ne requiert pas celle de $\theta \in \Theta$ (où $\dim(\Theta) \leq +\infty$).
- Hypothèses sur la fonction d'estimation
  - $\beta \mapsto C_n(\beta)$ fonction d'estimation de β_0, vraie valeur de β,
  - $C_n \in \mathscr{C}^2(\mathbb{R}^p)$,
  - C_n concave sur un voisinage de β_0.
- Estimation

$$\hat{\beta}_n = \text{argmax}_{\beta \in \mathbb{R}^p} C_n(\beta).$$

Notion de convergence complète

Définition

On dit que $(X_n)_{n\geq 1}$, suite de variables aléatoires réelles converge complètement vers X et on note $X_n \xrightarrow{c} X$, si :

$$\forall \varepsilon > 0, \sum_{n\geq 1} P\left(|X_n - X| > \varepsilon\right) < \infty.$$

Remarque

La convergence complète entraîne la convergence presque sûre.

Notation

On note U_n et $-\mathcal{I}_n$, les dérivées première et seconde de C_n/n.

Théorème1 *(adapté de Dmitrienko et Govindarajulu, 2000)*

Sous les conditions suivantes

- (C0) $\sqrt{n}U_n(\beta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathcal{I}(\beta_0))$, lorsque $n \to +\infty$;
- (C1) $|\mathcal{I}_n(\beta) - \mathbb{E}\mathcal{I}_n(\beta_0)| \xrightarrow{c} 0$ dans un voisinage de plus en plus petit de β_0, i.e. pour $\varepsilon > 0$,

$$\exists \delta_\varepsilon > 0; \sum_{n \geq 0} P\left(\sup_{\beta \in B(\beta_0; \delta_\varepsilon)} |\mathcal{I}_n(\beta) - \mathbb{E}\mathcal{I}_n(\beta_0)| > \varepsilon \right) < +\infty;$$

- (C2) $\mathbb{E}\mathcal{I}_n(\beta_0) \underset{n \to \infty}{\to} \mathcal{I}(\beta_0) > 0$;
- (C3) $(|U_n(\beta_0)|)_{n \geq 1}$ converge complètement vers 0 ;

Théorème1 *(adapté de Dmitrienko et Govindarajulu, 2000)*

Sous les conditions suivantes

- (C0) $\sqrt{n}U_n(\beta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathcal{I}(\beta_0))$, lorsque $n \to +\infty$;
- (C1) $|\mathcal{I}_n(\beta) - \mathbb{E}\mathcal{I}_n(\beta_0)| \xrightarrow{c} 0$ dans un voisinage de plus en plus petit de β_0, i.e. pour $\varepsilon > 0$,

$$\exists \delta_\varepsilon > 0; \sum_{n \geq 0} P\left(\sup_{\beta \in B(\beta_0; \delta_\varepsilon)} |\mathcal{I}_n(\beta) - \mathbb{E}\mathcal{I}_n(\beta_0)| > \varepsilon \right) < +\infty;$$

- (C2) $\mathbb{E}\mathcal{I}_n(\beta_0) \underset{n \to \infty}{\to} \mathcal{I}(\beta_0) > 0$;
- (C3) $(|U_n(\beta_0)|)_{n \geq 1}$ converge complètement vers 0 ;

Alors : $\hat{\beta}_n \to \beta_0$ pour la convergence complète.

Hypothèses et résultats de convergence

- Définition

  Soit $(T_n)_{n\geq 1}$, une suite de variables aléatoires entières. On dira qu'elle est régulière s'il existe une suite d'entiers positifs $(t_n)_{n\geq 1}$ telle que $T_n/t_n \xrightarrow{P} 1$ et $t_n \to \infty$.

Hypothèses et résultats de convergence

- Définition

  Soit $(T_n)_{n\geq 1}$, une suite de variables aléatoires entières. On dira qu'elle est régulière s'il existe une suite d'entiers positifs $(t_n)_{n\geq 1}$ telle que $T_n/t_n \xrightarrow{P} 1$ et $t_n \to \infty$.

- Théorème 2 *(adapté de Dmitrienko et Govindarajulu, 2000)*

  Sous les hypothèses (C0)-(C3), et si, en outre :

$$\text{(C4)} \qquad \sqrt{T_n} U_{T_n}(\beta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathcal{I}(\beta_0)),$$

  pour toute suite régulière de variables aléatoires entières $(T_n)_{n\geq 1}$, alors :

$$\sqrt{T_n}(\hat{\beta}_{T_n} - \beta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathcal{I}^{-1}(\beta_0)).$$

Notations pour le modèle de Cox ($p = 1$)

- $\frac{1}{n}\frac{\partial C_n}{\partial \beta}(\beta, \tau) = U_n(\beta, \tau) = U_n(\beta)$ *(score)*,
- $\frac{-1}{n}\frac{\partial^2 C_n}{\partial \beta^2}(\beta, \tau) = \mathcal{I}_n(\beta, \tau) = \mathcal{I}_n(\beta)$ *(information de Fisher)*,
- $S^{(m)}(\beta, t) = \frac{1}{n}\sum_{l=1}^{n} Z_l(t)^m Y_l(t) e^{\beta Z_l(t)}; m \in \{0, 1, 2\}$,
- $Y_l(t) = \mathbb{1}_{\{X_l \geq t\}}$.

Hypothèses (Andersen et Gill, 1982)

- (H1) $\int_0^\tau \lambda_0(t)dt$ est finie ;
- (H2) Il existe un voisinage $\mathcal{B}$ de β_0, des fonctions $s^{(0)}, s^{(1)}$ et $s^{(2)}$ définies sur $\mathcal{B} \times [0, \tau]$ telles que, pour $j \in [\![0, 2]\!]$,

$$\sup_{t\in[0,\tau],\beta\in\mathcal{B}} \left| \quad S^{(j)}(\beta, t) - s^{(j)}(\beta, t) \right| \longrightarrow 0;$$

- (H3) En posant $e = s^{(1)}/s^{(0)}$ et $v = s^{(2)}/s^{(0)} - e^2$, $s^{(0)}, s^{(1)}$ et $s^{(2)}$ sont bornées, continues unif. en $t \in [0, \tau]$, $s^{(0)} > \gamma > 0$ sur $\mathcal{B} \times [0, \tau]$, et $\int_0^\tau v(\beta_0, t)s^{(0)}(\beta_0, t)\lambda_0(t)dt > 0$;

Hypothèses modifiées

- (H1) $\int_0^\tau \lambda_0(t)dt$ est finie ;
- (H2bis) Il existe un voisinage $\mathcal{B}$ de β_0, des fonctions $s^{(0)}, s^{(1)}$ et $s^{(2)}$ définies sur $\mathcal{B} \times [0, \tau]$ telles que, pour $j \in [\![0, 2]\!]$,

$$\sup_{t\in[0,\tau],\beta\in\mathcal{B}} \left| \mathbb{E} S^{(j)}(\beta, t) - s^{(j)}(\beta, t) \right| \longrightarrow 0;$$

- (H3) En posant $e = s^{(1)}/s^{(0)}$ et $v = s^{(2)}/s^{(0)} - e^2$, $s^{(0)}, s^{(1)}$ et $s^{(2)}$ sont bornées, continues unif. en $t \in [0, \tau]$, $s^{(0)} > \gamma > 0$ sur $\mathcal{B} \times [0, \tau]$, et $\int_0^\tau v(\beta_0, t)s^{(0)}(\beta_0, t)\lambda_0(t)dt > 0$;
- (H4) $\{Z_i(t); t \in [0; \tau]\}$ est à variations bornées (unif. en i), i.e. $\exists B < \infty; \forall i \geq 1, \int_0^\tau |Z_i(ds)| \leq B$.

Hypothèses modifiées

- (H1) $\int_0^\tau \lambda_0(t)dt$ est finie ;
- (H2bis) Il existe un voisinage $\mathcal{B}$ de β_0, des fonctions $s^{(0)}, s^{(1)}$ et $s^{(2)}$ définies sur $\mathcal{B} \times [0, \tau]$ telles que, pour $j \in [\![0, 2]\!]$,

$$\sup_{t\in[0,\tau],\beta\in\mathcal{B}} \left| \mathbb{E} S^{(j)}(\beta, t) - s^{(j)}(\beta, t) \right| \longrightarrow 0;$$

- (H3) En posant $e = s^{(1)}/s^{(0)}$ et $v = s^{(2)}/s^{(0)} - e^2$, $s^{(0)}, s^{(1)}$ et $s^{(2)}$ sont bornées, continues unif. en $t \in [0, \tau]$, $s^{(0)} > \gamma > 0$ sur $\mathcal{B} \times [0, \tau]$, et $\int_0^\tau v(\beta_0, t)s^{(0)}(\beta_0, t)\lambda_0(t)dt > 0$;
- (H4) $\{Z_i(t); t \in [0; \tau]\}$ est à variations bornées (unif. en i), i.e. $\exists B < \infty; \forall i \geq 1, \int_0^\tau |Z_i(ds)| \leq B$.

Sous (H1-H4), les conditions (C0-C4) sont vérifiées.

Proposition

Si $(X_n)_{n\geq 1}$ et $(Y_n)_{n\geq 1}$ sont telles que $X_n \xrightarrow{c} X$ et $Y_n \xrightarrow{c} Y$, alors :

- $X_n + Y_n \xrightarrow{c} X + Y$;
- si X et Y sont bornées, $X_n Y_n \xrightarrow{c} XY$;
- si X est p.s. minorée par $x > 0$, alors $1/X_n \xrightarrow{c} 1/X$;
- Soit I un intervalle réel borné et $(f_n(t); t \in I)_{n\geq 1}$, $(g_n(t); t \in I)_{n\geq 1}$, tels que :
  - $\sup_{t\in I} |f_n(t) - f(t)| \xrightarrow{c} 0$, et $\sup_{t\in I} |g_n(t) - g(t)| \xrightarrow{c} 0$, où f et g sont déterministes et g est continue sur I,
  - $\exists B > 0$ telle que $\forall n \geq 1$, $\int_I |df_n| \leq B$ et $\int_I |df| \leq B$,

Alors $$\int_I g_n df_n \xrightarrow{c} \int_I g df.$$

Preuve de (C3) : notations

Notons, pour $t \in [0, \tau], i \in [\![1, n]\!]$:

$$\begin{aligned} N_i(t) &= \mathbb{1}_{\{X_i \leq t, \delta_i = 1\}}; \\ \overline{N}(t) &= \sum_{i=1}^{n} N_i(t); \\ Y_i(t) &= \mathbb{1}_{\{X_i \geq t\}}; \\ U_n(\beta, t) &= \frac{\beta}{\partial \beta} C_n(\beta, t). \end{aligned}$$

$Y_i(t) = 1$ si i est encore dans l'étude au temps t et N_i change de valeur seulement quand $\delta_i = 1$.

Squelette de démonstration de (C3)

On sépare $U_n(\beta_0, t)$ en trois parties :

$$U_n(\beta_0, t) = \frac{1}{n} \sum_{i=1}^{n} \int_0^t Z_i(s) dM_i(s) \text{ (T1)}$$

Squelette de démonstration de (C3)

On sépare $U_n(\beta_0, t)$ en trois parties :

$$
\begin{aligned}
U_n(\beta_0, t) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t Z_i(s)dM_i(s) \text{ (T1)} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\int_0^t \left(\frac{S_n^{(1)}(\beta_0, s)}{S_n^{(0)}(\beta_0, s)} - \frac{s^{(1)}(\beta_0, s)}{s^{(0)}(\beta_0, s)}\right) dM_i(s) \text{ (T2)}
\end{aligned}
$$

Squelette de démonstration de (C3)

On sépare $U_n(\beta_0, t)$ en trois parties :

$$
\begin{aligned}
U_n(\beta_0, t) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t Z_i(s)dM_i(s) \text{ (T1)} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\int_0^t \left(\frac{S_n^{(1)}(\beta_0, s)}{S_n^{(0)}(\beta_0, s)} - \frac{s^{(1)}(\beta_0, s)}{s^{(0)}(\beta_0, s)}\right) dM_i(s) \text{ (T2)} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\int_0^t \frac{s^{(1)}(\beta_0, s)}{s^{(0)}(\beta_0, s)} dM_i(s) \text{ (T3)},
\end{aligned}
$$

où $M_i(t) = N_i(t) - \int_0^t Y_i(s)\exp(\beta_0 Z_i)\lambda_0(s)ds$.

Squelette de démonstration de (C3)

On sépare $U_n(\beta_0, t)$ en trois parties :

$$
\begin{aligned}
U_n(\beta_0, t) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t Z_i(s)dM_i(s) \text{ (T1)} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\int_0^t \left(\frac{S_n^{(1)}(\beta_0, s)}{S_n^{(0)}(\beta_0, s)} - \frac{s^{(1)}(\beta_0, s)}{s^{(0)}(\beta_0, s)}\right) dM_i(s) \text{ (T2)} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\int_0^t \frac{s^{(1)}(\beta_0, s)}{s^{(0)}(\beta_0, s)} dM_i(s) \text{ (T3)},
\end{aligned}
$$

où $M_i(t) = N_i(t) - \int_0^t Y_i(s)\exp(\beta_0 Z_i)\lambda_0(s)ds$. On considère ensuite les trois termes séparément. (T1) se traite d'une façon similaire à (T2).

Preuve de (C3) : terme (T2)

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^t \left(\frac{S_n^{(1)}(\beta_0,s)}{S_n^{(0)}(\beta_0,s)} - \frac{s^{(1)}(\beta_0,s)}{s^{(0)}(\beta_0,s)}\right) dM_i(s)$$

Utilisation de la stabilité de la convergence complète par passage à l'intégrale :

- $\sup_{s\in[0,\tau]} \left|\frac{1}{n}\sum_{i=1}^{n} M_i(s)\right| \xrightarrow{c} 0$ et $\frac{1}{n}\sum_{i=1}^{n} M_i$ à variations totales bornées

Preuve de (C3) : terme (T2)

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^t \left(\frac{S_n^{(1)}(\beta_0,s)}{S_n^{(0)}(\beta_0,s)} - \frac{s^{(1)}(\beta_0,s)}{s^{(0)}(\beta_0,s)}\right) dM_i(s)$$

Utilisation de la stabilité de la convergence complète par passage à l'intégrale :

- $\sup_{s\in[0,\tau]} \left|\frac{1}{n}\sum_{i=1}^{n} M_i(s)\right| \xrightarrow{c} 0$ et $\frac{1}{n}\sum_{i=1}^{n} M_i$ à variations totales bornées
- $\sup_{s\in[0,\tau]} \left|\frac{S_n^{(1)}(\beta_0,s)}{S_n^{(0)}(\beta_0,s)} - \frac{s^{(1)}(\beta_0,s)}{s^{(0)}(\beta_0,s)}\right| \xrightarrow{c} 0$

Preuve de (C3) : terme (T2)

$$\frac{1}{n}\sum_{i=1}^{n}\int_0^t \left(\frac{S_n^{(1)}(\beta_0,s)}{S_n^{(0)}(\beta_0,s)} - \frac{s^{(1)}(\beta_0,s)}{s^{(0)}(\beta_0,s)}\right) dM_i(s)$$

Utilisation de la stabilité de la convergence complète par passage à l'intégrale :

- $\sup_{s\in[0,\tau]} \left|\frac{1}{n}\sum_{i=1}^{n} M_i(s)\right| \xrightarrow{c} 0$ et $\frac{1}{n}\sum_{i=1}^{n} M_i$ à variations totales bornées
- $\sup_{s\in[0,\tau]} \left|\frac{S_n^{(1)}(\beta_0,s)}{S_n^{(0)}(\beta_0,s)} - \frac{s^{(1)}(\beta_0,s)}{s^{(0)}(\beta_0,s)}\right| \xrightarrow{c} 0$

vient du fait que (Pollard, 1990 ; van der Vaart et Wellner, 1996) :

$$\forall \varepsilon > 0, \forall l \in [\![0,2]\!],$$
$$P\left(\sup_{s\in[0,\tau]} \left|S_n^{(l)}(\beta_0,s) - I\!E S_n^{(l)}(\beta_0,s)\right| > \varepsilon\right) \leq C_1 \exp(-C_2 n).$$

Preuve de (C3) : terme (T3)

On utilise le résultat suivant (Shorack et Wellner, 1986):
Soit $(Q_n)_{n\geq 1}$, une suite de martingales locales, localement uniformément intégrables et nulles en 0. Supposons que l'on ait, pour $n \geq 1$,

$$\|\Delta Q_n\|_0^\tau \leq \frac{c}{\sqrt{n}},$$

où c est une constante positive. Si

$$\sum_{n\geq 1} I\!P\left(\langle Q_n\rangle(\tau) \geq n^{-1/4}\right) < +\infty,$$

alors $\|M_n\|_0^\tau \xrightarrow{c} 0$.
On applique cela à $Q_n(t) = n^{-1}\sum_{i=1}^n \int_0^t \frac{s^{(1)}(\beta_0,s)}{s^{(0)}(\beta_0,s)} dM_i(s)$.

Preuve de (C4) : condition d'Anscombe

Il suffit de montrer :

$$\forall \varepsilon > 0, \exists \delta > 0;$$

$$\lim_{n\to\infty} P\left(\max_{0\leq k\leq \delta n} \left| \sqrt{n+k} U_{n+k}(\beta_0, t) - \sqrt{n} U_n(\beta_0, t) \right| > \varepsilon \right) < \varepsilon;$$

pour montrer (C4).
On dira aussi que $\left(\sqrt{n} U_n(\beta_0, t)\right)_{n\geq 1}$ est une suite uniformément continue en probabilité.

Résultats supplémentaires

- Sous les hypothèses (H1-H4) et si $\beta_0 \in \mathbb{R}^p$, $p > 1$, pour les ellipsoïdes de confiance par

$$CR_n = \{\beta \in R^p; (\hat{\beta}_n - \beta)^T n\mathcal{I}_n(\hat{\beta}_n)(\hat{\beta}_n - \beta) \leq \chi^2_{p,1-\alpha}\},$$

on a des résultats de convergence pour N_d définie par :

$$N_d = \min \left\{ n \geq n_0; n \geq \frac{\chi^2_{p,1-\alpha}}{d^2 \phi^*_{\min}\left(\mathcal{I}_n(\hat{\beta}_n)\right)} \right\},$$

où $\phi_{\min}(A)$ est la valeur propre minimale d'une matrice A.

Résultats supplémentaires

- Sous les hypothèses (H1-H4) et si $\beta_0 \in \mathbb{R}^p$, $p > 1$, pour les ellipsoïdes de confiance par

$$CR_n = \{\beta \in R^p; (\hat{\beta}_n - \beta)^T n\mathcal{I}_n(\hat{\beta}_n)(\hat{\beta}_n - \beta) \leq \chi^2_{p,1-\alpha}\},$$

on a des résultats de convergence pour N_d définie par :

$$N_d = \min \left\{ n \geq n_0; n \geq \frac{\chi^2_{p,1-\alpha}}{d^2 \phi^*_{\min}\left(\mathcal{I}_n(\hat{\beta}_n)\right)} \right\},$$

où $\phi_{\min}(A)$ est la valeur propre minimale d'une matrice A.

- Avec quelques hypothèses supplémentaires, on a $\sqrt{N_d} - \sqrt{n_d} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \gamma(\beta_0, \tau)\right), d \to 0^+,$ où $\gamma(\beta_0, \tau) > 0$.

Simulations numériques

Nature des données simulées

- les variables de censure $C_i \sim \mathcal{W}(10^{-6}; 2)$,
- la fonction de risque de base $\lambda_0 \sim \mathcal{W}(10^{-2}; 0,8)$,
- les variables explicatives $Z_i \sim \mathcal{U}([0,1])$.

On obtient un échantillon de (X_i, δ_i, Z_i), pour i variant de 1 à N.

Paramètres donnés par l'utilisateur :

- d, amplitude de l'intervalle de confiance,
- N et n_0, les tailles totale et initiale de l'échantillon simulé,
- β_0, paramètre à estimer,
- $1 - \alpha$, taux de confiance asymptotique.

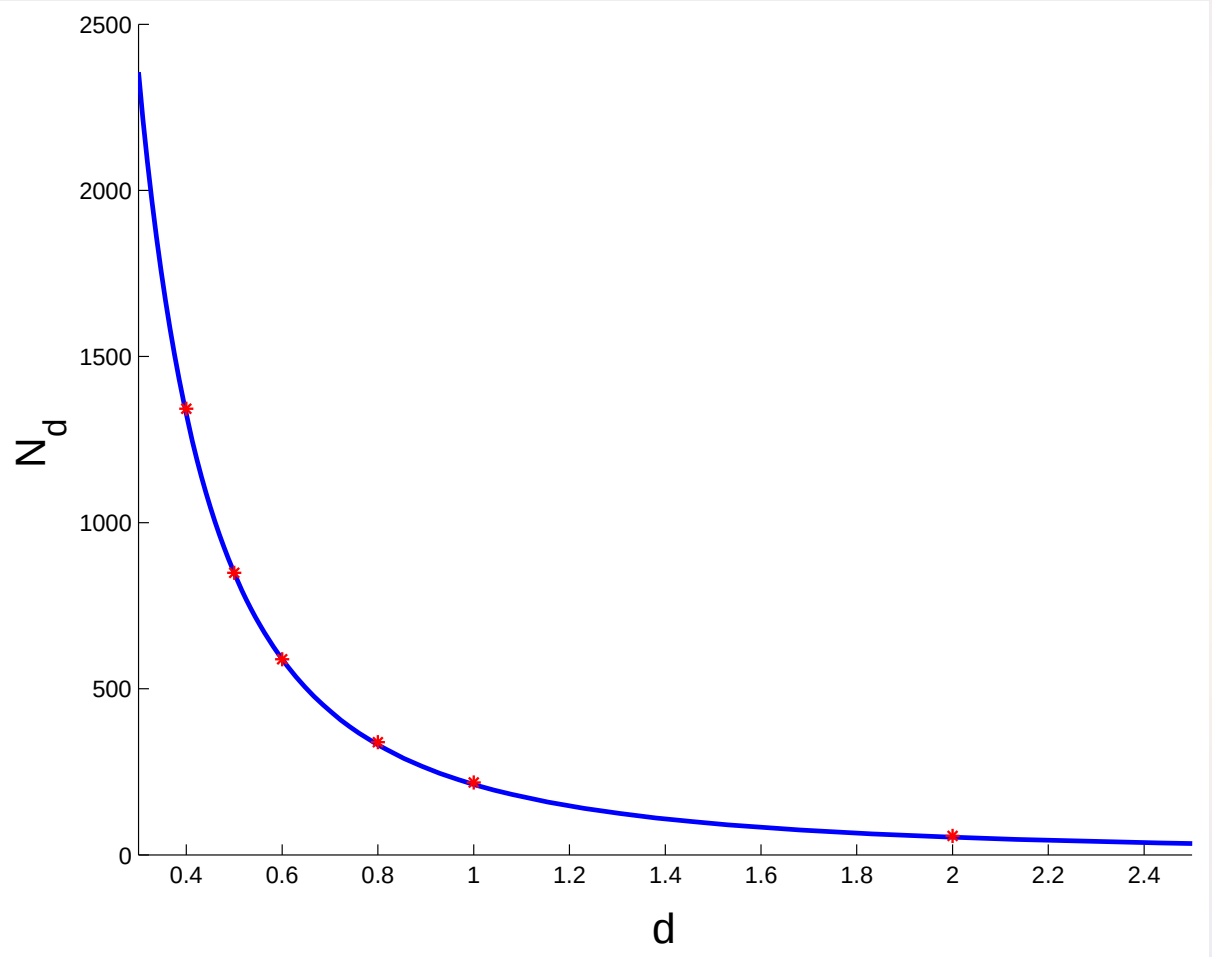
L'algorithme calcule $N_d, \hat{\sigma}_{N_d}, \hat{\beta}_{N_d}, \hat{\beta}_N, IC$ et répète cette procédure $A = 100$ fois.

Estimations de $\beta_0 = 1$*.*

d	2	1	0.8	0.6	0.5	0.4	0.05
$\hat{\beta}^m_{N_d}$	1.00	1.04	1.02	1.01	1.00	1.00	1.00
(e-t)	(0.56)	(0.26)	(0.17)	(0.16)	(0.15)	(0.09)	(0.01)
N^m_d	57	218	339	589	849	1343	84624
(e-t)	(8)	(19)	(21)	(26)	(31)	(25)	(367)
I.C.	91	95	99	94	93	98	94

Estimations séquentielles de $\beta_0 = 1$ pour différentes valeurs de d, $A = 100$

N_d^m **fonction de** d**, pour** $\beta_0 = 1$ **et** $A = 100$



Courbe tracée : $d \mapsto \hat{\sigma}_N^2 u_\alpha^2 / d^2$.

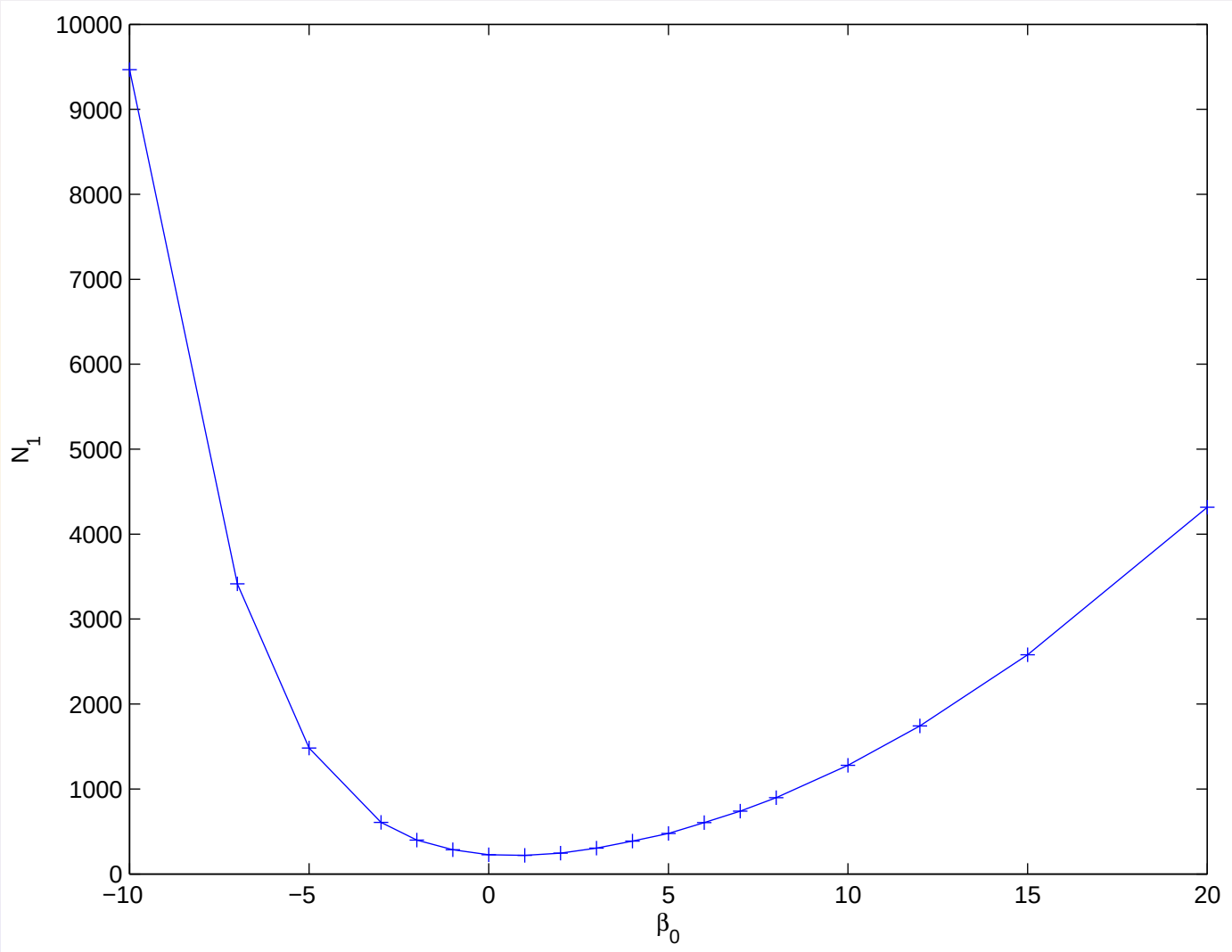
Influence de β_0 sur les estimations

- Méthode 1 : variance des A observations de $\sqrt{N_d}(\hat{\beta}_{N_d} - \beta_0)$
- Méthode 2: moyenne empirique des $\hat{\sigma}^2_{N_d}$

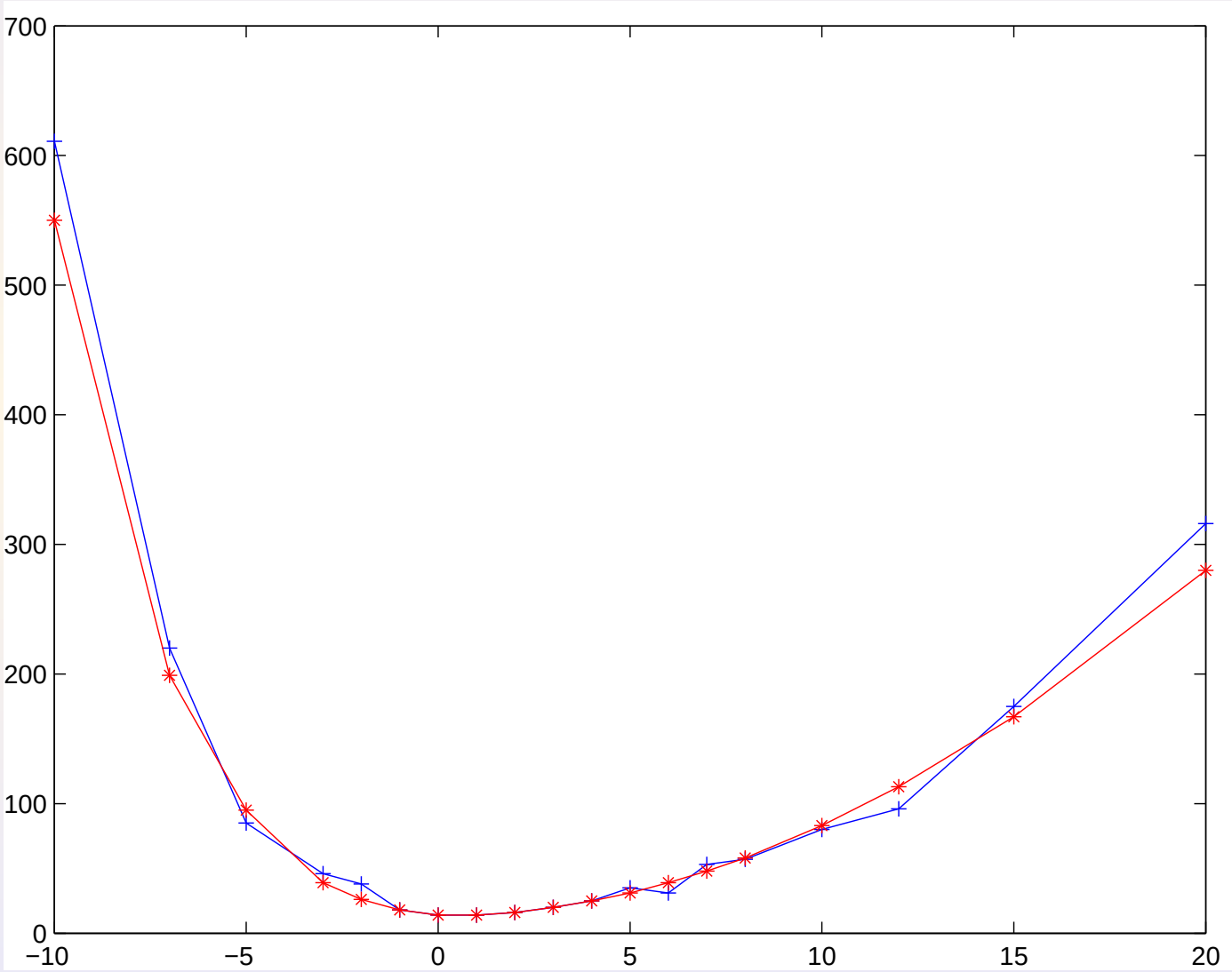
β_0	-5	-1	0	1
$\hat{\beta}^m_{N_1}$	-5.01(0.25)	-0.94(0.27)	0.09(0.20)	1.03(0.25)
N^m_1	564(41)	276(23)	258(21)	272(19)
Mét.1	34.97	19.61	10.73	16.95
Mét.2	36.43	17.76	16.63	17.54

Taux de censure de 30% , $A = 100, d = 1$

N_1^m *en fonction de* β_0



Variance estimée en fonction de β_0



Conclusions et perspectives - partie 1

On a montré des propriétés intéressantes pour notre variable d'arrêt (il suffit d'obtenir la régularité de la variable d'arrêt pour avoir les résultats de convergence). Nous prévoyons de :

- mettre en oeuvre ces techniques et étudier les résultats pour des données réelles ;
- procéder à une telle étude dans le cas d'entrées retardées ;
- considérer les estimations pour l'estimateur de Breslow de la fonction de risque cumulé et montrer des résultats similaires ;
- étudier ces procédures pour d'autres modèles (modèle à risques additifs, ...).

Seconde partie : Tests d'homogénéité

Modèle adopté

- Données
  - Z variable explicative, réelle, à valeurs dans $[0, 1]$,
  - T durée de vie,
  - Censure à droite : $X = T \wedge C, \delta = \mathbb{1}_{\{T \leq C\}}$.

Modèle adopté

- Données
  - Z variable explicative, réelle, à valeurs dans $[0,1]$,
  - T durée de vie,
  - Censure à droite : $X = T \wedge C, \delta = \mathbb{1}_{\{T \leq C\}}$.
- Hypothèse (McKeague et Utikal, 90; Beran, 81)
  La fonction de risque cumulé conditionnelle à $Z = z$ est donnée par

$$A(t;z) = \int_0^t \alpha(u;z)du,$$

où α est lipschitzienne sur $[0,1]^2$, de support supp(α) inclus dans $[0,1]^2$.

Estimation : notations

Pour $n \geq 1$, $z \in [0,1]$, $\left\{\chi_i^{(n)}; 1 \leq i \leq k_n\right\}$ une partition de $[0,1]$, on note $\chi^{(n)}(z) = \chi_i^{(n)}$, avec i tel que $z \in \chi_i^{(n)}$. On définit :

$$\mathcal{F}_z^{(n)} = \left\{s \in [0,1]; \chi^{(n)}(z) \subset \text{supp}\left(\alpha(s;\cdot)\right)\right\},$$
$$\overline{\mathcal{F}}_z^{(n)} = \left\{s \in [0,1]; \exists u, v \in \chi^{(n)}(z) \text{ avec } \alpha(s;u) > 0 \text{ et } \alpha(s;v) = 0\right\},$$

Estimation : notations

Pour $n \geq 1$, $z \in [0,1]$, $\left\{\chi_i^{(n)}; 1 \leq i \leq k_n\right\}$ une partition de $[0,1]$, on note $\chi^{(n)}(z) = \chi_i^{(n)}$, avec i tel que $z \in \chi_i^{(n)}$. On définit :

$$\mathcal{F}_z^{(n)} = \left\{s \in [0,1]; \chi^{(n)}(z) \subset \text{supp}\left(\alpha(s;\cdot)\right)\right\},$$
$$\overline{\mathcal{F}}_z^{(n)} = \left\{s \in [0,1]; \exists u, v \in \chi^{(n)}(z) \text{ avec } \alpha(s;u) > 0 \text{ et } \alpha(s;v) = 0\right\},$$

On définit les martingales M_i, pour $1 \leq i \leq n$, pour $t \in [0,1]$, par :

$$M_i(t) = M(t; z_i) = N_i(t) - \Lambda(t; z_i),$$

où

$$N_i(t) = \delta_i \mathbb{1}_{\{X_i \leq t\}}, Y_i(t) = \mathbb{1}_{\{X_i \geq t\}}, \text{ et } \Lambda(t; z_i) = \int_0^t Y_i(s)\alpha(s; z_i)ds.$$

Estimation : définitions

On définit alors des comptages par tranches donnés par :

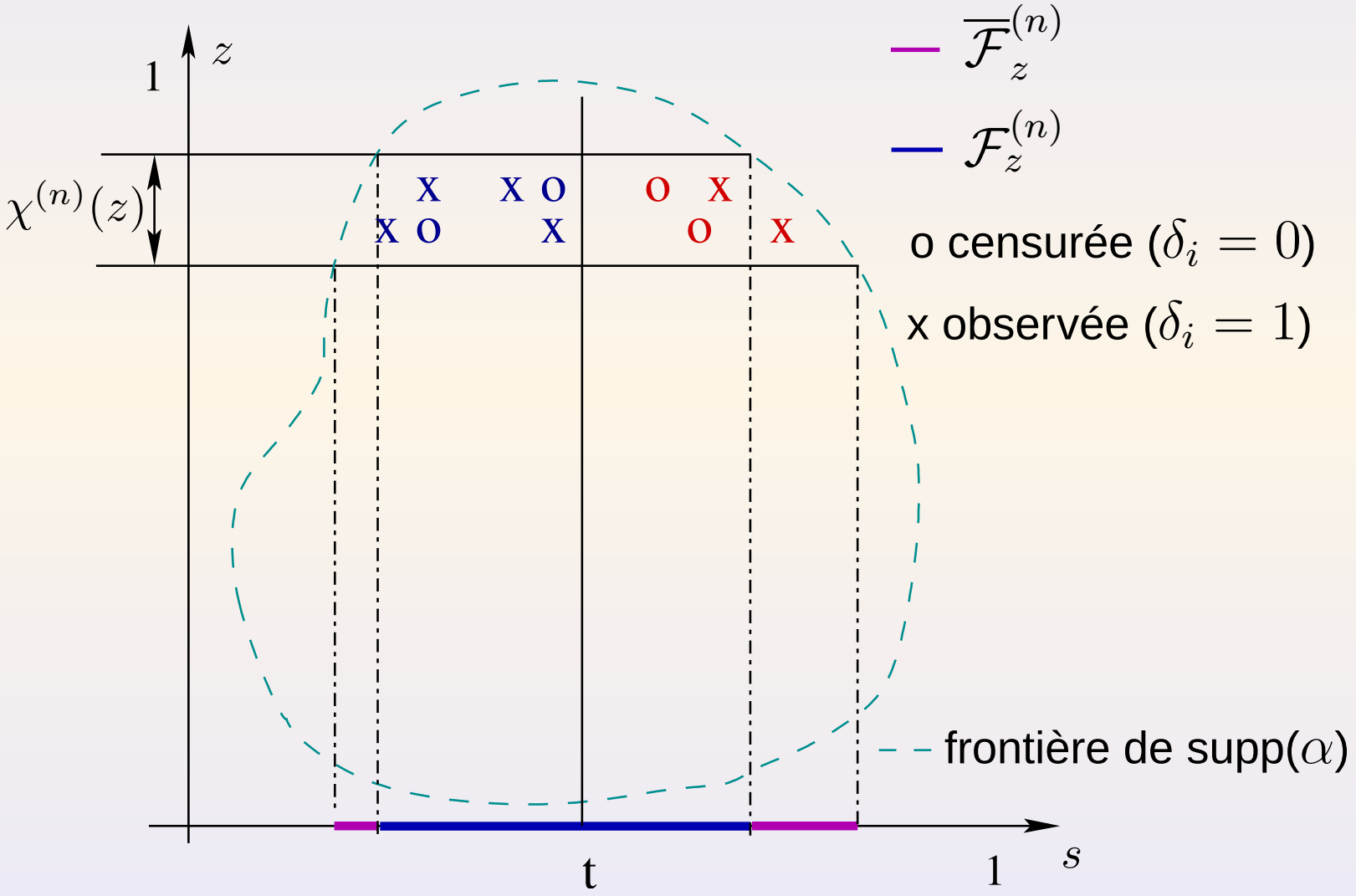
$$N^{(n)}(t;z) = \sum_{j=1}^{n} \mathbb{1}_{\{z_j \in \chi^{(n)}(z)\}} N_j(t),$$

$$Y^{(n)}(t;z) = \sum_{j=1}^{n} \mathbb{1}_{\{z_j \in \chi^{(n)}(z)\}} Y_j(t).$$

On définit $\hat{A}^{(n)}$, l'estimateur de Beran de A par :

$$\hat{A}^{(n)}(t;z) = \int_0^t \frac{N^{(n)}(ds;z)}{Y^{(n)}(s;z)}.$$

Illustration de l'estimation définie



Estimations : comportement

Si w_n est telle que $nw_n \to +\infty$ et $nw_n^2 \to 0$, avec des hypothèses supplémentaires, on a (McKeague et Utikal, 90) :

$$\sqrt{nw_n}\left(\hat{A}(\cdot, z) - A(\cdot, z)\right) \xrightarrow{\mathcal{D}} U(\cdot, z),$$

dans $\mathcal{D}([0,1])$, avec $U(\cdot, z)$ martingale gaussienne continue, centrée et de variance (g sera définie ultérieurement)

$$\text{Var}\left(U(t; z)\right) = \int_0^t \alpha(s; z) g(s; z) ds, t \in [0,1].$$

Hétérogénéité

- En notant $\boldsymbol{V} = (V_1, \ldots, V_n)^T$ de distribution G inconnue, on suppose que :
  - $\mathbb{E}\, V_i = 0$ et $\mathbb{E} V_i^2 = 1$ pour $1 \leq i \leq n$,
  - $\mathbb{E}(\boldsymbol{V}\boldsymbol{V}^T) = W$, où W est une matrice connue.

Hétérogénéité

- En notant $\boldsymbol{V} = (V_1, \ldots, V_n)^T$ de distribution G inconnue, on suppose que :
  - $\mathbb{E}\, V_i = 0$ et $\mathbb{E} V_i^2 = 1$ pour $1 \leq i \leq n$,
  - $\mathbb{E}(\boldsymbol{V}\boldsymbol{V}^T) = W$, où W est une matrice connue.
- Hypothèse pour $s \in [0, 1]$, l'intervalle d'étude, on a :

$$\alpha(s; z_i, \varepsilon_i) = \exp(\varepsilon_i)\alpha(s; z_i), \text{ où } \varepsilon_i = \sqrt{\theta} V_i \text{ et } \theta > 0.$$

Hétérogénéité

- En notant $\boldsymbol{V} = (V_1, \ldots, V_n)^T$ de distribution G inconnue, on suppose que :
  - $\mathbb{E}\, V_i = 0$ et $\mathbb{E} V_i^2 = 1$ pour $1 \leq i \leq n$,
  - $\mathbb{E}(\boldsymbol{V}\boldsymbol{V}^T) = W$, où W est une matrice connue.
- Hypothèse pour $s \in [0, 1]$, l'intervalle d'étude, on a :

$$\alpha(s; z_i, \varepsilon_i) = \exp(\varepsilon_i)\alpha(s; z_i), \text{ où } \varepsilon_i = \sqrt{\theta} V_i \text{ et } \theta > 0.$$

- Le test consiste alors à tester $\mathcal{H}_0 : \theta = 0$.

Hétérogénéité

- En notant $\boldsymbol{V} = (V_1, \ldots, V_n)^T$ de distribution G inconnue, on suppose que :
  - $\mathbb{E}\, V_i = 0$ et $\mathbb{E} V_i^2 = 1$ pour $1 \leq i \leq n$,
  - $\mathbb{E}(\boldsymbol{V}\boldsymbol{V}^T) = W$, où W est une matrice connue.
- Hypothèse pour $s \in [0, 1]$, l'intervalle d'étude, on a :

$$\alpha(s; z_i, \varepsilon_i) = \exp(\varepsilon_i)\alpha(s; z_i), \text{ où } \varepsilon_i = \sqrt{\theta} V_i \text{ et } \theta > 0.$$

- Le test consiste alors à tester $\mathcal{H}_0 : \theta = 0$.
- Exemple (voir Commenges et Jacqmin-Gadda, 97)
  On prend W telle que
  - $w_{ij} = 1$ si i et j sont dans la même famille,
  - $w_{ij} = 0$ sinon.

Construction de la statistique de test

- On calcule la fonction score τ_n:

$$\tau_n = \frac{\partial L}{\partial \theta}(\theta)\Big|_{\theta=0} = \lim_{t \to +\infty} T^{(n)}(t),$$

avec L, la log-vraisemblance, et l'on obtient :

$$T^{(n)}(t) = \sum_{i=1}^{n} \sum_{j=1}^{n} \int_0^t w_{ij} M_j(s-) dM_i(s) + \frac{1}{2} \sum_{i=1}^{n} \int_0^t dM_i(s).$$

Construction de la statistique de test

- $$T^{(n)}(t) = \sum_{i=1}^{n}\sum_{j=1}^{n}\int_0^t w_{ij} M_j(s-) dM_i(s) + \frac{1}{2}\sum_{i=1}^{n}\int_0^t dM_i(s).$$

- On remplace les martingales M_i par leur résidus $\hat{M}_i$ définis par :

  $$\hat{M}^{(n)}(t; z_i) = N_i(t) - \int_0^t Y_i(s)\frac{N^{(n)}(ds; z_i)}{Y^{(n)}(s; z_i)},$$

  on obtient :

  $$\hat{T}^{(n)}(t) = \sum_{i=1}^{n}\sum_{j=1}^{n}\int_0^t w_{ij}\hat{M}^{(n)}(s-; z_j)\hat{M}^{(n)}(ds; z_i).$$

Hypothèses $(n \to +\infty)$

- (B0) Leb $\left\{ s \in \mathcal{F}_z^{(n)}; Y^{(n)}(s;z) = 0 \right\} = o_P(1/\sqrt{n\omega_n})$.
- (B1) $\exists K > 1;\ \forall i \in [1,n],$ Card$\{j \in [1,n], w_{ij} \neq 0\} \leq K$.
- (B2a) $\sup_{z\in[0,1]}$ Leb$\{\overline{\mathcal{F}}_z^{(n)}\} = \mathcal{O}(\omega_n)$.
- (B2b) $\sup_{z\in[0,1]}$ Card $\left\{1 \leq i \leq n; z_i \in \chi^{(n)}(z)\right\} = \mathcal{O}(n\omega_n)$.
- (B3) Il existe une fonction g mesurable, positive et bornée, définie sur le support de α telle que

$$\sup_{z\in[0,1]} \int_{\mathcal{F}_z^{(n)}} \left(\frac{n\omega_n}{Y^{(n)}(s;z)} - g(s;z) \right)^2 ds \xrightarrow{P} 0.$$

Théorème

- (B4)

$$\sup_{s\in[0,1]}\left|a_n^2\sum_{i=1}^{n}\sum_{j=1}^{n}(1-\delta_{ij})w_{ij}^2\Lambda_j(s)Y_i(s)\alpha(s;z_i)-\gamma(s)\right|\xrightarrow{P}0,$$

où γ intégrable positive. On note $\Gamma(t)=\int_0^t\gamma(s)ds$.

Théorème

- (B4)

$$\sup_{s\in[0,1]} \left| a_n^2 \sum_{i=1}^{n}\sum_{j=1}^{n} (1-\delta_{ij}) w_{ij}^2 \Lambda_j(s) Y_i(s) \alpha(s;z_i) - \gamma(s) \right| \xrightarrow{P} 0,$$

où γ intégrable positive. On note $\Gamma(t) = \int_0^t \gamma(s)ds$.

Sous les hypothèses (B0-B4), pour $a_n^{-2} = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^2$, on a

$$a_n \hat{T}^{(n)} \xrightarrow{\mathcal{D}} \mathcal{U},$$

dans $\mathcal{D}([0,1])$, où $\mathcal{U}$ est une martingale gaussienne centrée de fonction de variance $\mathrm{Var}\left(\mathcal{U}(t)\right) = \Gamma(t)$.

Conclusions et perspectives - partie 2

L'étude menée sur les tests d'homogénéité présente un résultat de convergence dans un cadre non paramétrique, pour des groupes de taille finie. On poursuivra cette étude par

- des simulations numériques qui nous permettront de déceler les limites de validité du test et des hypothèses émises ;
- une approche dans le cadre de groupes de tailles croissant vers l'infini ;
- la prise en compte de variables explicatives dépendantes du temps.