# Modèles de régression pour données de survie

Feriel BOUHADJERA

EPN6 / CEDRIC (MSDMA), CNAM

28–10–2022

# Plan

- Méthodes non-paramétriques pour des données de survie
  - Méthodes à noyaux
  - Zoom sur le modèle LLRER

- Méthodes paramétriques pour des données fonctionnelles
  - Régression logistique sous CBS
  - Extension du modèle Bliss

# Contexte

Soit

- $T$ une variable aléatoire réelle (v.a.r) d'intérêt
- $X$ sa co-variable associée de densité $f_X$
- Le couple $(X, T)$ admet une densité $f_{X,T}$ par rapport à la mesure de Lebesgue
- L'équation de régression entre $T$ et $X$ est :

$$T = m(X) + \varepsilon \tag{1}$$

Paramétrique : $m(\cdot)$ possède une forme connue avec des paramètres inconnus.

Non paramétrique : Aucune hypothèse sur la forme de $m(\cdot)$ n'est faite.

# Régression classique de N-W

Le critère utilisé pour estimer la fonction de régression est le critère des moindres carrées donné par:

$$\arg\min_{m} \mathbb{E}[(T - m(X))^2 | X]. \tag{2}$$

La fonction qui réalise pour tout $x$ fixé la meilleure approximation de $T$ sachant $X = x$, au sens des moindres carrées est donnée par :

$$m_{CR}(x) = \mathbb{E}[T|X = x] = \frac{\int t \; f_{X,T}(x,t)dt}{f_X(x)}.$$

On considère un $n$-échantillon $(X_i, T_i)$ de v.a. indépendantes de même loi (i.i.d) que le couple $(X, T)$. Le problème de minimisation devient :

$$\arg\min_m \sum_{1\leq i\leq n} (T_i - m)^2 K\left(\frac{X_i - x}{h_n}\right)$$

où $K$ est une densité de probabilité (appelée : noyau) et $h_n$ est une suite positive qui converge vers 0 à l'infini (appelée : fenêtre).
Un calcul algébrique nous donne le célèbre estimateur de Nadaraya et Watson (1964) donné par :

$$\widehat{m}_{NW}(x) := \frac{\sum_{1\leq i\leq n} T_i K\left(\frac{X_i - x}{h_n}\right)}{\sum_{1\leq i\leq n} K\left(\frac{X_i - x}{h_n}\right)}.$$

---

E. A. Nadaraya. On estimating regression. *Theor. Probab. Appl.*, 9 :141–142, 1964.

G. S. Watson. Smooth regression analysis. Sankhyà, 26 :359–372, 1964.

# Régression locale linéaire

L'idée principale de l'estimation de la fonction de régression par la méthode locale linéaire (LL) est la suivante :

- Un développement de Taylor à l'ordre 1 au voisinage du point $x$ donne :

$$m(X) \approx m(x) + m'(x)(X - x) =: \alpha + \beta(X - x)$$

- Estimer $m$ en trouvant $\alpha$ et $\beta$ qui minimise l'erreur quadratique moyenne pondérées par un noyau $K$.

---

J. Fan. Design adaptative nonparametric regression. J. of the American Statist. Association, 87 : 998–1004, 1992.

Le problème de minimisation devient :

$$\arg \min_{(\alpha,\beta)\in\mathbb{R}^2} \sum_{1\leq i\leq n} (T_i - \alpha - \beta(X_i - x))^2 K\left(\frac{X_i - x}{h_n}\right).$$

Un calcul algébrique donne :

$$\widehat{m}_{LL}(x) =: \frac{\sum_{1\leq i,j\leq n} T_i v_{i,j}(x)}{\sum_{1\leq i,j\leq n} v_{i,j}(x)}, \tag{3}$$

où

$$v_{i,j}(x) = (X_i - x)((X_i - x) - (X_j - x)) K\left(\frac{X_i - x}{h_n}\right) K\left(\frac{X_j - x}{h_n}\right). \tag{4}$$

# Régression relative

On considère la fonction de perte suivante :

$$\arg\min_{m} \mathbb{E}\left[\left(\frac{T - m(X)}{T}\right)^2 | X\right], \qquad \text{pour } T > 0.$$

La solution de ce problème pour tout $x \in \mathbb{R}$ est donnée par :

$$m_{RER}(x) = \frac{\mathbb{E}\left[T^{-1}|X = x\right]}{\mathbb{E}\left[T^{-2}|X = x\right]} =: \frac{m_1(x)}{m_2(x)} \leq m_{CR}(x).$$

Les deux premiers moments inverses conditionnels existent et soient finies. Notons par $\mu_\ell(\cdot) = m_\ell(\cdot) f_X(\cdot) = \int t^{-\ell} f_{X,T}(\cdot, t)dt$ pour $\ell = 1, 2$.

---

H. Park and L. A. Stefanski. Relative error prediction. Statist. & Probab. Lett., 40 : 227–236, 1998.

Le problème de minimisation (5) devient :

$$\arg\min_{m} \sum_{1\leq i\leq n} \left(\frac{T_i - m}{T_i}\right)^2 K\left(\frac{X_i - x}{h_n}\right),$$

ce qui conduit à

$$\widehat{m}_{RER}(x) := \frac{\sum_{1\leq i\leq n} T_i^{-1} K\left(\frac{x - X_i}{h_n}\right)}{\sum_{1\leq i\leq n} T_i^{-2} K\left(\frac{x - X_i}{h_n}\right)}.$$

Cet estimateur a été défini par Jones et al. (2008).

---

M. C. Jones, H. Park, K. I. Shin, S. K. Vines, and S. O. Jeong. Relative error prediction via kernel regression smoothers. *Journal of Statist. Plann. and Infer.*, 138 :2887–2898, 2008.

# Modèle LLRER

L'estimateur de la fonction de régression relative est solution de :

$$\arg \min_{(\alpha,\beta)\in\mathbb{R}^2} \sum_{1\leq i\leq n} \left( \frac{T_i - \alpha - \beta(X_i - x)}{T_i} \right)^2 K\left( \frac{X_i - x}{h_n} \right) \qquad (5)$$

où $K$ est une densité de probabilité (appelée : noyau) et $h_n$ est une suite positive qui converge vers 0 à l'infini (appelée : fenêtre).

# Modèle LLRER

La solution à ce problème de minimisation est donnée par :

$$m_{LLRER}(x) =: \frac{\mu_1(x)}{\mu_2(x)},$$

avec

$$\mu_\ell(x) = \frac{1}{(nh_n)^2} \sum_{1 \le i,j \le n} w_{i,j}^{\ell}(x), \qquad \textit{pour } \ell = 1, 2,$$

où

$$\begin{aligned} w_{i,j}^{\ell}(x) &= (X_i - x)^2 K\left(\frac{X_i - x}{h_n}\right) K\left(\frac{X_j - x}{h_n}\right) T_i^{-2} T_j^{-\ell} \\ &- (X_i - x)(X_j - x) K\left(\frac{X_i - x}{h_n}\right) K\left(\frac{X_j - x}{h_n}\right) T_i^{-2} T_j^{-\ell}. \end{aligned}$$

# Analyse des données de survie

- On s'intéresse au temps d'apparition d'un événement. Application : fiabilité, économie, finance, médecine ...
- L'événement ne s'est pas nécessairement produit à la fin de la période d'étude !
- Les données incomplètes ont une incidence sur les estimations.
- Types de données incomplètes : Troncature et Censure .

# Types de données incomplètes

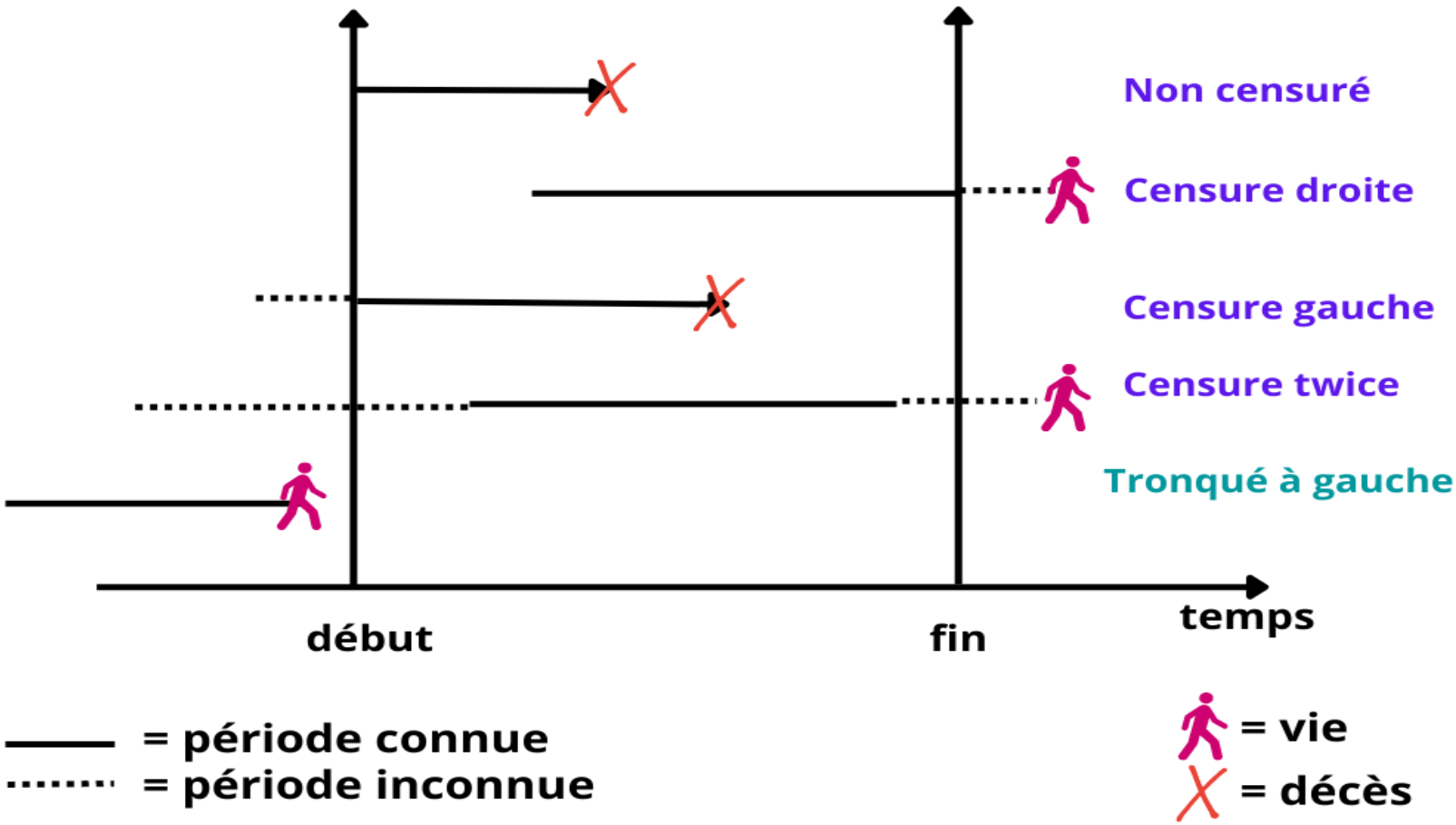
- Troncature : La variable d'intérêt $T$ doit être supérieure à une certaine variable de troncature $Y$ pour pouvoir être observée $(T \geq Y)$.

  Exemple : L'étude de la durée de vie après la retraite.

- Censure : Soit $C$ une variable de censure. Dans le cas d'une censure à droite ( resp. censure à gauche), on n'observe que le $T \wedge C$ (resp. $T \vee C$). La variable d'intérêt $T$ doit être supérieure (resp. inférieure) à une certaine variable $C$ pour pouvoir observée $T \geq C$ (resp. $T \leq C$).

  Exemple : L'étude de la durée de vie des habitants d'une petite ville suivies pour un traitement contre l'obésité.

# Types de censure




Construire des estimateurs robustes aux outliers, effets de bords et données incomplètes.

# Modèle de censure à droite

On introduit les variables de censure $C_1, \cdots, C_n$ indépendantes et de même loi que $C$ de f.d.r. $G$. Dans le modèle de censure, nos observations sont :

$$\begin{cases} Y_i = T_i \wedge C_i & 1 \leq i \leq n, \\ \delta_i = \mathbb{1}_{\{T_i \leq C_i\}} & 1 \leq i \leq n, \\ X_i \in \mathbb{R} & 1 \leq i \leq n. \end{cases} \tag{6}$$

## Hypothèse

*Nous supposons que $(X_i, T_i)_i$ et $C_i$ sont indépendants.*

# Modèle de censure à droite

Données synthétiques : une idée de Carbonez et al.(1995)

$$\widetilde{T}_i^{-\ell} = \frac{\delta_i Y_i^{-\ell}}{\overline{G}(Y_i)}, \tag{7}$$

pour $\ell = 1, 2$ où $\overline{G}(\cdot) = 1 - G(\cdot)$ est la fonction de survie de la variable de censure $C$. En utilisant la propriété de l'espérance conditionnelle, nous avons :

$$\begin{aligned}
\mathbb{E}\left[\widetilde{T}_1^{-\ell} | X_1 = x\right] &= \mathbb{E}\left[\frac{\delta_1 Y_1^{-\ell}}{\overline{G}(Y_1)} \Big| X_1 = x\right] \\
&= \mathbb{E}\left[\frac{T_1^{-\ell}}{\overline{G}(T_1)} \mathbb{E}\left[\mathbb{1}_{\{T_1 \leq C_1\}} | T_1\right] | X_1 = x\right] \\
&= \mathbb{E}\left[T_1^{-\ell} | X_1 = x\right] = m_\ell(x).
\end{aligned}$$

---

A. Carbonez, L. Gyorfi, E.C. Van Der Meulen. Partitioning estimates of a regression function under random censoring. *Statist. and Decisions.* 76, 1335–1344, 1995.

# Plan

Méthodes non-paramétriques pour des données de survie
- Modèles de régression
- Modèles de survie

Zoom sur le modèle LLRER pour des données censurées à droite
- Estimateur
- Hypothèses et résultat principal
- Étude numérique
  - Performance de l'estimateur
  - Comparaison avec d'autres méthodes
  - Exemple sur un jeu de données sur la transplantation rénale

Méthodes paramétriques pour des données fonctionnelles
- Extension du modèle Bliss
  - Modèle bayésien
  - Distributions à priori
  - Distributions à posteriori
  - Application sur un jeu de donnée viticole

# Pseudo-Estimateur

Le problème de minimisation (5) devient :

$$\arg \min_{(\alpha,\beta)\in\mathbb{R}^2} \sum_{1\leq i\leq n} \widetilde{T}_i^{-2} \left(Y_i - \alpha - \beta(X_i - x)\right)^2 K\left(\frac{X_i - x}{h_n}\right).$$

Un calcul algébrique donne :

$$\widetilde{m}_{LLRER}(x) =: \frac{\widetilde{\mu}_1(x)}{\widetilde{\mu}_2(x)}, \tag{8}$$

avec

$$\widetilde{\mu}_\ell(x) = \frac{1}{(nh_n)^2} \sum_{1\leq i,j\leq n} \widetilde{w}_{i,j}^{\ell}(x), \qquad \textit{pour } \ell = 1, 2,$$

où

$$\begin{aligned} \widetilde{w}_{i,j}^{\ell}(x) &= (X_i - x)^2 K\left(\frac{X_i - x}{h_n}\right) K\left(\frac{X_j - x}{h_n}\right) \widetilde{T}_i^{-2} \widetilde{T}_j^{-\ell} \\ &- (X_i - x)(X_j - x) K\left(\frac{X_i - x}{h_n}\right) K\left(\frac{X_j - x}{h_n}\right) \widetilde{T}_i^{-2} \widetilde{T}_j^{-\ell}. \end{aligned} \tag{9}$$

# Estimateur de Kaplan Meier

Pour obtenir un estimateur quantifiable, nous remplaçons la fonction de survie $\overline{G}(\cdot)$ par l'estimateur de Kaplan et Meier (1958) défini par :

$$\overline{G}_n(t) := \begin{cases} \prod_{i=1}^{n} \left(1 - \frac{1-\delta_{(i)}}{n-i+1}\right)^{\mathbb{1}_{\{Y_{(i)} \leq t\}}} & \text{si} \quad t < Y_{(n)}, \\ 0 & \text{sinon} \end{cases} \tag{10}$$

où $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$ sont les statistiques d'ordre des $Y_i$ et $\delta_{(i)}$ sont les concomitants de $Y_i$.

---

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *J. Amer. Stat. Assoc.*, 53 :458–481, 1958.

# Estimateur LLRER

Données synthétiques calculables :

$$\widehat{T}_i^{-\ell} = \frac{\delta_i Y_i^{-\ell}}{\overline{G}_n(Y_i)}, \qquad \text{pour} \quad 1 \leq i \leq n. \tag{11}$$

En substituant (11) dans (9), on obtient :

$$\widehat{m}_{LLRER}(x) = \frac{\widehat{\mu}_1(x)}{\widehat{\mu}_2(x)}, \tag{12}$$

avec

$$\widehat{\mu}_\ell(x) = \frac{1}{(nh_n)^2} \sum_{1 \leq i,j \leq n} \widehat{w}_{i,j}^{\ell}(x), \qquad \textit{pour} \;\; \ell = 1, 2,$$

où

$$\begin{aligned} \widehat{w}_{i,j}^{\ell}(x) &= (X_i - x)^2 K\left(\frac{X_i - x}{h_n}\right) K\left(\frac{X_j - x}{h_n}\right) \widehat{T}_i^{-2} \widehat{T}_j^{-\ell} \\ &- (X_i - x)(X_j - x) K\left(\frac{X_i - x}{h_n}\right) K\left(\frac{X_j - x}{h_n}\right) \widehat{T}_i^{-2} \widehat{T}_j^{-\ell}. \end{aligned}$$

## Hypothèse

*Si nous supposons que $T_i$ et $C_i$ sont indépendants conditionnellement à $X_i$.*

Les données synthétiques deviennent :

$$\widehat{T}_i^{-\ell} = \frac{\delta_i Y_i^{-\ell}}{\overline{G}_n(Y_i|X_i)},$$

pour $\ell = 1, 2$ où $\overline{G}_n$ est l'estimateur de $\mathrm{Beran}$ de la fonction de survie conditionnelle de la variable de censure $C$ sachant $X$.

---

R. Beran. Nonparametric regression with randomly censored survival data. Technical report, Department of Statistics, University of California, Berkeley., 1981.

## Remarque

*Si $\beta = 0$, nous retombons sur l'estimateur de la fonction de régression relative (RER pour "relative error regression") par la méthode à noyau défini dans Khardani et Slaoui (2019) par :*

$$\widehat{m}_{RER}(x) = \frac{\sum\limits_{1\leq i\leq n} \widehat{T}_i^{-1} K\left(\frac{X_i - x}{h_n}\right)}{\sum\limits_{1\leq i\leq n} \widehat{T}_i^{-2} K\left(\frac{X_i - x}{h_n}\right)}, \tag{13}$$


S. Khardani and Y. Slaoui. Nonparametric relative regression under random censorship model. *Probab. and Statist. Letters*, 151:116–122, 2019.

F. Bouhadjera, E. Ould Saïd. M.R. Remita. On the strong uniform consistency for relative error of the regression function estimator for censoring times series model. *Communication in Statistics : Theory and Methods*.

F. Bouhadjera & E. Ould Saïd. Asymptotic normality of the relative error regression function estimator for censored time series data. *Dependence Modeling*.

## Remarque

*L'estimateur de la fonction de régression classique par la méthode linéaire locale (LLR pour "local linear regression") défini dans Bouhadjera et al. (2021) par :*

$$\widehat{m}_{LLR}(x) = \frac{\sum\limits_{1\leq i,j\leq n} v_{i,j}(x)\,\widehat{T}_j}{\sum\limits_{1\leq i,j\leq n} v_{i,j}(x)}, \tag{14}$$

*où*

$$v_{i,j}(x) = (X_i - x)\left((X_i - x) - (X_j - x)\right) K\left(\frac{X_i - x}{h_n}\right) K\left(\frac{X_j - x}{h_n}\right).$$

---

F. Bouhadjera, E. Ould Saïd & R. M. Remita. Strong consistency of the nonparametric local linear regression estimation under censorship model. *Communication in Statist. : Theory and Methods.*

O. Benrabah, F. Bouhadjera & E. Ould Saïd. Local linear estimation of the regression function for twice censored data. *Statistical Papers.*

## Remarque

*L'estimateur de la fonction de régression classique par la méthode à noyau (CR pour "classical regression") défini dans Guessoum et Ould Saïd (2008) est donné par :*

$$\widehat{m}_{CR}(x) = \frac{\sum_{1\leq i\leq n} \widehat{T}_i K\left(\frac{x - X_i}{h_n}\right)}{\sum_{1\leq i\leq n} K\left(\frac{x - X_i}{h_n}\right)}. \tag{15}$$

Z. Guessoum and E. Ould Saïd. On nonparametric estimation of the regression function under random censorship model. *Statist. and Decisions,* 26 : 1001–1020, 2008.

Soit $\mathcal{C}_0 = \{x \in \mathbb{R} / f_X(x) > 0\}$ et $\mathcal{C}$ un sous ensemble compact de $\mathcal{C}_0$.

## Hypothèses

**(A1)** *La fenêtre* $h_n$ *satisfait* $\lim_{n\to\infty} \dfrac{nh_n^2}{\log n} = +\infty$ *et* $\lim_{n\to\infty} \dfrac{nh_n^6}{\log n} = 0$.

**(A2)** *Le noyau* $K$ *est une fonction de densité non-négative et symétrique. De plus, pour* $\gamma = 2, 3$

- **i.** $\displaystyle\int |t|^\gamma K(t)dt < \infty,$
- **ii.** $\displaystyle\int |t|^\gamma K^2(t)dt < \infty.$

**(A3)** *La fonction de densité* $f_X(\cdot)$ *est continûment différentiable.*

**(A4)** *La fonction* $\mu_\ell(\cdot)$, *pour* $\ell = 1, 2$, *est continûment différentiable.*

**(A5)** *La fonction* $\displaystyle\int \frac{t^{-\ell}}{\overline{G}(t)} f_{X,T}(x, t)dt$, *pour* $\ell = 2, 3, 4$, *est continûment différentiable.*

# Convergence uniforme presque sûre (p.s.)

Théorème

Sous les hypothèses (A1)–(A5), pour $n$ assez grand, on a :

$$\sup_{x\in\mathcal{C}} |\widehat{m}_{LLRER}(x) - m_{RER}(x)| = \mathrm{O}\left(h_n^2\right) + \mathrm{O}_{p.s.}\left(\sqrt{\frac{\log n}{nh_n^2}}\right).$$

F. Bouhadjera and E. Ould Saïd. Strong consistency of the local linear relative error regression for censored data. *Opuscula Mathematica*, 42(6), 2022

La preuve du théorème est basée sur la décomposition suivante :

$$\begin{aligned}
\widehat{m}_{LLRER}(x) - m_{RER}(x) &= \frac{1}{\widehat{\mu}_2(x)}\Big\{\widehat{\mu}_1(x) - \widetilde{\mu}_1(x) + \widetilde{\mu}_1(x) - \mathbb{E}[\widetilde{\mu}_1(x)] \\
&+ \mathbb{E}[\widetilde{\mu}_1(x)] - m_1(x)m_2(x) \\
&+ m_{RER}(x)\Big(m_2^2(x) - \mathbb{E}[\widetilde{\mu}_2(x)] \\
&+ \mathbb{E}[\widetilde{\mu}_2(x)] - \widetilde{\mu}_2(x) + \widetilde{\mu}_2(x) - \widehat{\mu}_2(x)\Big)\Big\}.
\end{aligned}$$

$$\begin{aligned}
\widehat{\mu}_\ell(x) - \widetilde{\mu}_\ell(x) &= \frac{1}{(nh_n)^2} \sum_{1\le i,j\le n} \left(\widehat{w}_{i,j}^{\ell}(x) - \widetilde{w}_{i,j}^{\ell}(x)\right) \\
&= \widehat{S}_{2,2}(x)\widehat{S}_{\ell,0}(x) - \widetilde{S}_{2,2}(x)\widetilde{S}_{\ell,0}(x) \\
&- \left(\widehat{S}_{2,1}(x)\widehat{S}_{\ell,1}(x) - \widetilde{S}_{2,1}(x)\widetilde{S}_{\ell,1}(x)\right) \\
&=: \mathcal{B}_{\ell,1}(x) - \mathcal{B}_{\ell,2}(x),
\end{aligned}$$

où pour $\ell = 1, 2$ et $\gamma = 0, \ell$

$$\begin{aligned}
\widehat{S}_{\ell,\gamma}(x) &:= \frac{1}{nh_n} \sum_{i=1}^{n} \widehat{T}_i^{-\ell} (X_i - x)^\gamma K\left(\frac{X_i - x}{h_n}\right), \\
\widetilde{S}_{\ell,\gamma}(x) &:= \frac{1}{nh_n} \sum_{i=1}^{n} \widetilde{T}_i^{-\ell} (X_i - x)^\gamma K\left(\frac{X_i - x}{h_n}\right).
\end{aligned}$$

D'un coté, pour $\ell = 1, 2$, on a

$$\begin{aligned}
\mathcal{B}_{\ell,1}(x) &= \left(\widehat{S}_{2,2}(x) - \widetilde{S}_{2,2}(x)\right)\left(\widehat{S}_{\ell,0}(x) - \widetilde{S}_{\ell,0}(x)\right) \\
&+ \left(\widetilde{S}_{\ell,0}(x) - \mathbb{E}[\widetilde{S}_{\ell,0}(x)]\right)\left(\widehat{S}_{2,2}(x) - \widetilde{S}_{2,2}(x)\right) \\
&+ \mathbb{E}[\widetilde{S}_{\ell,0}(x)]\left(\widehat{S}_{2,2}(x) - \widetilde{S}_{2,2}(x)\right) \\
&+ \left(\widetilde{S}_{2,2}(x) - \mathbb{E}[\widetilde{S}_{2,2}(x)]\right)\left(\widehat{S}_{\ell,0}(x) - \widetilde{S}_{\ell,0}(x)\right) \\
&+ \mathbb{E}[\widetilde{S}_{2,2}(x)]\left(\widehat{S}_{\ell,0}(x) - \widetilde{S}_{\ell,0}(x)\right).
\end{aligned}$$

D'un autre coté, pour $\ell = 1, 2$, on a

$$\begin{aligned}
\mathcal{B}_{\ell,2}(x) &= \left(\widehat{S}_{2,1}(x) - \widetilde{S}_{2,1}(x)\right)\left(\widehat{S}_{\ell,1}(x) - \widetilde{S}_{\ell,1}(x)\right) \\
&+ \left(\widetilde{S}_{2,1}(x) - \mathbb{E}[\widetilde{S}_{2,1}(x)]\right)\left(\widehat{S}_{\ell,1}(x) - \widetilde{S}_{\ell,1}(x)\right) \\
&+ \mathbb{E}[\widetilde{S}_{2,1}(x)]\left(\widehat{S}_{\ell,1}(x) - \widetilde{S}_{\ell,1}(x)\right) \\
&+ \left(\widetilde{S}_{\ell,1}(x) - \mathbb{E}[\widetilde{S}_{\ell,1}(x)]\right)\left(\widehat{S}_{2,1}(x) - \widetilde{S}_{2,1}(x)\right) \\
&+ \mathbb{E}[\widetilde{S}_{\ell,1}(x)]\left(\widehat{S}_{2,1}(x) - \widetilde{S}_{2,1}(x)\right).
\end{aligned}$$

## Proposition

*Sous les hypothèses (A1), (A2) i), (A3) et (A4), pour* $\ell = 1, 2$ *et* $n$ *assez grand, nous avons :*

$$\sup_{x \in \mathcal{C}} |\widehat{\mu}_\ell(x) - \widetilde{\mu}_\ell(x)| = O_{p.s.}\left(\sqrt{\frac{\log \log n}{n}}\right).$$

## Proposition

*Sous les hypothèses (A1), (A2) i), (A3)–(A5), pour* $\ell = 1, 2$ *et* $n$ *assez grand, nous avons :*

$$\sup_{x \in \mathcal{C}} \left| \widetilde{\mu}_\ell(x) - \mathbb{E}[\widetilde{\mu}_\ell(x)] \right| = O_{p.s.} \left( \sqrt{\frac{\log n}{n h_n^2}} \right).$$

## Proposition

*Sous les hypothèses (A1), (A2) et (A4), pour $\ell = 1, 2$ et $n$ assez grand, nous avons :*

$$\sup_{x \in \mathcal{C}} |\mathbb{E}[\widetilde{\mu}_\ell(x)] - m_\ell(x) m_2(x)| = O\left(h_n^2\right).$$

## Corollaire

*Sous les hypothèses (A1)–(A5), il existe un nombre réel $C > 0$ tel que :*

$$\sum_{n=1}^{\infty} \mathbb{P}\left( \inf_{x \in \mathcal{C}} |\widehat{\mu}_2(x)| \leq C \right) < \infty.$$

# Plan

Méthodes non-paramétriques pour des données de survie
- Modèles de régression
- Modèles de survie

Zoom sur le modèle LLRER pour des données censurées à droite
- Estimateur
- Hypothèses et résultat principal
- Étude numérique
  - Performance de l'estimateur
  - Comparaison avec d'autres méthodes
  - Exemple sur un jeu de données sur la transplantation rénale

Méthodes paramétriques pour des données fonctionnelles
- Extension du modèle Bliss
  - Modèle bayésien
  - Distributions à priori
  - Distributions à posteriori
  - Application sur un jeu de donnée viticole

# Algorithme

Entrées : Générer $n$ v.a. i.i.d. $\{X_i \rightsquigarrow \mathcal{W}(1,1),\ C_i \rightsquigarrow \exp(\lambda)$ et $\varepsilon_i \rightsquigarrow \mathcal{N}(0, 0.2)\}$ pour $1 \leq i \leq n$ où $\lambda$ est une constante qui permet d'ajuster le pourcentage de censure (C.P.).

Étape 1 : Calculer la variable d'intérêt selon les modèles suivants :

$$\begin{aligned}
\text{Modèle 0 (M0):} \quad T_i &= 2X_i + 1 + \varepsilon_i, \\
\text{Modèle 1 (M1):} \quad T_i &= 3 + \sin(X_i) + \varepsilon_i, \\
\text{Modèle 2 (M2):} \quad T_i &= 0.9X_i^2 + \frac{5}{2} + \varepsilon_i.
\end{aligned}$$

pour $1 \leq i \leq n$.

Étape 2 : Déterminer les observées $Y_i$ et l'indicatrice de non censure $\delta_i$ selon le modèle (6).

# Algorithme (suite)

Étape 2′ : Pour $1 \leq j \leq \frac{n}{20}$, nous considérons $i = 20 \times j$ et nous multiplions $Y_i$ par un constante multiplicative (M.C.) que nous varions pour créer l'effet des valeurs aberrantes.

Étape 3 : Estimer les valeurs de $\overline{G}(\cdot)$.

Étape 4 : Déterminer les données synthétiques calculables $\{\widehat{T}_i^{-\ell}, 1 \leq i \leq n\}$ pour $\ell = -1, 1, 2$.

Étape 5 : Le noyau $K$ est une gaussienne standard et la fenêtre optimale est sélectionnée par validation croisée sur une séquence $[0.01, 2]$ avec un pas de 0.01.

Sorties : Compiler les estimateurs CR, LLR, RER et LLRER donnés par les formules (15), (14), (13) et (12) pour $h_{opt}$ et $x \in [1, 4]$.

# Méthode de la validation croisée

La méthode la plus utilisée pour déterminer la fenêtre optimale est la méthode de validation croisée. L'idée principale est de minimiser les critères suivants :

## Régression classique

$$CV_{CR} = \frac{1}{n-1} \sum_{1 \leq i \leq n} \left( Y_i - \widehat{m}_{CR}^{-i}(X_i) \right)^2,$$

$$CV_{LLR} = \frac{1}{n-1} \sum_{1 \leq i \leq n} \left( Y_i - \widehat{m}_{LLR}^{-i}(X_i) \right)^2,$$

## Régression relative

$$CV_{RER} = \frac{1}{n-1} \sum_{1 \leq i \leq n} \left( \frac{Y_i - \widehat{m}_{RER}^{-i}(X_i)}{Y_i} \right)^2,$$

$$CV_{LLRER} = \frac{1}{n-1} \sum_{1 \leq i \leq n} \left( \frac{Y_i - \widehat{m}_{LLRER}^{-i}(X_i)}{Y_i} \right)^2,$$

où $\widehat{m}_{CR}^{-i}(\cdot)$,$\widehat{m}_{LLR}^{-i}(\cdot)$, $\widehat{m}_{RER}^{-i}(\cdot)$ et $\widehat{m}_{LLRER}^{-i}(\cdot)$ sont respectivement les estimateurs CR, LLR, RER et LLRER en supprimant le $i$ ème triplet d'observations $(X_i, Y_i, \delta_i)$.

# Performance du LLRER pour **(M0).**

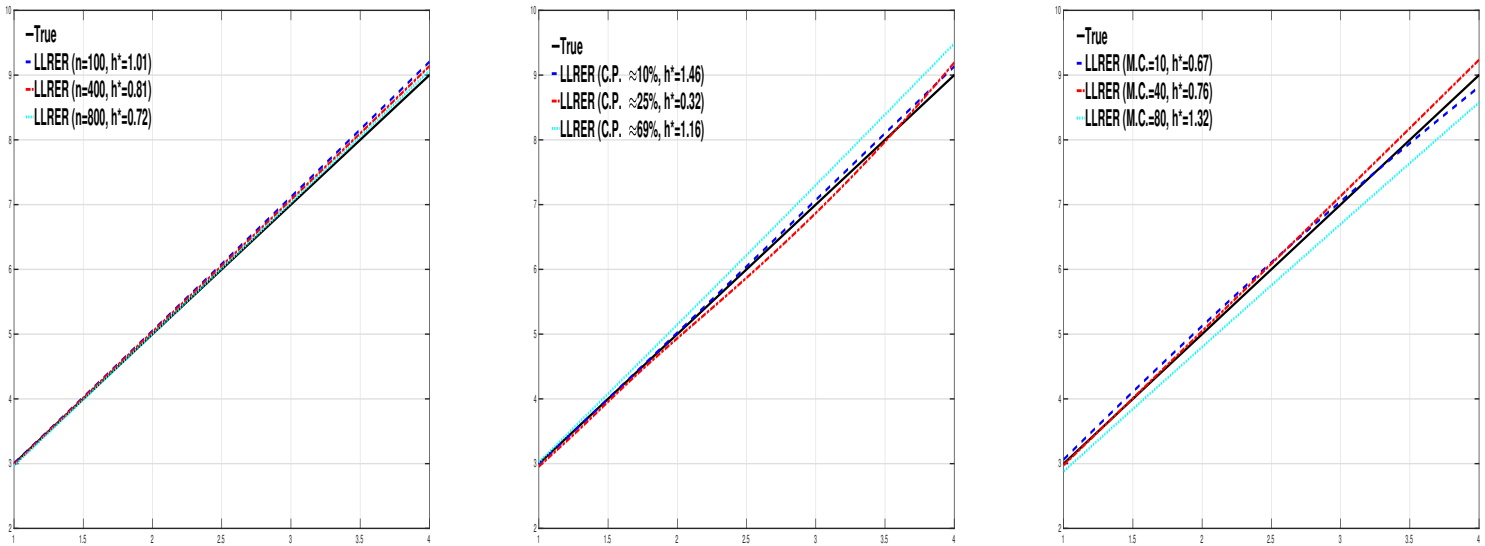
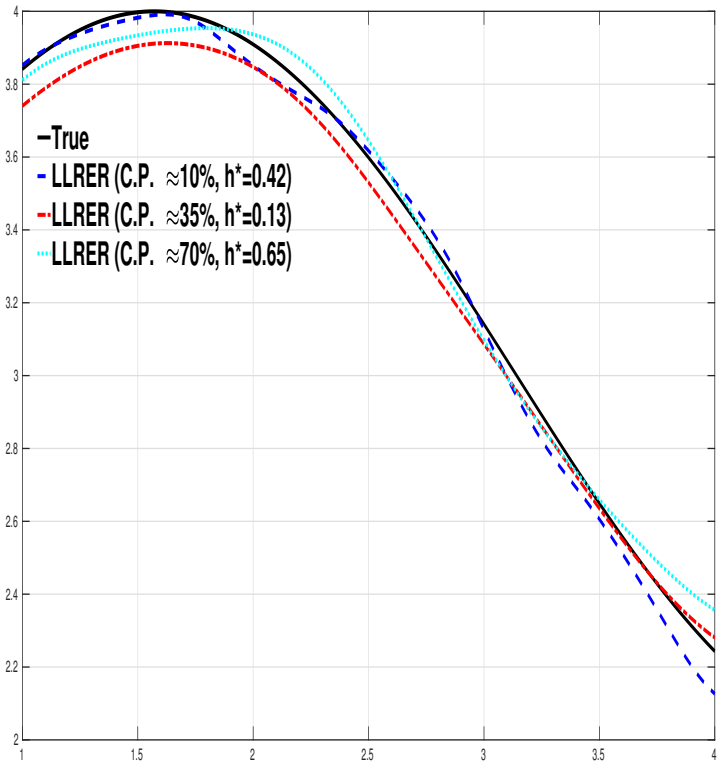



Figure: Effet de la taille d'échantillon (C.P. $\approx 25\%$), du taux de censure (pour $n = 400$) et des effets de valeurs aberrantes (pour $n = 400$ et C.P. $\approx 60\%$ ).

# Performance du LLRER pour **(M1)**



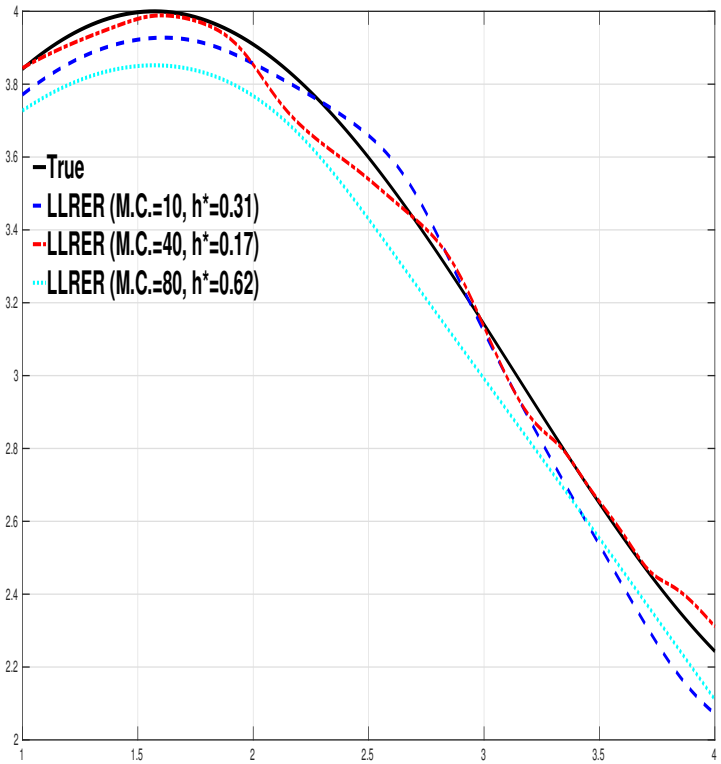




Figure: Effet de le censure (pour $n = 400$) et effet des outliers (pour $n = 400$ et C.P.$\approx$ 34%).

# Comparaison avec d'autres méthodes

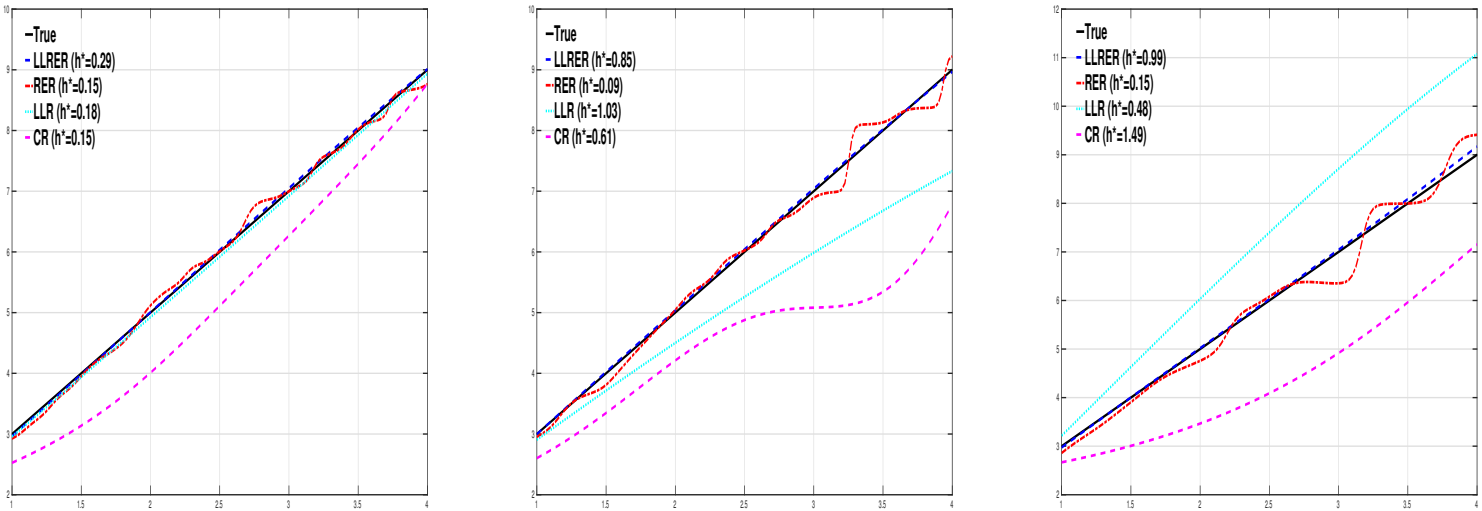



Figure: Comparaison des modèles CR, LLR, RER et LLRER en terme d'effet de la censure pour le modèle **(M0)** avec $n = 400$ et C.P.$\approx$ 13, 40 et 69% respectivement.

# Comparaison avec d'autres méthodes

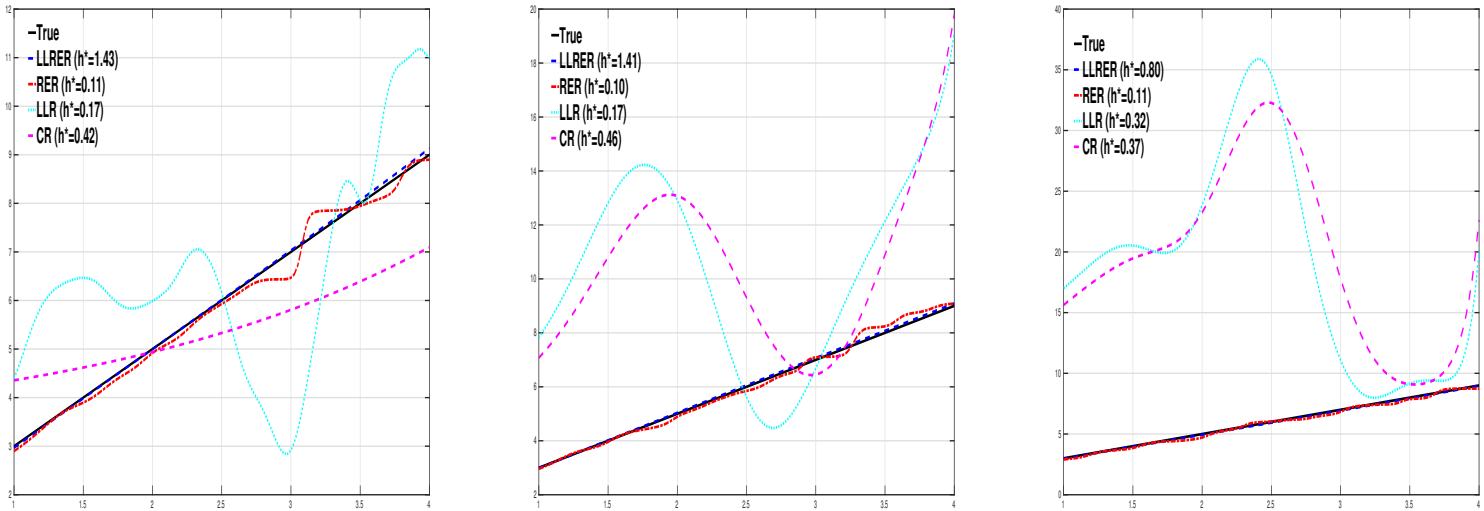
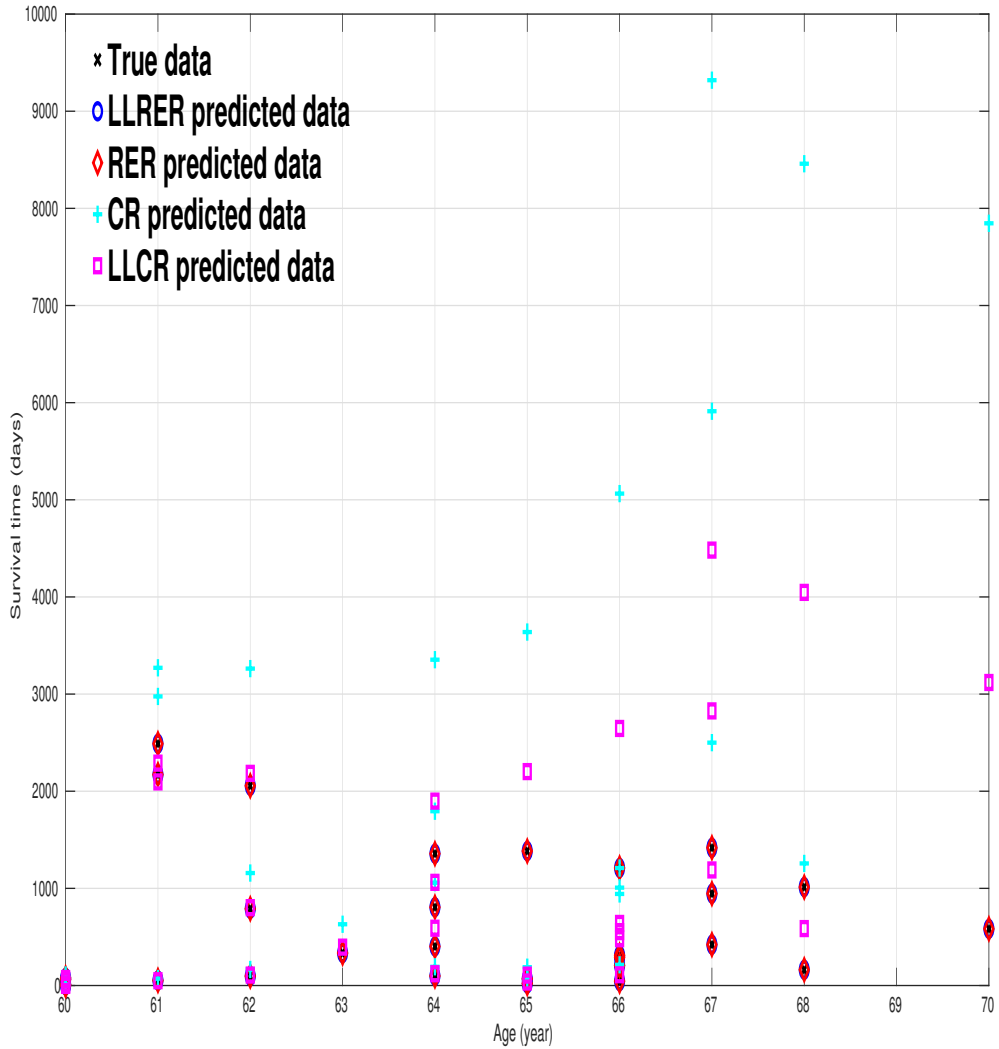



Figure: Comparaison des estimateurs CR, LLR, RER et LLRER en terme de valeurs aberrantes pour le modèle **(M0)** avec $n = 400$, C.P. $\approx 42.5\%$ et M.C. $= 10, 40$ et $80$ respectivement.

# Exemple sur un jeu de données sur la transplantation rénale

- Test d'efficacité du prédicteur LLRER en présence de censure.
- Les données consistent en $n = 863$ temps de survie de patients qui ont subit une transplantation rénale.
- Lieu : *The Ohio State University Transplant Center* durant la période de $1982 - 1992$
- Les patients sont censurées si : ils déménagement (perte de vue) ou ils sont encore en vie le 30 juin 1992.
- Les covariables sont : le sexe, age (en année), à l'année de la transplantation et on sait quand est ce que un temps de survie est censuré ou non. Le taux de censure est de 84%.
- La variable temporelle la plus importante dans les modèles de survie est le temps depuis la transplantation. On prend comme covariable, l'age au moment de la transplantation.
- L'échantillon d'apprentissage est $n^\star = 763$ sélectionné aléatoirement. $(X_i, Y_i, \delta_i),\ i = 1, \ldots, n^\star$

Les données viennent du livre de Klein et Moshberger (2004)


True data
LLRER predicted data
RER predicted data
CR predicted data
LLCR predicted data
Survival time (days)
Age (year)
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
60
61
62
63
64
65
66
67
68
69
70

# Plan

Méthodes non-paramétriques pour des données de survie
- Modèles de régression
- Modèles de survie

Zoom sur le modèle LLRER pour des données censurées à droite
- Estimateur
- Hypothèses et résultat principal
- Étude numérique
  - Performance de l'estimateur
  - Comparaison avec d'autres méthodes
  - Exemple sur un jeu de données sur la transplantation rénale

Méthodes paramétriques pour des données fonctionnelles
- Extension du modèle Bliss
  - Modèle bayésien
  - Distributions à priori
  - Distributions à posteriori
  - Application sur un jeu de donnée viticole

# Contexte

- La variable d'interêt $Y$ prend ses valeurs dans $\{0, 1\}$
- Les $q$-covariables $X_1(t), \cdots, X_q(t)$ indexé par $(t_1, \cdots, t_q)$

$$Y = \beta_0 + \sum_{v=1}^{q} \int_{\mathcal{C}_v} \beta_v(t_v) X_v(t_v) dt_v + \epsilon$$

- Choice-Based-Sampling (CBS) : Supposer que la population est divisée en fonction des valeurs de la variable d'interêt $Y$.
  - Soit $Q(i) = \mathbb{P}[Y = i]$ avec $i \in \{0, 1\}$
  - Les deux stratas :

$$\begin{aligned} \mathcal{S}(0) &= \{(0, X), X \in \mathcal{H}\} \\ \mathcal{S}(1) &= \{(1, X), X \in \mathcal{H}\} \end{aligned}$$

  - Soit $0 < H(i) < 1$ la proba. de tirer de la strata $\mathcal{S}(i)$

- La densité conditionnelle de $Y$ sachant $X = x$ est définie par :

$$h(Y = i|X = x, \beta_0, \beta) = \frac{\mathbb{P}(Y = i|X = x, \beta_0, \beta)\frac{H(i)}{Q(i)}}{\sum_{l=0}^{1} \mathbb{P}(Y = l|X = x, \beta_0, \beta)\frac{H(l)}{Q(l)}}, \ x \in \mathcal{H}, i \in \{0, 1\}$$

- Soit un $N$-échantillon i.i.d. d'observations $(Y^{(n)} = i_n, \{X^{(n)}(\mathbf{t}), \mathbf{t} \in \mathbf{C}\})$ , $n = 1, \ldots, N$, de même loi que le couple $(Y, X)$. L'estimateur du max. de vraisemblance conditionnelle :

$$\begin{aligned} \boldsymbol{l}(\beta_0, \beta) &= \prod_{n=1}^{N} h(Y^{(n)}|X^{(n)}, \beta_0, \beta) \\ &= \prod_{n=1}^{N} \left[h(1|X^{(n)}, \beta_0, \beta))\right]^{Y^{(n)}} \left[1 - h(1|X^{(n)}, \beta_0, \beta)\right]^{1-Y^{(n)}} \end{aligned}$$

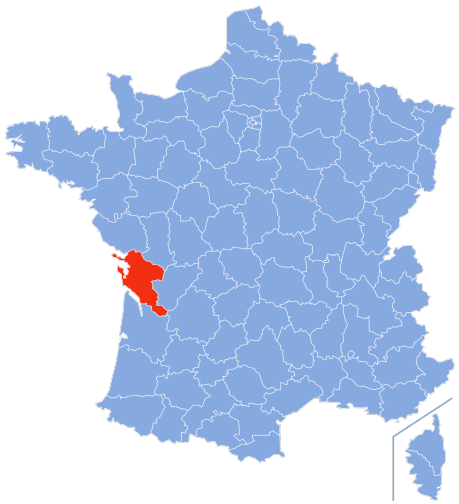
# Projet de post-doc

- Extension d'un modèle de régression fonctionnelle linéaire bayésienne aux variables scalaires et catégorielles.
- **La motivation de notre étude vient d'un projet sur le dépérissement viticole.**
- Une base unique obtenue du Bureau national interprofessionnel du Cognac (BNIC)
- Les objectifs sont d'être capable d'identifier :
  - Les facteurs et interactions de facteurs qui contribuent aux dépérissement d'une parcelle et la mortalité d'une grappe de vigne.
  - Les périodes de temps sur lesquelles ces facteurs ont un impact, sur le court et long terme.

# Données

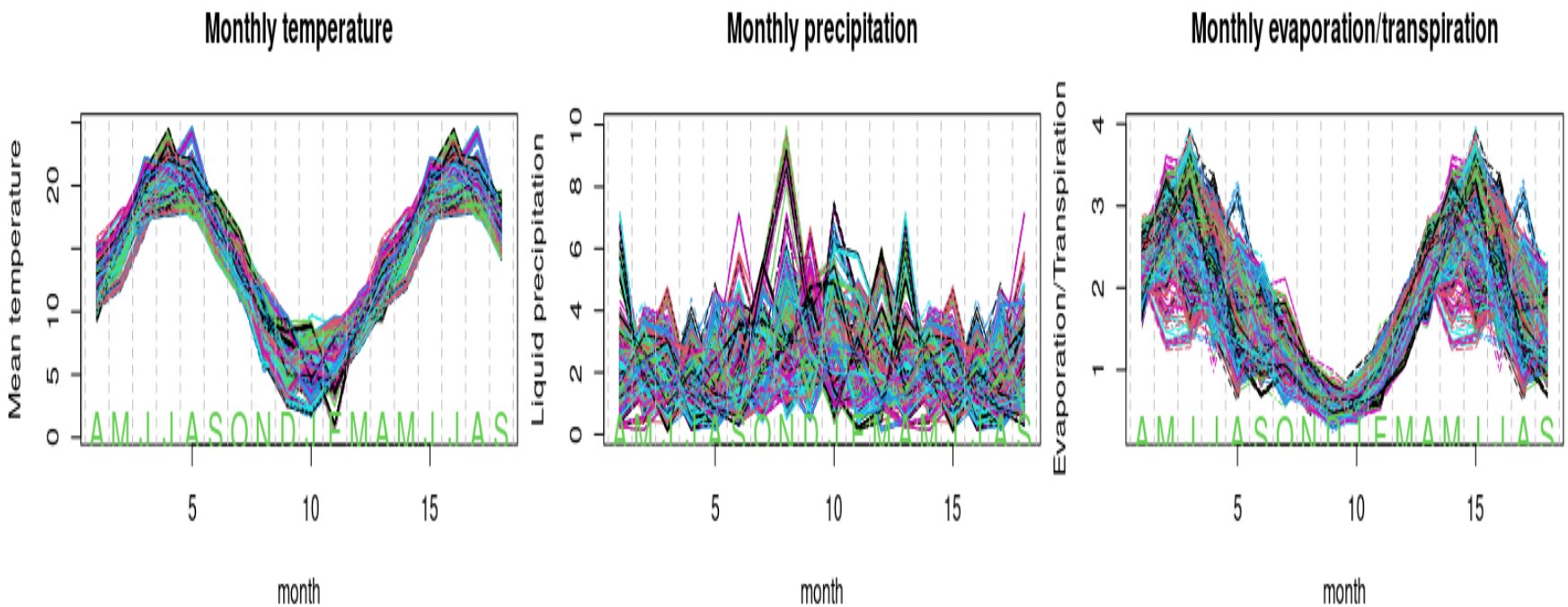
Jeu de données sur 55 parcelles de vigne suivies durant la période $1986 - -2018$.

- $Y$ : le rendement (v. d'intérêt réelle)
- $(X_v,\ v = 1, \cdots, q)$ : données climatiques (température min, max et moyenne, précipitation et évaporation-transpiration durant le cycle de vie de la vigne (covariables–fonctionnelles))



| année | id | rendement | | année | id | rendement |
|---|---|---|---|---|---|---|
| 1995 | P10G1 | 7.836 | | 1995 | P9G2 | 7.407 |
| 1996 | P10G1 | 6.548 | ... ... | 1996 | P9G2 | 7.774 |
| ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ | ⋮ |
| 2018 | P10G1 | 4.687 | | 2018 | P9G2 | 3.755 |

# Données climatiques




Nous avons une courbe par parcelle par année associée à un rendement.

# + Données

- Situation topographique
- Type de sol
- Sensibilité au froid, l'humidité, pourriture, ...
- Méthodes de plantation
- Précédent cultural
- Année d'entrée et de sortie d'une parcelle de la base

...

# Goal

- Estimation des $\beta_v(t),\ v = 1, \cdots, q$
- On cherche à identifier des périodes de temps (on cherche un modèle plus explicatif que prédictif). Les intervalles de temps où $\beta_v \neq 0$
- Une méthode bayésienne qui prend en compte les connaissances à priori.
- Considérer d'autres types de co-variables dans le modèle.

# Extension du modèle Bliss

- $(Z_1, \cdots, Z_p)$ un vecteur de scalaires de dimension $p$,
- $(X_1(t), \cdots, X_q(t))$ un vecteur de variables fonctionnelles de dimension on $\mathcal{C}$, un intervalle de $\mathbb{R}$,
- $\alpha$ une variable catégorielle de $r$ modalités

$$Y^{(ij)} = \beta_0 + \alpha_i + \sum_{s=1}^{p} \gamma_s Z_s + \sum_{v=1}^{q} \int_{\mathcal{C}} \beta_v(t) X_v^{(ij)}(t) dt + \epsilon^{(ij)},$$

où $\beta_0$ est l'intercepte, $(\alpha_1, \cdots, \alpha_r)$ et $(\gamma_1, \ldots, \gamma_p)$ sont des paramètres scalaires et $(\beta_1(t), \cdots, \beta_q(t))$ sont des paramètres fonctionnelles.
$\epsilon$ un bruit gaussien.

## Bayesian Functional Linear Regression with Sparse Step functions (Bliss)

$$\beta_v(t) = \sum_{k=1}^{K_v} \frac{b_{k,v}}{|\mathcal{I}_{k,v}|} \mathbb{1}_{\{t \in \mathcal{I}_{k,v}\}}, \qquad \text{for } v = 1, \cdots, q$$

où

- $\mathcal{I}_{1,v}, \cdots, \mathcal{I}_{K_v,v}$ sont de intervalles inclus dans $\mathcal{C}$ (Les intervalles $\mathcal{I}_{k,v}$ peuvent se chevaucher). De plus,

  $$\mathcal{I}_{k,v} = [m_{1,v} \pm \ell_{1,v}, \cdots, m_{K_v,v} \pm \ell_{K_v,v}]$$

  où pour un $v$ fixé, $m_{k,v}$ est le centre et $\ell_{k,v}$ sont les demi-longueurs des intervalles $\mathcal{I}_{k,v}$.
- $|\mathcal{I}_{k,v}|$ est la longueur de l'intervalle,
- $b_{1,v} \cdots, b_{K_v,v}$ sont de paramètres réelles.

---

Grollemund, P. M., Abraham, C., Baragatti, M. and Pudlo, P. (2019) Bayesian Functional Linear Regression with Sparse Step Functions. *Bayesian Analysis*, 14(1) :111–135.

# Modèle

$$Y^{(ij)} = \beta_0 + \alpha_i + \sum_{s=1}^{p} \gamma_s Z_s^{(ij)} + \sum_{v=1}^{q} \sum_{k=1}^{K_v} b_{k,v} X_v^{(ij)}(\mathcal{I}_{k,v}) + \epsilon^{(ij)},$$

avec

$$X_v^{(ij)}(\mathcal{I}_{k,v}) = \frac{1}{|\mathcal{I}_{k,v}|} \int_{\mathcal{I}_{k,v}} X_v^{(ij)}(t) dt,$$

Les paramètres à estimer sont :

- L'intercepte $\beta_0$
- La variance $\sigma^2$
- Paramètres scalaires : $\gamma_1, \cdots, \gamma_p$
- Paramètres catégorielles : $\alpha_1, \cdots, \alpha_r$
- Paramètres fonctionnelles : pour $v = 1, \cdots, q$
  - $b_v = (b_{1,v}, \cdots, b_{K_v,v})$ et $b = (b_1, \cdots, b_q)$,
  - $m_v = (m_{1,v}, \cdots, m_{K_v,v})$ et $m = (m_1, \cdots, m_q)$,
  - $\ell_v = (\ell_{1,v}, \cdots, \ell_{K_v,v})$, et $\ell = (\ell_1, \cdots, \ell_q)$,

# Paramètres du modèle

$$\theta = \left(\beta_0, \alpha_1, \cdots, \alpha_r, \gamma_1, \cdots, \gamma_p, \theta_1, \cdots, \theta_q, \sigma^2\right)$$

où pour $v = 1, \cdots, q$

$$\theta_v = \left(b_{1,v}, \cdots, b_{K_v,v}, m_{1,v}, \cdots, m_{K_v,v}, \ell_{1,v}, \cdots, \ell_{K_v,v}\right)$$

# Distributions à priori

$$
\begin{aligned}
\beta_0|\sigma^2 &\rightsquigarrow \mathcal{N}\left(0, u_0\sigma^2\right), \\
\alpha_i|\sigma^2 &\rightsquigarrow \mathcal{N}\left(0, v_0\sigma^2\right), \qquad \text{for } i = 1, \cdots, r \\
\gamma_s|\sigma^2 &\rightsquigarrow \mathcal{N}\left(0, w_0\sigma^2\right), \qquad \text{for } s = 1, \cdots, p \\
b_v|\sigma^2, m_v, \ell_v &\rightsquigarrow \mathcal{N}_{K_v}\left(0, n\sigma^2(G_v + \nu\lambda_{max}(G_v)I_{K_v})^{-1}\right), \\
\pi(\sigma^2) &\propto 1/\sigma^2, \\
m_v &\overset{i.i.d.}{\rightsquigarrow} \mathcal{U}\left(\mathcal{C}\right), \\
\ell_v &\overset{i.i.d.}{\rightsquigarrow} \exp\left(a \times |\mathcal{C}|\right),
\end{aligned}
$$

où $G_v = X_{\cdot,v}(\mathcal{I}_{\cdot,v})^t X_{\cdot,v}(\mathcal{I}_{\cdot,v})$ avec

$$
X_v^{\cdot}(\mathcal{I}_{\cdot,v}) = \left\{ X_v^{(ij)}(\mathcal{I}_{k,v}),\ \ 1 \leq i \leq r,\ 1 \leq j \leq n_i,\ 1 \leq k \leq K_v \right\}
$$

Les hyperparamètres sont $u_0, v_0, w_0, \nu, a, K$.

La full distribution à posteriori peut être écrit explicitement

## Distribution à posteriori

$$
\begin{aligned}
\beta|Y,\sigma^2,m,\ell &\rightsquigarrow \mathcal{N}_{K_v+p+r+1}\left((\underline{X}^t\underline{X}+\underline{V})^{-1}\underline{X}^tY,\sigma^2(\underline{X}^t\underline{X}+\underline{V})^{-1}\right),\\
\sigma^2|Y,\beta,m,\ell &\rightsquigarrow \Gamma^{-1}\left(\frac{n+K+r+p+1}{2},\frac{1}{2}\left\{\mathrm{RSS}+\beta^t\underline{V}\beta\right\}\right),\\
f(m_v|Y,\beta,\sigma^2,m_{-v},\ell) &\propto \exp\left\{-\frac{\mathrm{RSS}}{2\sigma^2}\right\}\times f(b_v|m_v,\ell_v,\sigma^2)\times f(m_v),\\
f(\ell_v|Y,\beta,\sigma^2,\ell_{-v},m) &\propto \exp\left\{-\frac{\mathrm{RSS}}{2\sigma^2}\right\}\times f(b_v|m_v,\ell_v,\sigma^2)\times f(\ell_v),
\end{aligned}
$$

où

- $\underline{X}=(A\mid Z_1,\cdots,Z_p\mid X_1^{\cdot}(\mathcal{I}_{\cdot,1}),\cdots,X_q^{\cdot}(\mathcal{I}_{\cdot,q}))$,
- $\beta^t=(\beta_0,\alpha_1,\cdots,\alpha_r,\gamma_1,\cdots,\gamma_p,b_1,\cdots,b_q)^t$,
- $\mathrm{RSS}=||Y-\underline{X}\beta||^2$.

$\hookrightarrow$ échantillonneur de Gibbs.

# Application sur un jeu de donnée viticole

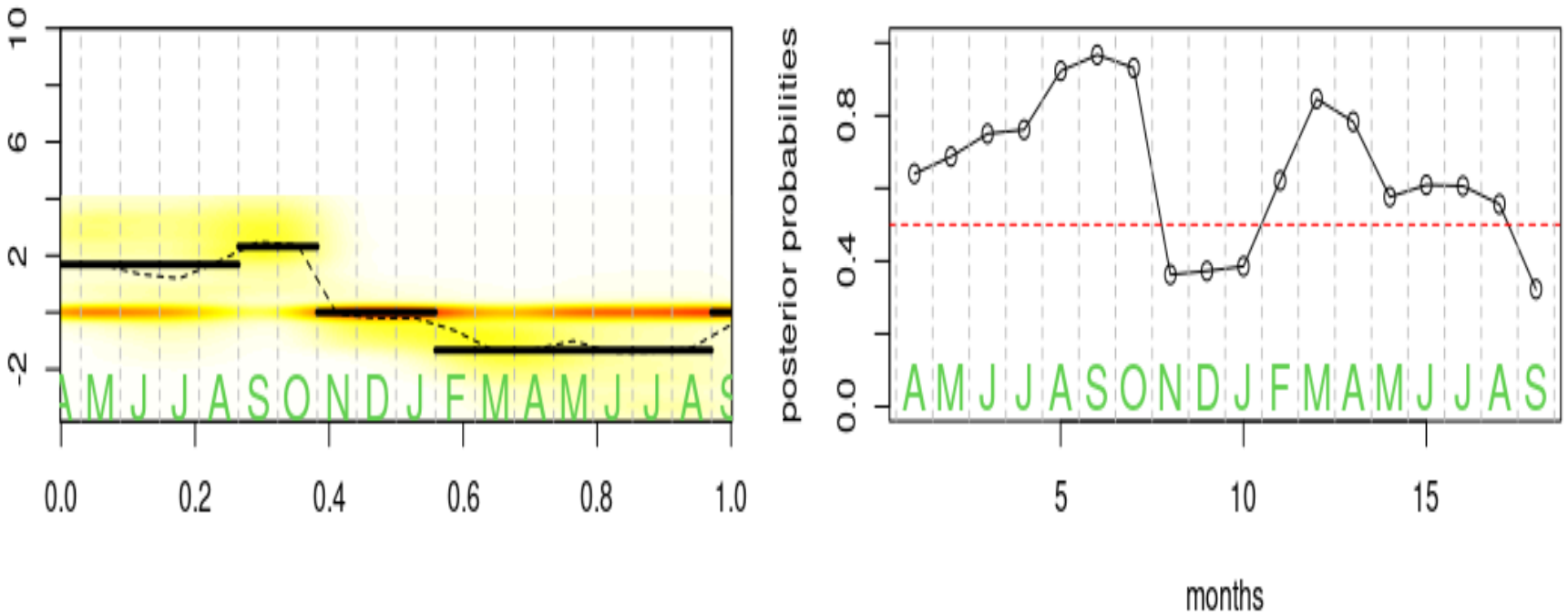
Effet des variations climatiques sur la production de vin.

- $\{Y_i,\ i = 1, \cdots, 1309\}$ : rendement d'une parcelle / an (en septembre).

Pour les 18 précédents mois :

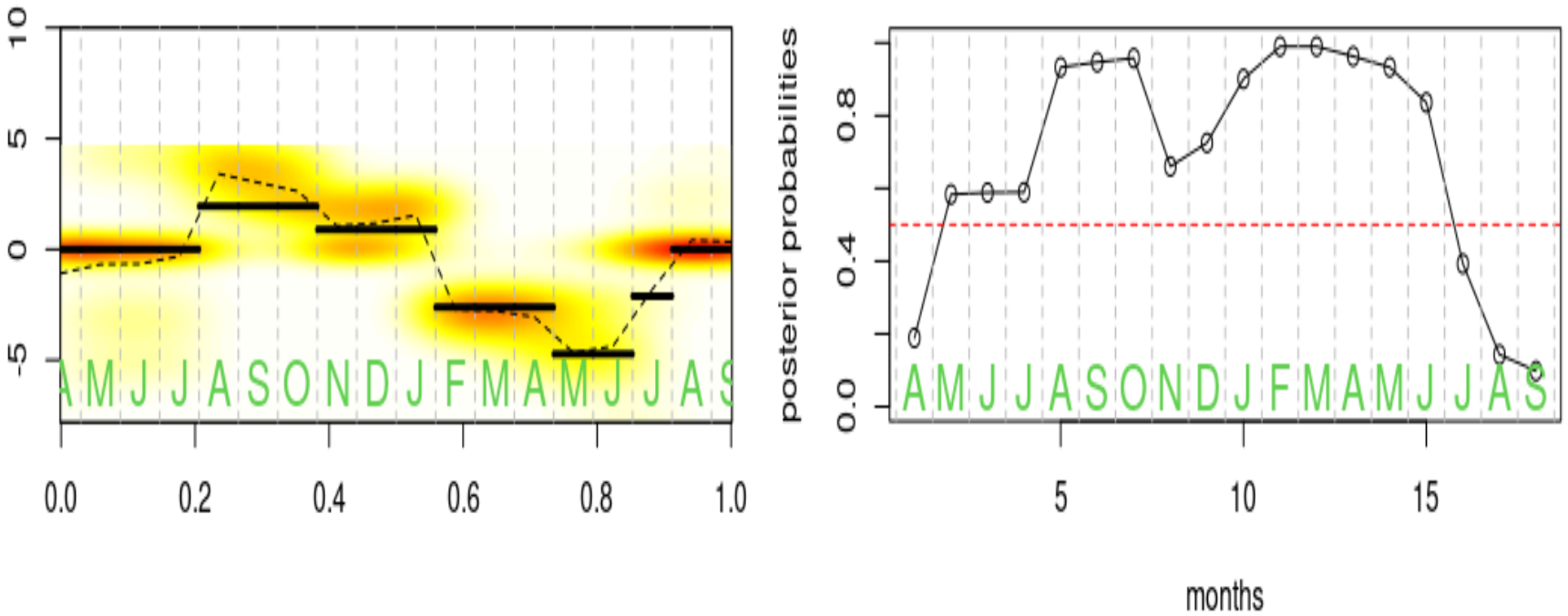
- $X_1^i(t)$ : Température mensuelle,
- $X_2^i(t)$ : Précipitation mensuelle,
- $X_3^i(t)$ : Évaporation-Transpiration mensuelle.

# $X_1(t)$ : Température moyenne



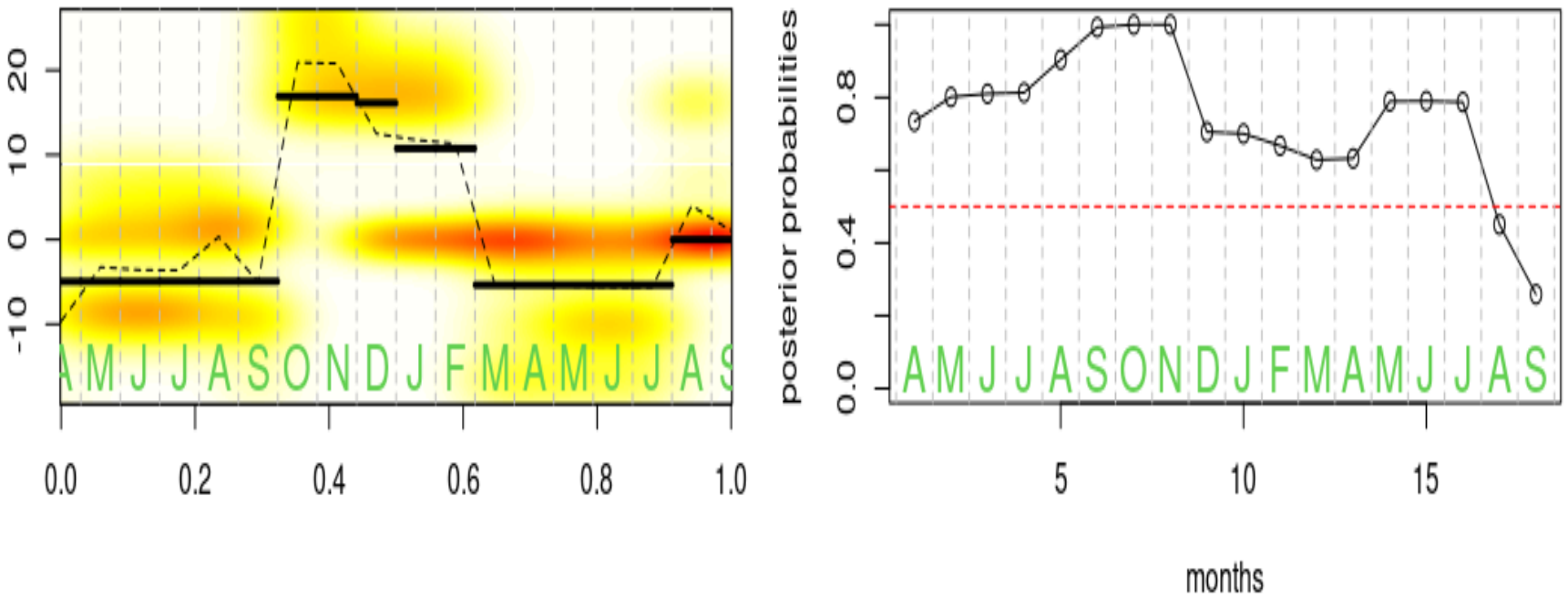

- Effet + de $X_1(t)$ d'avril à octobre de l'année n-1.
- Effet - de $X_1(t)$ de février à août de l'année n.

# $X_2(t)$ : Précipitation liquide




- Effet + de $X_2(t)$ d'aout à janvier de l'année n-1.
- Effet - de $X_2(t)$ de février à juillet de l'année n.

# $X_3(t)$ : Évaporation – Transpiration




- Effet - de $X_3(t)$ d'avril à septembre de l'année n-1 et de mars à juillet de l'année n.
- Effet + de $X_3(t)$ d'octobre à Février des années n-1/n.

## Conclusion

- Produire un estimateur de la fonction coefficient avec une forme très simple et surtout interprétable par les non-statisticiens.
- Package Bliss (la version étendue) : en cours.

## Perspectives

- Inclure les connaissances à priori,
- Considérer $Y$ comme une variable de survie pour étudier la mortalité des vignes. Identifier le modèle (censure (quel type)/ Troncature) à partir des données.
- Considérer $Y$ comme une variable binaire.