Estimation of the intensity of a counting process with high-dimensional covariates

Sarah Lemler

Chaire Modélisation Mathématique de la Biodiversité

CentraleSupélec

Outline

1. Framework and objectives
   - Context
   - Framework
   - Objectives

2. Estimation in the Cox model in high dimension
   - A two-step procedure
   - First step : estimation of the regression parameter in high dimension
   - Second step : estimation of the baseline intensity
   - Non-asymptotic oracle inequality

3. Simulations and application to a real dataset

Context

We want to estimate/learn the link between covariates

$$\boldsymbol{Z} = (Z_1, ..., Z_p)^T \in \mathbb{R}^p \quad \text{in high-dimension}$$

and the intensity with which some events occur :

- deaths/births
- asthma attacks
- purchases
- blog entries
- disasters
- ...

Application in ecology

- Birth-Death process :
  $N_t = n$ is the size of the population at time t
  - the population grows by 1 with a rate of birth $\lambda(n)$
  - the population decreases by 1 with a rate of death $\mu(n)$

Application in ecology

- Birth-Death process :
  $N_t = n$ is the size of the population at time t
  - the population grows by 1 with a rate of birth $\lambda(n)$
  - the population decreases by 1 with a rate of death $\mu(n)$
- Predator-Prey model : the Lokta-Volterra model
  $N(t) = [N_1(t), \ldots, N_k(t)]$, $N_i(t)$: size of the population i at time t
  - $n_1(t)$: number of preys
  - $n_2(t)$: number of predators

  Deterministic version of the model :

$$\begin{cases} n_1'(t) &= +an_1(t) - bn_1(t)n_2(t) \\ n_2'(t) &= -cn_2(t) + dn_1(t)n_2(t) \end{cases}$$

  Interpretation of the parameters :
  - $a = \lambda_1(t)$ rate of birth of the preys
  - $bn_2(t) = \mu_1(t)$ rate of death of the preys
  - $dn_1(t) = \lambda_2(t)$ rate of birth of the predators
  - $c = \mu_2(t)$ rate of death of the predators

Application in survival analysis

Example :

- 246 breast cancer patients :
  - 142 patients receiving Tamoxifen
  - 104 untreated patients

- Variables of interest : relapse free survival time (RFS), that can be right-censored
  - T_i relapse free survival time for individual i
  - C_i censoring time for individual i
  - $\delta_i = \mathbb{1}_{T_i \leq C_i}$ censoring indicator for individual i

- Covariates : 6 clinical variables, $\mathbf{44928}$ levels of gene expression

- Observations : for $i = 1, ..., 246$ and $p = 44934$
  - $X_i = \min(T_i, C_i)$ and δ_i
  - $\boldsymbol{Z_i} = (Z_{i,1}, ..., Z_{i,p})^T \in \mathbb{R}^p$ the vector of covariates

Hazard risk

Goal : to predict the RFS for the breast cancer adjusted on covariates

Conditional hazard : For $i = 1, ..., n$,

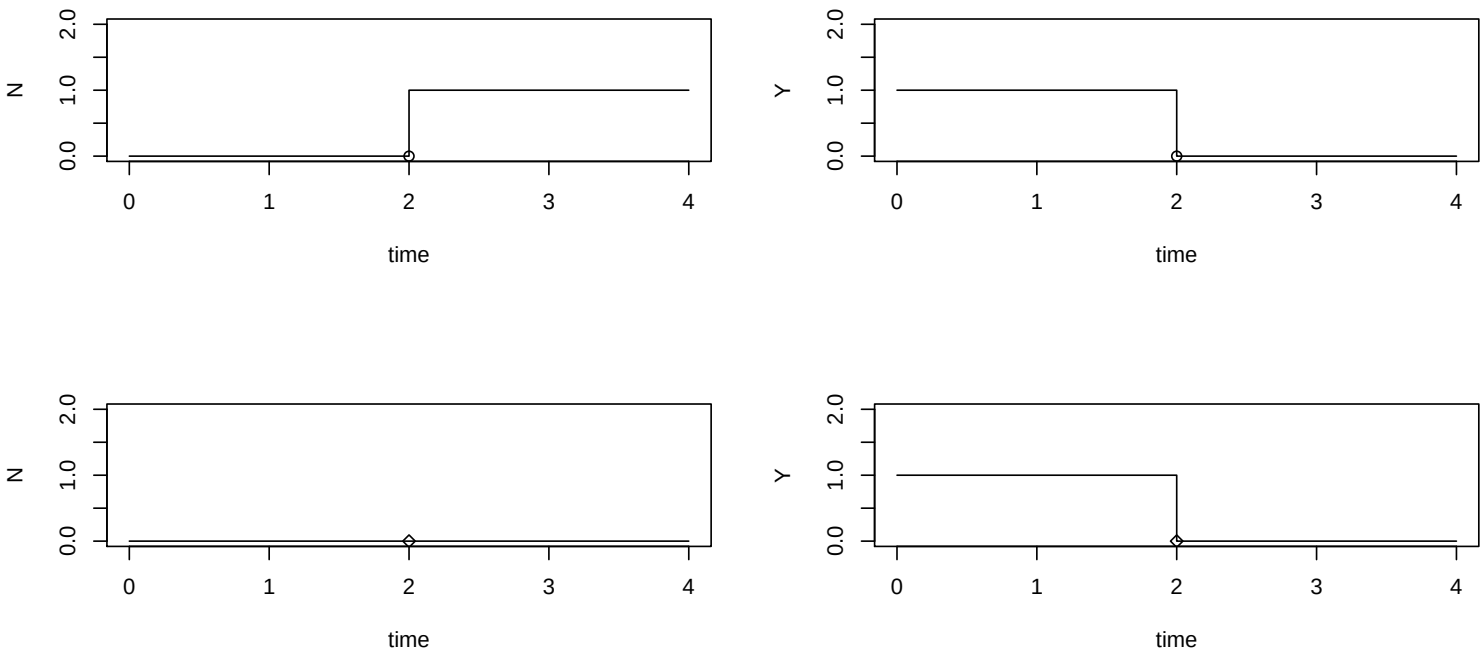
$$\lambda_0(t, \boldsymbol{Z_i}) = \frac{f_{T_i|\boldsymbol{Z}_i}(t)}{1 - F_{T_i|\boldsymbol{Z_i}}(t)}$$

- $f_{T_i|\boldsymbol{Z_i}}$ conditional density of T_i given $\boldsymbol{Z_i}$
- $F_{T_i|\boldsymbol{Z_i}}$ conditional distribution function of T_i given $\boldsymbol{Z_i}$

Counting processes in the specific case of right censoring

Counting processes [Aalen, 1980] :

- $N_i(t) = \mathbb{1}_{\{X_i \leq t,\ \delta_i = 1\}}$ counting process
- $Y_i(t) = \mathbb{1}_{\{X_i \geq t\}}$ at-risk process



Framework : the counting processes

For $i = 1, ..., n$,

- N_i counting process
- Y_i predictable random process with values in $[0, 1]$
- $[0, \tau]$ time interval between the beginning and the end of the study
- Observations : $(\boldsymbol{Z_i}, N_i(t), Y_i(t), i = 1, ..., n, 0 \leq t \leq \tau)$

Λ_i compensator of N_i, so that

$$M_i = N_i - \Lambda_i \in \mathcal{M}^2_{loc} \quad \text{(Doob-Meyer decomposition)}$$

Assumption 1. *N_i satisfies the Aalen multiplicative intensity model :*

$$\Lambda_i(t) = \int_0^t \lambda_0(s, \boldsymbol{Z_i}) Y_i(s) ds,$$

where λ_0 is an unknown nonnegative function called intensity

The Cox model

The Cox model :

$$\lambda_0(t, \boldsymbol{Z_i}) = \alpha_0(t) \exp(\boldsymbol{\beta_0^T Z_i})$$

- $\alpha_0 : \mathbb{R}^+ \to \mathbb{R}^+$ baseline intensity
- $\boldsymbol{\beta_0} \in \mathbb{R}^p$ regression parameter

$\Rightarrow$ semi-parametric model

The Cox partial log-likelihood

For a function $\lambda(t, \boldsymbol{Z_i}) = \alpha(t)\mathrm{e}^{\boldsymbol{\beta^T Z_i}}$, the log-likelihood [Jacod, 1973] is defined by

$$
\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^{n}\Big\{\int_0^\tau \log\lambda(t,\boldsymbol{Z_i})\mathrm{d}N_i(t) - \int_0^\tau \lambda(t,\boldsymbol{Z_i})Y_i(t)\mathrm{d}t\Big\} \\
=&-\frac{1}{n}\sum_{i=1}^{n}\Big\{\int_0^\tau \log\big(\alpha(t)S_n(t,\boldsymbol{\beta})\big)\mathrm{d}N_i(t)\Big\} - \int_0^\tau \alpha(t)S_n(t,\boldsymbol{\beta})\mathrm{d}t \\
&\underbrace{-\frac{1}{n}\sum_{i=1}^{n}\Big\{\int_0^\tau \log\frac{\mathrm{e}^{\boldsymbol{\beta^T Z_i}}}{S_n(t,\boldsymbol{\beta})}\mathrm{d}N_i(t)\Big\}}_{l_n^*} \text{ with } S_n(t,\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}Y_i(t)\mathrm{e}^{\boldsymbol{\beta^T Z_i}}
\end{aligned}
$$

l_n^* : Cox partial log-likelihood [Cox, 1972]

When $p < n$,

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\max}\{l_n^*(\boldsymbol{\beta})\}$$

Usual procedure to estimate parameters in the Cox model

The Cox model :

$$\lambda_0(t, \boldsymbol{Z_i}) = \alpha_0(t) \mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}, \quad \boldsymbol{Z_i} \in \mathbb{R}^p$$

Usual two-step procedure, when $p < n$:

1. First step : estimation of $\boldsymbol{\beta_0}$ using the Cox partial log-likelihood
2. Second step : estimation of α_0
   - Kernel estimator $\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}$ [see Ramlau-Hansen, 1983] :

$$\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}(t) = \sum_{i=1}^{n} \int_0^\tau \frac{1}{h} K\Big(\frac{t-u}{h}\Big) \frac{1}{\sum_{j=1}^n \mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{Z_j}} Y_j(u)} \mathrm{d}N_i(u),$$

     for some bandwidths $h > 0$ and $K : \mathbb{R} \to \mathbb{R}$ a kernel with integral 1.
   - Cross-validation to select the bandwidth [Ramlau-Hansen, 1983 and Grégoire, 1993]

Estimation of the regression parameter in high-dimension

When $p \gg n$:

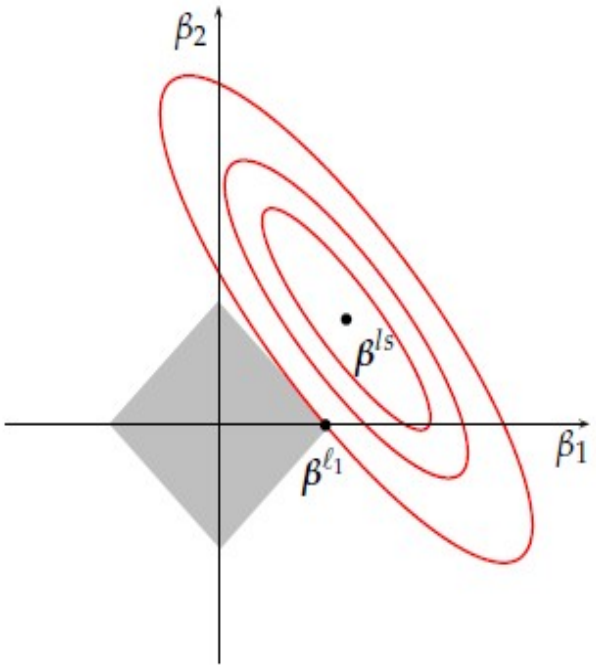
Lasso procedure in the Cox model [Tibshirani (1997)] :

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{L}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \{ -l_n^*(\boldsymbol{\beta}) + \Gamma_n |\boldsymbol{\beta}|_1 \},$$

$\Gamma_n > 0$ regularization parameter :

- in theory of order $\sqrt{\log(p)/n}$
- chosen in practice by cross-validation

Lasso procedure



$$\hat{\boldsymbol{\beta}}^{ls} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{C_n(\boldsymbol{\beta}) + \Gamma \sum_{j=1}^{p} |\beta_j|\}$$

$$\Leftrightarrow$$

$$\begin{cases} \hat{\boldsymbol{\beta}}^{ls} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{C_n(\boldsymbol{\beta})\} \\ \text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| \leq b \end{cases}$$

Advantages of the Lasso procedure :

- convex minimization problem $\Rightarrow$ computable in practice
- sparsity of the Lasso estimator $\Rightarrow$ results easily interpretable

State-of-the-art : estimation of an intensity with covariates

1. **In the Cox model :**
   - When $p < n$:
     - ⋆ Ramlau-Hansen (Ann. Stat., 1983) and Grégoire (Scand. J. Stat., 1993)

   ↪ Asymptotic results for the resulting estimators

   - In high dimension :
     - ⋆ Kong and Nan (Stat. Sin., 2014)
     - ⋆ Bradic and Song (Elec. Journ. Stat., 2015)
     - ⋆ Huang et al. (Ann. Stat., 2013)

   ↪ Non-asymptotic results for the Lasso estimator of $\boldsymbol{\beta_0}$

2. **General intensity :**
   - Comte et al. (AIHP, 2011) : results for a small number of covariates

   ↪ Procedure non-adapted to high-dimensional covariates

   - Lemler (AIHP, 2014) : approximation of a general intensity by a Cox model using a simultaneaous weighted Lasso procedure

   ↪ Estimation of both parameters $\boldsymbol{\beta_0}$ and α_0 with a Lasso procedure

Objectives

$$\lambda_0(t, \boldsymbol{Z_i}) = \alpha_0(t) \mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}, \quad \boldsymbol{Z_i} \in \mathbb{R}^p$$

$\Rightarrow$ Estimation of both parameters separately

- high-dimensional covariates $\Rightarrow$ high-dimension on $\boldsymbol{\beta}_0$
- estimation of $\alpha_0 : \mathbb{R}^+ \to \mathbb{R}^+$ with a procedure not specific to high dimension

Objectives

$$\lambda_0(t, \boldsymbol{Z_i}) = \alpha_0(t) \mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}, \quad \boldsymbol{Z_i} \in \mathbb{R}^p$$

$\Rightarrow$ Estimation of both parameters separately

- high-dimensional covariates $\Rightarrow$ high-dimension on $\boldsymbol{\beta}_0$
- estimation of $\alpha_0 : \mathbb{R}^+ \to \mathbb{R}^+$ with a procedure not specific to high dimension

$\Rightarrow$ Possibly in high dimension

- $p < n$ small dimension
- $p \approx \sqrt{n}$ intermediate case
- $p \geq n$ high dimension
- $p \gg n$ ultra-high dimension

Objectives

$$\lambda_0(t, \boldsymbol{Z_i}) = \alpha_0(t) \mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}, \quad \boldsymbol{Z_i} \in \mathbb{R}^p$$

$\Rightarrow$ Estimation of both parameters separately

- high-dimensional covariates $\Rightarrow$ high-dimension on $\boldsymbol{\beta}_0$
- estimation of $\alpha_0 : \mathbb{R}^+ \to \mathbb{R}^+$ with a procedure not specific to high dimension

$\Rightarrow$ Possibly in high dimension

- $p < n$ small dimension
- $p \approx \sqrt{n}$ intermediate case
- $p \geq n$ high dimension
- $p \gg n$ ultra-high dimension

$\Rightarrow$ Measurement of the influence of high-dimension on the estimation of α_0

Objectives

$$\lambda_0(t, \boldsymbol{Z_i}) = \alpha_0(t) \mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}, \quad \boldsymbol{Z_i} \in \mathbb{R}^p$$

$\Rightarrow$ Estimation of both parameters separately

- high-dimensional covariates $\Rightarrow$ high-dimension on $\boldsymbol{\beta}_0$
- estimation of $\alpha_0 : \mathbb{R}^+ \to \mathbb{R}^+$ with a procedure not specific to high dimension

$\Rightarrow$ Possibly in high dimension

- $p < n$ small dimension
- $p \approx \sqrt{n}$ intermediate case
- $p \geq n$ high dimension
- $p \gg n$ ultra-high dimension

$\Rightarrow$ Measurement of the influence of high-dimension on the estimation of α_0

$\Rightarrow$ Non-asymptotic result for the estimator of the baseline function α_0

1. Framework and objectives
   - Context
   - Framework
   - Objectives

2. Estimation in the Cox model in high dimension
   - A two-step procedure
   - First step : estimation of the regression parameter in high dimension
   - Second step : estimation of the baseline intensity
   - Non-asymptotic oracle inequality

3. Simulations and application to a real dataset

A two-step procedure

The Cox model in high dimension : $p \gg n$

$$\lambda_0(t, \boldsymbol{Z_i}) = \alpha_0(t) \mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}, \quad \boldsymbol{Z_i} \in \mathbb{R}^p$$

A two-step procedures :

- First step : estimation of $\boldsymbol{\beta_0}$ via a Lasso procedure
- Second step : estimation of α_0 via kernel estimation with a bandwidth selected by the Goldenshluger and Lepski method

First step :

Estimation of the regression parameter

First step : estimation of the regression parameter

Lasso estimation of $\boldsymbol{\beta_0}$:

$$\hat{\boldsymbol{\beta}} = \arg\min\{-l_n^*(\boldsymbol{\beta}) + \Gamma_n|\boldsymbol{\beta}|_1\},$$

where l_n^* is the Cox partial log-likelihood

First step : estimation of the regression parameter

Lasso estimation of $\boldsymbol{\beta_0}$:

$$\hat{\boldsymbol{\beta}} = \arg\min\{-l_n^*(\boldsymbol{\beta}) + \Gamma_n|\boldsymbol{\beta}|_1\},$$

where l_n^* is the Cox partial log-likelihood

Proposition [Huang et al. (2013)/Guilloux, L., Taupin (2014)]

Let $k > 0$, $c > 0$ and $s := \mathrm{Card}\{j : \beta_{0_j} \neq 0\}$. Under some Assumptions, with probability larger than $1 - cn^{-k}$,

$$|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}|_1 \leq C(s)\sqrt{\frac{\log(pn^k)}{n}}$$

Second step :

Estimation of the baseline intensity

The kernel estimator

$$\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}(t) = \frac{1}{nh} \sum_{i=1}^{n} \int_0^\tau \frac{K_h(t-u)}{S_n(u, \hat{\boldsymbol{\beta}})} \mathrm{d}N_i(u),$$

where

- $S_n(u, \hat{\boldsymbol{\beta}}) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{e}^{\hat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{Z_i}} Y_i(u)$
- $\hat{\boldsymbol{\beta}}$ is the Lasso estimator from the first step
- $K_h(.) = \frac{1}{h} K\left(\frac{.}{h}\right)$

The kernel estimator

$$\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}(t) = \frac{1}{nh} \sum_{i=1}^{n} \int_0^\tau \frac{K_h(t-u)}{S_n(u, \hat{\boldsymbol{\beta}})} \mathrm{d}N_i(u),$$

where

- $S_n(u, \hat{\boldsymbol{\beta}}) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{e}^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z_i}} Y_i(u)$
- $\hat{\boldsymbol{\beta}}$ is the Lasso estimator from the first step
- $K_h(.) = \frac{1}{h} K\left(\frac{.}{h}\right)$

General idea to choose the bandwidth

- $\mathcal{H}_n$ a grid of bandwidths $h > 0$
- Set of estimators :

$$\mathcal{F}(\mathcal{H}_n) = \{\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}, h \in \mathcal{H}_n\}$$

- Oracle :

$$h^* = \underset{h \in \mathcal{H}_n}{\arg\min} \{ \underbrace{||\alpha_0 - \mathbb{E}[\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}]||^2}_{B(\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}})} + \underbrace{\mathbb{E}||\mathbb{E}[\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}] - \hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}||^2}_{V(h)} \}$$

- Estimated bias : $\hat{B}(\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}})$
- Selection procedure of h :

$$\hat{h}^{\hat{\boldsymbol{\beta}}} = \underset{h \in \mathcal{H}_n}{\arg\min} \{ \hat{B}(\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}) + V(h) \}$$

- Final estimator : $\hat{\alpha}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}^{\hat{\boldsymbol{\beta}}}$

The Goldenshluger and Lepski method (2011)

Selection of $\hat{h}^{\hat{\boldsymbol{\beta}}} \in \mathcal{H}_n$ s.t. the risk of $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}} \in \mathcal{F}(\mathcal{H}_n)$ is as close as possible to

$$\min_{h \in \mathcal{H}_n} \{||\alpha_0 - K_h * \alpha_0||_2^2 + V(h)\}$$

- Estimated bias :

$$\hat{B}(\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_h) = \sup_{h' \in \mathcal{H}_n} \left\{ ||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h'} - \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h,h'}||_2^2 - V(h') \right\}_+,$$

  with $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{h,h'} = K_{h'} * \hat{\alpha}^{\hat{\boldsymbol{\beta}}}_h$ and

$$V(h) \propto \kappa ||\alpha_0||_{\infty,\tau} \frac{||K||_2^2}{nh}$$

- Selected bandwidth :

$$\hat{h}^{\hat{\boldsymbol{\beta}}} = \arg\min_{h \in \mathcal{H}_n} \{\hat{B}(\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_h) + V(h)\}$$

- Resulting estimator : $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}$

Non-asymptotic oracle inequalities for $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}$

Theorem [Guilloux, L., Taupin (2016)]

Under classical assumptions, for some $a \geq 0$ and C, $C'(s)$ and L some contants,

$$\begin{aligned}\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}} - \alpha_0||_2^2] &\leq C \inf_{h \in \mathcal{H}_n} \left\{ ||\alpha_h - \alpha_0||_2^2 + V(h) \right\} + C'(s) \frac{\ln^a n \ln np}{n} \\ &\leq C \inf_{h \in \mathcal{H}_n} \left\{ ||\alpha_h - \alpha_0||_2^2 + \frac{L}{nh} \right\} + C'(s) \frac{\ln^a n \ln np}{n}\end{aligned}$$

Non-asymptotic oracle inequalities for $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}$

Theorem [Guilloux, L., Taupin (2016)]

Under classical assumptions, for some $a \geq 0$ and C, $C'(s)$ and L some contants,

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}} - \alpha_0||_2^2] \leq C \inf_{h \in \mathcal{H}_n} \left\{ ||\alpha_h - \alpha_0||_2^2 + V(h) \right\} + C'(s) \frac{\ln^a n}{n} \ln np$$

$$\leq C \inf_{h \in \mathcal{H}_n} \left\{ ||\alpha_h - \alpha_0||_2^2 + \frac{L}{nh} \right\} + C'(s) \frac{\ln^a n}{n} \ln np$$

Non-asymptotic oracle inequalities for $\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}}$

Theorem [Guilloux, L., Taupin (2016)]

Under classical assumptions, for some $a \geq 0$ and C, $C'(s)$ and L some contants,

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}^{\hat{\boldsymbol{\beta}}}} - \alpha_0||_2^2] \leq C \inf_{h \in \mathcal{H}_n} \left\{ ||\alpha_h - \alpha_0||_2^2 + V(h) \right\} + C'(s) \ln^a n \frac{\ln np}{n}$$

$$\leq C \inf_{h \in \mathcal{H}_n} \left\{ ||\alpha_h - \alpha_0||_2^2 + \frac{L}{nh} \right\} + C'(s) \ln^a n \frac{\ln np}{n}$$

Comparison model selection / kernel estimation

Comparison of the oracle inequalities

- Kernel estimation [Guilloux, L., Taupin (2016)]

$$V(h) = \kappa ||\alpha_0||_{\infty,\tau} \frac{||K||_2^2}{nh},$$

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}\hat{\boldsymbol{\beta}}} - \alpha_0||_2^2] \leq C \inf_{h \in \mathcal{H}_n} \left\{ ||\alpha_h - \alpha_0||_2^2 + V(h) \right\} + C'(s) \log^a(n) \frac{\log np}{n}$$

- Model selection [Guilloux, L., Taupin (2015)]

$$\text{pen}(m) := \kappa(1 + ||\alpha_0||_{\infty,\tau}) \frac{D_m}{n},$$

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}\hat{\boldsymbol{\beta}}} - \alpha_0||_{det}^2] \leq \tilde{C} \inf_{m \in \mathcal{M}_n} \{||\alpha_0 - \alpha_m||_{det}^2 + \text{pen}(m)\} + \tilde{C}'(s) \frac{\log np}{n}$$

Model selection

- $\mathcal{M}_n$ a set of indices,
- $\{S_m, m \in \mathcal{M}_n\}$ a collection of models such that

$$S_m = \{\alpha : \alpha = \sum_{j \in J_m} a_j^m \varphi_j^m, a_j^m \in \mathbb{R}\},$$

where $(\varphi_j^m)_{j \in J_m}$ is an orthonormal basis of $(L^2 \cap L^\infty)([0, \tau])$.

- $D_m := |J_m|$, dimension of S_m.

Comparison model selection / kernel estimation

Comparison of the oracle inequalities

- Kernel estimation [Guilloux, L., Taupin (2016)]

$$V(h) = \kappa ||\alpha_0||_{\infty,\tau} \frac{||K||_2^2}{nh},$$

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{h}\hat{\boldsymbol{\beta}}} - \alpha_0||_2^2] \leq C \inf_{h \in \mathcal{H}_n} \left\{ ||\alpha_h - \alpha_0||_2^2 + V(h) \right\} + C'(s) \log^a(n) \frac{\log np}{n}$$

- Model selection [Guilloux, L., Taupin (2015)]

$$\text{pen}(m) := \kappa(1 + ||\alpha_0||_{\infty,\tau}) \frac{D_m}{n},$$

$$\mathbb{E}[||\hat{\alpha}^{\hat{\boldsymbol{\beta}}}_{\hat{m}\hat{\boldsymbol{\beta}}} - \alpha_0||_{det}^2] \leq \tilde{C} \inf_{m \in \mathcal{M}_n} \{||\alpha_0 - \alpha_m||_{det}^2 + \text{pen}(m)\} + \tilde{C}'(s) \frac{\log np}{n}$$

1. Framework and objectives
   - Context
   - Framework
   - Objectives

2. Estimation in the Cox model in high dimension
   - A two-step procedure
   - First step : estimation of the regression parameter in high dimension
   - Second step : estimation of the baseline intensity
   - Non-asymptotic oracle inequality

3. Simulations and application to a real dataset

Simulations : simulated data

$$\lambda_0(t, \boldsymbol{Z_i}) = \alpha_0(t) \mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}, \quad \text{for } i = 1, ..., n$$

where $\boldsymbol{Z_i} = (Z_{i,1}, ..., Z_{i,p})^T$ is the vector of covariates of individual i.

- Sample sizes and number of covariates :
  - $n = 200$ or $n = 500$
    - ★ $p = 5 < n$ small dimension case
    - ★ $p = \sqrt{n}$ $(= 15$ or $22)$ intermediate case
    - ★ $p = n$ high dimension case
- Design matrix : $\boldsymbol{Z} \sim \mathcal{U}([-1, 1])$, $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$
- Rate of censoring : $\approx 20\%$ or 50%

True parameters :

- Baseline function : $\alpha_0(t) = q\lambda t^{q-1}$ (Weibull distribution),
- Regression parameter : $\boldsymbol{\beta_0} = (0.1, 0.3, 0.5, 0, ..., 0)^T \in \mathbb{R}^p$

Simulations : estimation of the parameters

- Estimation of $\boldsymbol{\beta_0}$
  - Lasso [Simon et al., 2011] :

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \{-l_n^*(\boldsymbol{\beta}) + \mu|\boldsymbol{\beta}|_1\}$$

    where μ is chosen by cross-validation

Simulations : estimation of the parameters

- Estimation of $\boldsymbol{\beta_0}$
  - Lasso [Simon et al., 2011] :

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \{-l_n^*(\boldsymbol{\beta}) + \mu|\boldsymbol{\beta}|_1\}$$

    where μ is chosen by cross-validation

- Kernel estimation of α_0
  - Epanechnikov kernel :

$$K(u) = \frac{3}{4}(1-u^2)\mathbb{1}_{\{|u|\leq 1\}}$$

  - Choice of the bandwidth :
    - ★ cross-validation of Ramlau-Hansen
    - ★ the Goldenshluger and Lepski method

Simulations : estimation of the parameters

- Estimation of $\boldsymbol{\beta_0}$
  - Lasso [Simon et al., 2011] :

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \{-l_n^*(\boldsymbol{\beta}) + \mu|\boldsymbol{\beta}|_1\}$$

    where μ is chosen by cross-validation

- Kernel estimation of α_0
  - Epanechnikov kernel :

$$K(u) = \frac{3}{4}(1-u^2)\mathbb{1}_{\{|u|\leq 1\}}$$

  - Choice of the bandwidth :
    - ★ cross-validation of Ramlau-Hansen
    - ★ the Goldenshluger and Lepski method
- Estimation of α_0 by model selection :
  - Histogram basis : $\varphi_j(t) = \frac{1}{\sqrt{\tau}} 2^{m/2} \mathbb{1}_{[(j-1)\tau/2^m, j\tau/2^m[}(t)$, for $j = 1, ..., 2^m$, $D_m = 2^m$.

Simulations : MISEs for a known regression parameter

For $\boldsymbol{\beta_0}$ known, $\alpha_0 \sim \mathcal{W}(1.5, 1)$, $n = 200$, $p = 15$:

$$\mathrm{MISE}(\hat{\alpha}^{GL}_{\hat{h}}) = 0.015$$
$$\mathrm{MISE}(\hat{\alpha}^{CV}_{\hat{h}}) = 0.018$$

Simulations : MISEs for the different estimation procedures

For $\alpha_0 \sim \mathcal{W}(1.5, 1)$

Estimators \ Dimensions	$n = 200$		$n = 500$	
	$p = 15$	$p = 200$	$p = 22$	$p = 500$
KernelCV	0.023	0.045	0.010	0.023
KernelGL	0.017	0.044	0.009	0.022

For $\alpha_0 \sim \mathcal{W}(0.5, 2)$

Estimators \ Dimensions	$n = 200$		$n = 500$	
	$p = 15$	$p = 200$	$p = 22$	$p = 500$
KernelCV	1.561	1.556	1.521	1.515
KernelGL	1.02	0.923	1.006	1.098

Simulations : MISEs for different rate of censoring

For $\alpha_0 \sim \mathcal{W}(1.5, 1)$ and two rates of censoring 20% and 50%

Dimensions \ MISEs		20%		50%	
		GL	CV	GL	CV
$n = 200$	$p = 15$	0.014	0.017	0.023	0.029
	$p = 500$	0.013	0.016	0.022	0.026
$n = 500$	$p = 22$	0.009	0.007	0.011	0.012
	$p = 1000$	0.008	0.008	0.011	0.013

Simulations : MISEs for the different estimation procedures

For $\alpha_0 \sim \mathcal{W}(1.5, 1)$

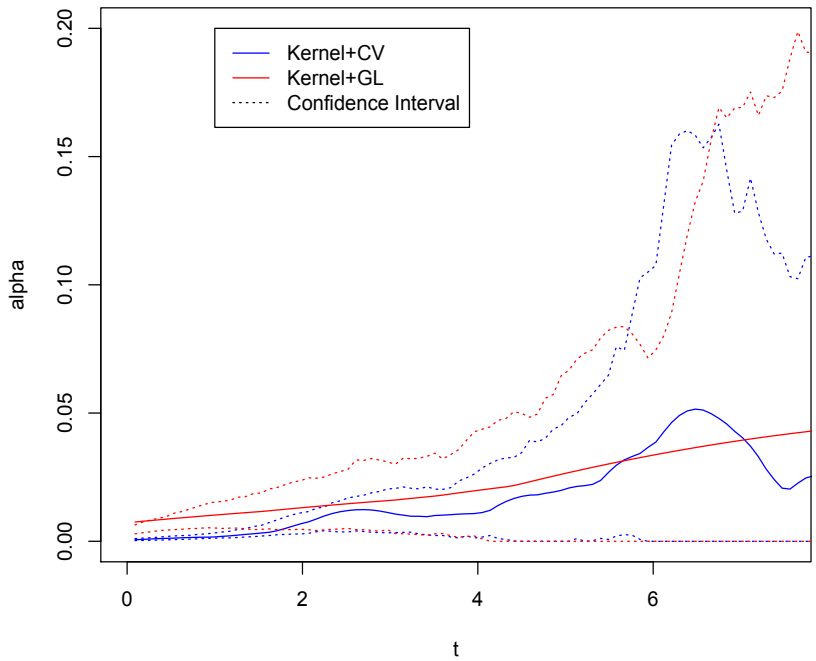
Estimators \ Dimensions	$n = 200$		$n = 500$	
	$p = 15$	$p = 200$	$p = 22$	$p = 500$
KernelCV	0.023	0.045	0.010	0.023
KernelGL	0.017	0.044	0.009	0.022
Model Selection Hist	0.072	0.071	0.055	0.059

Dataset and screening step

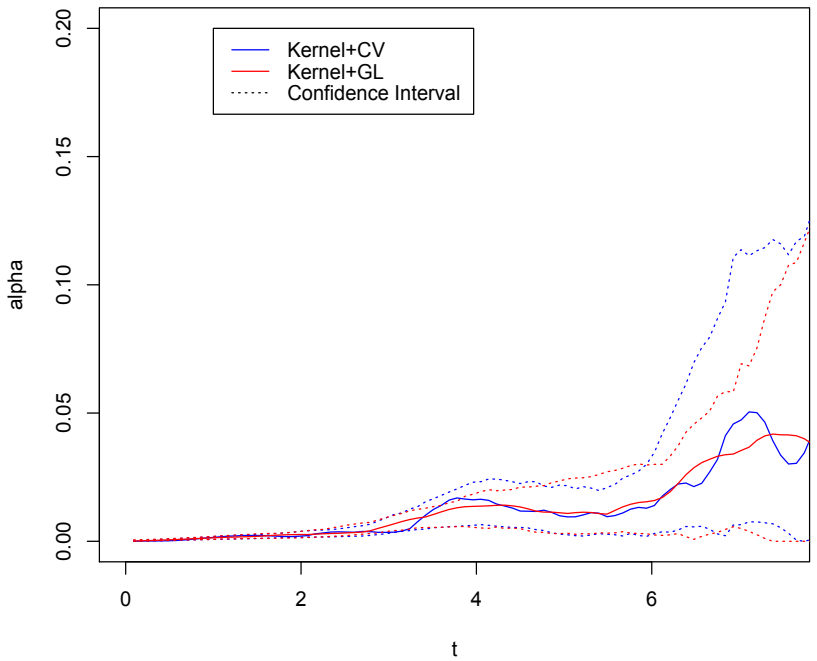
- 246 patients breast cancer patients :
  - 142 patients receiving Tamoxifen,
  - 104 untreated patients
- Variable of interest : time of relapse free survival, that can be right-censored
- Covariates : 6 clinical variables, $\mathbf{44928}$ levels of gene expression

Screening step : $p = 1000$

Dataset : plots of the estimators of the baseline functions



(a) Untreated patients (p=1000).



(b) Tamoxifen patients (p=1000).

Conclusion

- Two-step procedures to estimate both parameters Cox model with high-dimensional covariates :
  - Lasso procedure to estimate $\boldsymbol{\beta_0}$
  - Non-parametric procedures to estimate α_0 :
    - ⋆ Model selection
    - ⋆ Kernel estimation with the Goldenshluger and Lepski method
- Non-asymptotic oracle inequalities for the estimators of α_0
- Measurement of the influence of the high dimension on the estimation of α_0


Guilloux A., Lemler S. and Taupin M-L. Adaptive estimation of the baseline hazard function in the Cox model by model selection, with high-dimensional covariates. *JSPI* (2015).


Guilloux A., Lemler S. and Taupin M-L. Adaptive kernel estimation of the baseline function in the Cox model, with high-dimensional covariates. *JMVA* (2016).

Perspectives

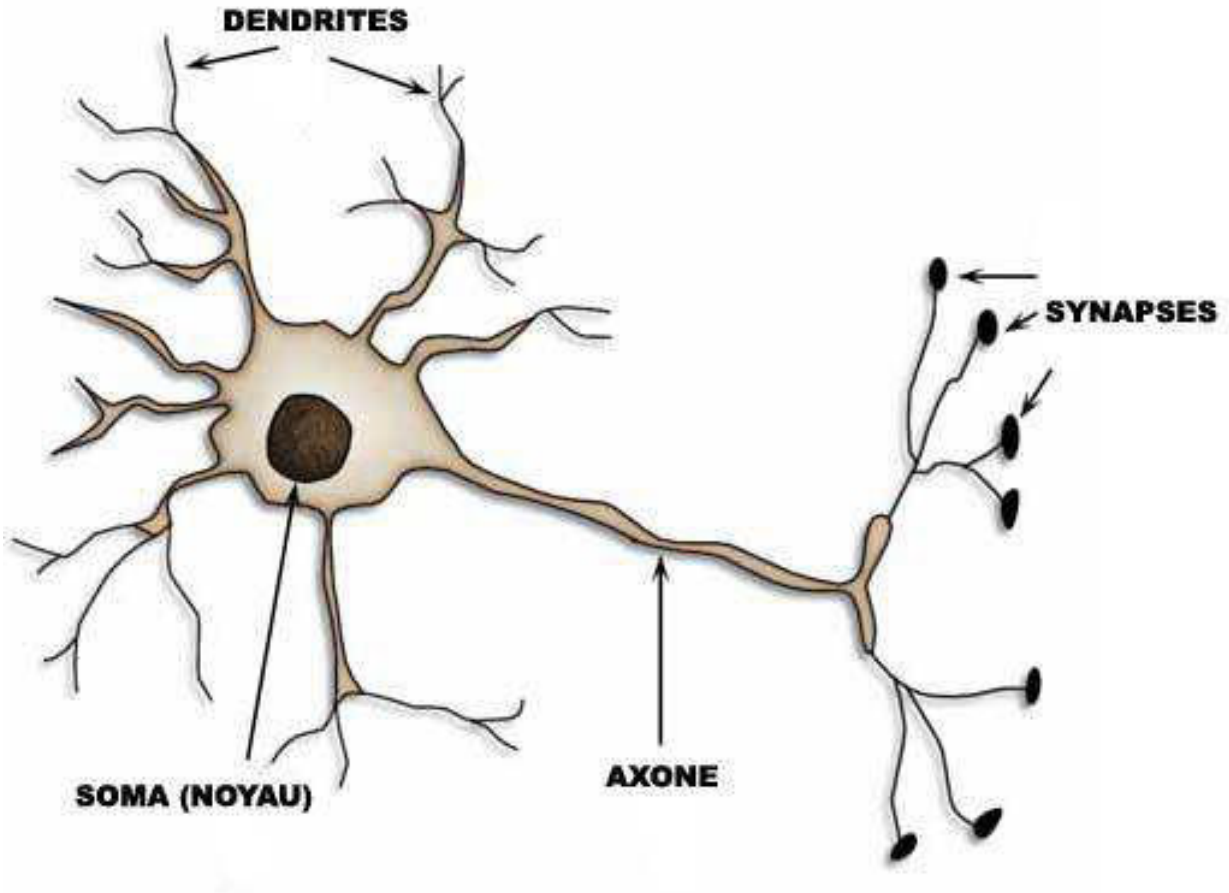
- Interaction models to investigate potential crossed effects of the treatment and some genes

- Cox model with covariates and a regression parameter that depend on the time

$$\lambda_0(t, \boldsymbol{Z_i}(t)) = \alpha_0(t) \mathrm{e}^{\boldsymbol{\beta_0}(\boldsymbol{t})^{\boldsymbol{T}} \boldsymbol{Z_i}(\boldsymbol{t})}$$

Perspectives : applications to SDE

Application in neuroscience :



Perspectives : applications to SDE

- **Spike train :** sequence of the times of occurence of the action potentials of a neuron
- **Spike :** time of occurence

Two kind of data :

- Extracellular signal : spike trains,
  action potential of several neurons $\Rightarrow$ spike sorting
  $\hookrightarrow$ discrete signal
- Intracellular signal : membrane potential in mV (all the electric fluctuations in the cellular membrane, including those who do not lead to action potential)
  $\hookrightarrow$ continuous signal

Perspectives : applications to SDE

- $X = (X_t)_{t \geq 0}$ action potential of a fixed neuron
- measurements of M spike trains from M different neurons

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma \mathrm{d}W_{t-} + \sum_{j=1}^{M} a(X_{t-}) dN_t^j$$

with $N = (N_t^1, \ldots, N_t^M)$ a multivariate Hawkes process

$$\mathrm{d}N_t^j = \lambda_t^j \mathrm{d}t + \mathrm{d}M_t^j$$

and λ_t^j an estimator of the intensity of a Hawkes process or a Cox model :

$$\lambda_t^j = \left(\nu_j + \sum_{\ell=1}^{M} \int_{-\infty}^{t-} h_{\ell \to j}(t-u) dN_u^\ell \right)_+$$

Appendix : assumptions on kernels and bandwidths

1) K has a compact support $[-1,1]$

$$\int_{-1}^{1} K(u)\mathrm{d}u = 1 \quad \text{and} \quad ||K||_2^2 = \int_{-1}^{1} K^2(u)\mathrm{d}u < \infty.$$

2) $nh \geq 1$ and $0 < h < 1$.

3) $\mathrm{Card}(\mathcal{H}_n) \leq n$

4) For some $a \geq 0$, $\sum_{h\in\mathcal{H}_n} \frac{1}{nh} = \mathcal{O}(\ln^a(n))$

5) For all $b > 0$, $\sum_{h\in\mathcal{H}_n} \exp(-b/h) < +\infty$

6) For $j \in \{0,1,2\}$ the function $x \mapsto x^j K(x)$ is integrable and

$$\int_{\mathbb{R}} xK(x)\mathrm{d}x = 0 \quad \text{and} \quad \int_{\mathbb{R}} x^2 K(x)\mathrm{d}x < \infty.$$

Appendix : Adaptive selection of the bandwidth

- Idea : introduction of a pseudo-estimator

$$\bar{\alpha}_h(t) = \frac{1}{nh}\sum_{i=1}^{n}\int_0^\tau \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathrm{d}N_i(u), \quad S(u,\boldsymbol{\beta_0}) = \mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}]$$

Appendix : Adaptive selection of the bandwidth

- Idea : introduction of a pseudo-estimator

$$\bar{\alpha}_h(t) = \frac{1}{nh}\sum_{i=1}^{n}\int_0^\tau \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathrm{d}N_i(u), \quad S(u,\boldsymbol{\beta_0}) = \mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}]$$

- Estimation of an upper bound of the excess risk : $\mathbb{E}(\bar{\alpha}_h) = K_h * \alpha_0$

$$\mathbb{E}||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \alpha_0||_2^2 \leq C\left\{\mathbb{E}||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h||_2^2 + \mathbb{E}||\bar{\alpha}_h - K_h * \alpha_0||_2^2 + ||K_h * \alpha_0 - \alpha_0||_2^2\right\}$$

Appendix : Adaptive selection of the bandwidth

- Idea : introduction of a pseudo-estimator

$$\bar{\alpha}_h(t) = \frac{1}{nh}\sum_{i=1}^{n}\int_0^\tau \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathrm{d}N_i(u), \quad S(u,\boldsymbol{\beta_0}) = \mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}]$$

- Estimation of an upper bound of the excess risk : $\mathbb{E}(\bar{\alpha}_h) = K_h * \alpha_0$

$$\mathbb{E}||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \alpha_0||_2^2 \leq C\left\{\mathbb{E}||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h||_2^2 + \mathbb{E}||\bar{\alpha}_h - K_h * \alpha_0||_2^2 + ||K_h * \alpha_0 - \alpha_0||_2^2\right\}$$

- $\mathbb{E}||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h||_2^2$ bounded by a constant that does not depend on h :

$$\mathbb{E}||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \alpha_0||_2^2 \leq C(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta_0}) + C\left\{||\alpha_0 - K_h * \alpha_0||_2^2 + \mathbb{E}||\bar{\alpha}_h - K_h * \alpha_0||_2^2\right\}$$

Appendix : Adaptive selection of the bandwidth

- Idea : introduction of a pseudo-estimator

$$\bar{\alpha}_h(t) = \frac{1}{nh}\sum_{i=1}^{n}\int_0^\tau \frac{K_h(t-u)}{S(u,\boldsymbol{\beta_0})}\mathrm{d}N_i(u), \quad S(u,\boldsymbol{\beta_0}) = \mathbb{E}[\mathrm{e}^{\boldsymbol{\beta_0^T Z_i}}]$$

- Estimation of an upper bound of the excess risk : $\mathbb{E}(\bar{\alpha}_h) = K_h * \alpha_0$

$$\mathbb{E}||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \alpha_0||_2^2 \leq C\Big\{\mathbb{E}||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h||_2^2 + \mathbb{E}||\bar{\alpha}_h - K_h * \alpha_0||_2^2 + ||K_h * \alpha_0 - \alpha_0||_2^2\Big\}$$

- $\mathbb{E}||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \bar{\alpha}_h||_2^2$ bounded by a constant that does not depend on h :

$$\mathbb{E}||\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}} - \alpha_0||_2^2 \leq C(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta_0}) + C\Big\{||\alpha_0 - K_h * \alpha_0||_2^2 + \mathbb{E}||\bar{\alpha}_h - K_h * \alpha_0||_2^2\Big\}$$

- Estimation of $\bar{h}^*$ instead of the oracle :

$$\bar{h}^* = \underset{h\in\mathcal{H}_n}{\arg\min}\Big\{||\alpha_0 - K_h * \alpha_0||_2^2 + \mathbb{E}||\bar{\alpha}_h - K_h * \alpha_0||_2^2\Big\}.$$