Estimation of the conditional cumulative distribution function from current status data by model selection

Sandra Plancade
INRA (Institut français de recherche en agronomie)

Séminaire de l'ISBA - 24 février 2017

1. Current status data
2. Model selection (univariate framework)
3. Adaptive estimation of the conditional c.d.f from current status data
4. Results on simulated data

1. Current status data

2. Model selection (univariate framework)

3. Adaptive estimation of the conditional c.d.f from current status data

4. Results on simulated data

Survival analysis

- **Survival analysis** : interest in distribution of time to an event T (depending on covariates)

Survival analysis

- **Survival analysis** : interest in distribution of time to an event T (depending on covariates)
- **Censoring** : T is not always/directly observed
$\hookrightarrow$ Most widely studies : right censoring $(Y = \min(C, T), \Delta = \mathbb{1}_{T \leqslant C})$, with C censoring time.

Survival analysis

- **Survival analysis** : interest in distribution of time to an event T (depending on covariates)

- **Censoring** : T is not always/directly observed
  $\hookrightarrow$ Most widely studies : right censoring $(Y = \min(C, T), \Delta = 1\!\!1_{T \leqslant C})$, with C censoring time.

- **Current status data** :

$$\begin{cases} C \text{ observation time} \\ \Delta = 1\!\!1_{T \leqslant C} \\ Z \text{ covariate} \end{cases}$$

$\hookrightarrow$ Require assumptions on $\mathbb{P}[C|T, Z]$ (most usual : $T \perp\!\!\!\perp C|Z$).

$\hookrightarrow$ Vocabulary of survival analysis : T is the time of death and Δ is the current status (dead or alive) at the observation time.

An example of current status data

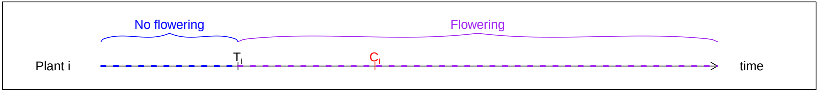
- Distribution of the time T of **floral initiation** in maize, as a function of covariates Z (génotype, etc)

An example of current status data

- Distribution of the time T of **floral initiation** in maize, as a function of covariates Z (génotype, etc)

- At time C_i, the plant i is dissected and we observe if the floral initiation has already occured ($\Delta_i = 1$) or not ($\Delta_i = 0$)

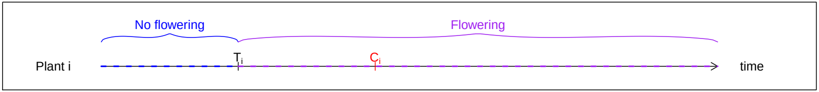
  $\hookrightarrow$ **the observation process destroys** the plant.



An example of current status data

- Distribution of the time T of **floral initiation** in maize, as a function of covariates Z (génotype, etc)
- At time C_i, the plant i is dissected and we observe if the floral initiation has already occured $(\Delta_i = 1)$ or not $(\Delta_i = 0)$

  $\hookrightarrow$ **the observation process destroys** the plant.

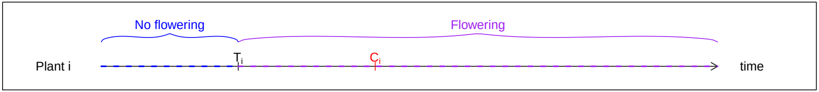


- Goal : infer $\mathbb{P}[T_i|Z_i]$ from a sample $i = 1, \ldots, n$ of plants with varying observation times $(C_i)_{i=1,\ldots,n}$.

An example of current status data

- Distribution of the time T of **floral initiation** in maize, as a function of covariates Z (génotype, etc)
- At time C_i, the plant i is dissected and we observe if the floral initiation has already occured ($\Delta_i = 1$) or not ($\Delta_i = 0$)

  $\hookrightarrow$ **the observation process destroys** the plant.



- Goal : infer $\mathbb{P}[T_i|Z_i]$ from a sample $i = 1, \ldots, n$ of plants with varying observation times $(C_i)_{i=1,\ldots,n}$.
- Δ_i indicates if T_i lies in $[0, C_i]$ or $[C_i, +\infty[$. Current status data framework is also called *interval censoring "case 1"*.

Brief review on estimation from current status data

$(T_i)_{i=1,\dots,n}$ unobserved, $(C_i, \Delta_i = 1\!\!1_{T_i \leqslant C_i}, Z_i)_{i,\dots,n}$ observations

- **Non Parametric Maximum Likelihood Estimator** of the c.d.f. The likelihood only depends on the value of the c.d.f. at the observation points (C_i), and admit a unique minimiser over $\{h : \mathbb{R}_+ \mapsto [0,1]\}$ with an explicit expression.

- **Quantile regression** is based on the invariance of quantiles by monotone tranformations, and the observation that the following function is decreasing.

$$h: \begin{array}{ccc} x & \mapsto & 1\!\!1_{x \leqslant C} \\ T & \mapsto & \Delta \end{array}$$

- **Inverse Probability of censoring Weighted Estimator (IPWE)** : the risk $\mathbb{E}[L((T,Z),h)]$ is estimated by an empirical contrast with observation Δ_i weighted by the inverse of the "probability to be observed" at C_i : $f_C(C_i)$.

- **Semi-parametric models** : the distribution of T is linearly related to the covariates Z

$$\mathbb{P}[T|Z] = \phi(T, \langle \beta, Z \rangle)$$

- **Dependent censoring** : models on the joint distribution of T and C

Framework of this presentation

- We consider an i.i.d. sample $(T_i, C_i, Z_i)_{i=1,\dots,n}$ and the observation sample

$$(C_i, Z_i, \Delta_i)_{i=1,\dots,n} \quad \text{with}$$

  - $\Delta_i = \mathbb{1}_{T_i \leqslant C_i}$.
  - $Z_i \in \mathbb{R}$.
  - $T_i, C_i \in \mathbb{R}_+$
  - $T_i \perp\!\!\!\perp C_i | Z_i$

Framework of this presentation

- We consider an i.i.d. sample $(T_i, C_i, Z_i)_{i=1,\dots,n}$ and the observation sample

$$(C_i, Z_i, \Delta_i)_{i=1,\dots,n} \quad \text{with}$$

  - $\Delta_i = \mathbb{1}_{T_i \leqslant C_i}$.
  - $Z_i \in \mathbb{R}$.
  - $T_i, C_i \in \mathbb{R}_+$
  - $T_i \perp\!\!\!\perp C_i | Z_i$

- **Goal** : Estimate the conditional c.d.f. of T_i given Z_i

$$F(t, z) = \mathbb{P}[T_i \leqslant t | Z_i = z]$$

on a compact $A = A_1 \times A_2 \subset \mathbb{R}_+ \times \mathbb{R}$.

Least-square contrast

- Let $(C_i, Z_i, \Delta_i)_{i=1,\dots,n}$ the i.i.d. observation sample with $\Delta_i = 1\!\!1_{T_i \leqslant C_i}$.
- By définition of Δ_i,

$$\begin{aligned}\mathbb{E}[\Delta_i | C_i = c, Z_i = z] &= \mathbb{E}[1\!\!1_{T_i \leqslant C_i} | C_i = c, Z_i = z] \\ &= \mathbb{P}[T_i \leqslant c | Z_i = z] \\ &= F(c, z)\end{aligned}$$

Least-square contrast

- Let $(C_i, Z_i, \Delta_i)_{i=1,\dots,n}$ the i.i.d. observation sample with $\Delta_i = 1\!\!1_{T_i \leqslant C_i}$.
- By définition of Δ_i,

$$\begin{aligned}\mathbb{E}[\Delta_i|C_i = c, Z_i = z] &= \mathbb{E}[1\!\!1_{T_i \leqslant C_i}|C_i = c, Z_i = z] \\ &= \mathbb{P}[T_i \leqslant c|Z_i = z] \\ &= F(c, z)\end{aligned}$$

- **Least-square contrat** : for $h : A \mapsto \mathbb{R}$ let

$$\gamma_n(h) = \frac{1}{n}\sum_{i=1}^{n} (\Delta_i - h(C_i, Z_i))^2$$

Least-square contrast

- Let $(C_i, Z_i, \Delta_i)_{i=1,\dots,n}$ the i.i.d. observation sample with $\Delta_i = 1\!\!1_{T_i \leqslant C_i}$.
- By définition of Δ_i,

$$\begin{aligned} \mathbb{E}[\Delta_i | C_i = c, Z_i = z] &= \mathbb{E}[1\!\!1_{T_i \leqslant C_i} | C_i = c, Z_i = z] \\ &= \mathbb{P}[T_i \leqslant c | Z_i = z] \\ &= F(c, z) \end{aligned}$$

- **Least-square contrat** : for $h : A \mapsto \mathbb{R}$ let

$$\gamma_n(h) = \frac{1}{n} \sum_{i=1}^{n} (\Delta_i - h(C_i, Z_i))^2 \quad \Rightarrow \quad \mathbb{E}[\gamma_n(h)] = \int (h - F)^2(t, z) f_{(C,Z)}(t, z) dt dz + C$$

Least-square contrast

- Let $(C_i, Z_i, \Delta_i)_{i=1,\dots,n}$ the i.i.d. observation sample with $\Delta_i = \mathbb{1}_{T_i \leqslant C_i}$.
- By définition of Δ_i,

$$\begin{aligned}
\mathbb{E}[\Delta_i | C_i = c, Z_i = z] &= \mathbb{E}[\mathbb{1}_{T_i \leqslant C_i} | C_i = c, Z_i = z] \\
&= \mathbb{P}[T_i \leqslant c | Z_i = z] \\
&= F(c, z)
\end{aligned}$$

- **Least-square contrat** : for $h : A \mapsto \mathbb{R}$ let

$$\gamma_n(h) = \frac{1}{n} \sum_{i=1}^{n} (\Delta_i - h(C_i, Z_i))^2 \quad \Rightarrow \quad \mathbb{E}[\gamma_n(h)] = \int (h - F)^2 (t, z) f_{(C,Z)}(t, z) dt dz + C$$

- Then

$$\widehat{F} = \arg\min_{h \in \mathcal{F}} \gamma_n(h) \quad \rightsquigarrow \quad \text{Choice of } \mathcal{F} \text{ ?}$$

Least-square contrast

- Let $(C_i, Z_i, \Delta_i)_{i=1,\ldots,n}$ the i.i.d. observation sample with $\Delta_i = 1\!\!1_{T_i \leqslant C_i}$.
- By définition of Δ_i,

$$\begin{aligned} \mathbb{E}[\Delta_i | C_i = c, Z_i = z] &= \mathbb{E}[1\!\!1_{T_i \leqslant C_i} | C_i = c, Z_i = z] \\ &= \mathbb{P}[T_i \leqslant c | Z_i = z] \\ &= F(c, z) \end{aligned}$$

- **Least-square contrat** : for $h : A \mapsto \mathbb{R}$ let

$$\gamma_n(h) = \frac{1}{n} \sum_{i=1}^{n} (\Delta_i - h(C_i, Z_i))^2 \quad \Rightarrow \quad \mathbb{E}[\gamma_n(h)] = \int (h - F)^2(t, z) f_{(C,Z)}(t, z) dt dz + C$$

- Then

$$\widehat{F} = \arg\min_{h \in \mathcal{F}} \gamma_n(h) \quad \rightsquigarrow \quad \text{Choice of } \mathcal{F} \text{ ?}$$

- **Model selection** :
  - build a collection of estimators on finite dimensional linear subspaces of $L^2(A)$ called *models*
  - select a model by a data-driven criterion.

General context

- Consider an i.i.d. observation sample of size n
- Let $F \in L^2(I)$ with $I \subset \mathbb{R}$ a function related to the distribution of the sample.

General context

- Consider an i.i.d. observation sample of size n
- Let $F \in L^2(I)$ with $I \subset \mathbb{R}$ a function related to the distribution of the sample.
- Consider a semi-norm $\| \cdot \|_0$ on $L^2(I)$ so that we want to control $\mathbb{E}[\|\hat{F} - F\|_0^2]$.

General context

- Consider an i.i.d. observation sample of size n
- Let $F \in L^2(I)$ with $I \subset \mathbb{R}$ a function related to the distribution of the sample.
- Consider a semi-norm $\| \cdot \|_0$ on $L^2(I)$ so that we want to control $\mathbb{E}[\|\hat{F} - F\|_0^2]$.
- Consider an empirical contrast $\gamma_n : L^2(I) \mapsto \mathbb{R}$ such that

$$\mathbb{E}[\gamma_n(h)] = \|F - h\|_0^2 + cte$$

with cte independent of h.

Collection of estimators

- Consider a collection $\mathcal{M}_n = \{S_m, m \in I_n\}$ of finite dimensional linear subpaces $L^2(I)$ called **models** :

$$S_m = \text{vect}\{\phi_1^m, \ldots, \phi_{D_m}^m\} = \left\{ h = \sum_{k=1}^{D_m} a_k \phi_k^m, (a_1, \ldots, a_{D_m}) \in \mathbb{R}^{D_m} \right\}$$

$\hookrightarrow$ histogram, wavelets...

Collection of estimators

- Consider a collection $\mathcal{M}_n = \{S_m, m \in I_n\}$ of finite dimensional linear subpaces $L^2(I)$ called **models** :

$$S_m = \text{vect}\{\phi_1^m, \ldots, \phi_{D_m}^m\} = \left\{ h = \sum_{k=1}^{D_m} a_k \phi_k^m, (a_1, \ldots, a_{D_m}) \in \mathbb{R}^{D_m} \right\}$$

$\hookrightarrow$ histogram, wavelets...

- Once the collection $\mathcal{M}_n$ fixed, "model" refers to either m or S_m.

Collection of estimators

- Consider a collection $\mathcal{M}_n = \{S_m, m \in I_n\}$ of finite dimensional linear subpaces $L^2(I)$ called **models** :

$$S_m = \text{vect}\{\phi_1^m, \dots, \phi_{D_m}^m\} = \left\{ h = \sum_{k=1}^{D_m} a_k \phi_k^m, (a_1, \dots, a_{D_m}) \in \mathbb{R}^{D_m} \right\}$$

$\hookrightarrow$ histogram, wavelets...

- Once the collection $\mathcal{M}_n$ fixed, "model" refers to either m or S_m.
- For each $m \in I_n$, let

$$\widehat{F}_m = \arg\min_{h \in S_m} \gamma_n(h) = \sum_{k=1}^{D_m} \widehat{a}_k^m \phi_k^m$$

$\hookrightarrow$ For a given model S_m,the estimation of F reduces to estimate a finite number of parameters $(\hat{a}_k^m)$.

How to choose an estimator among the collection $\{\widehat{F}_m, m \in I_n\}$?

- Oracle : best model in the collection
$$m_{oracle} = \arg\min_{m \in I_n} \mathbb{E}\left[\|\widehat{F}_m - F\|_0^2\right]$$
- Oracle unknown (depends on the true function F)
- Idea : estimate $\mathbb{E}\left[\|\widehat{F}_m - F\|_0^2\right]$ up to a constant independent of m.

Bias-variance decomposition. Let $F_m = \arg\min_{h \in S_m} \|h - F\|_0^2 = \arg\min_{h \in S_m} \mathbb{E}[\gamma_n(h)]$ the $\|\cdot\|_0$-projection of F on S_m.

Thus $\langle \underbrace{\widehat{F}_m - F_m}_{\in S_m} , \underbrace{F_m - F}_{\in S_m^\perp} \rangle_0 = 0$ therefore

$$\mathbb{E}\left[\|\widehat{F}_m - F\|_0^2\right] = \|F_m - F\|_0^2 + \mathbb{E}\left[\|\widehat{F}_m - F_m\|_0^2\right]$$

Bias-variance decomposition. Let $F_m = \arg\min_{h \in S_m} \|h - F\|_0^2 = \arg\min_{h \in S_m} \mathbb{E}[\gamma_n(h)]$ the $\|\cdot\|_0$-projection of F on S_m.

$$\text{Thus} \quad \Big\langle \underbrace{\widehat{F}_m - F_m}_{\in S_m} \, , \, \underbrace{F_m - F}_{\in S_m^\perp} \Big\rangle_0 = 0 \quad \text{therefore}$$

$$\mathbb{E}\left[\|\widehat{F}_m - F\|_0^2 \right] = \underbrace{\|F_m - F\|_0^2}_{\substack{\text{bias = distance} \\ \text{between } F \text{ and } S_m}} + \underbrace{\mathbb{E}\left[\|\widehat{F}_m - F_m\|_0^2 \right]}_{\substack{\text{variance = cost of} \\ \text{estimating the model}}}$$

Bias-variance decomposition. Let $F_m = \arg\min_{h \in S_m} \|h - F\|_0^2 = \arg\min_{h \in S_m} \mathbb{E}[\gamma_n(h)]$ the $\|\cdot\|_0$-projection of F on S_m.

$$\text{Thus} \quad \Big\langle \underbrace{\widehat{F}_m - F_m}_{\in S_m} \ , \ \underbrace{F_m - F}_{\in S_m^\perp} \Big\rangle_0 = 0 \quad \text{therefore}$$

$$\mathbb{E}\left[\|\widehat{F}_m - F\|_0^2\right] = \underbrace{\|F_m - F\|_0^2}_{\text{bias = distance between } F \text{ and } S_m} + \underbrace{\mathbb{E}\left[\|\widehat{F}_m - F_m\|_0^2\right]}_{\text{variance = cost of estimating the model}}$$

- **Bias** $\|F_m - F\|_0^2 = \min_{h \in S_m} \mathbb{E}[\gamma_n(h)] + C$ estimated by $\gamma_n(\widehat{F}_m) = \min_{h \in S_m} \gamma_n(h)$.
- **Variance** $\mathbb{E}\left[\|\widehat{F}_m - F_m\|_0^2\right] \leqslant AD_m/n$

Bias-variance decomposition. Let $F_m = \arg\min_{h \in S_m} \|h - F\|_0^2 = \arg\min_{h \in S_m} \mathbb{E}[\gamma_n(h)]$ the $\|\cdot\|_0$-projection of F on S_m.

$$\text{Thus} \quad \langle \underbrace{\widehat{F}_m - F_m}_{\in S_m} , \underbrace{F_m - F}_{\in S_m^{\perp}} \rangle_0 = 0 \quad \text{therefore}$$

$$\mathbb{E}\left[\|\widehat{F}_m - F\|_0^2\right] = \underbrace{\|F_m - F\|_0^2}_{\substack{\text{bias = distance} \\ \text{between } F \text{ and } S_m}} + \underbrace{\mathbb{E}\left[\|\widehat{F}_m - F_m\|_0^2\right]}_{\substack{\text{variance = cost of} \\ \text{estimating the model}}}$$

- **Bias** $\|F_m - F\|_0^2 = \min_{h \in S_m} \mathbb{E}[\gamma_n(h)] + C$ estimated by $\gamma_n(\widehat{F}_m) = \min_{h \in S_m} \gamma_n(h)$.
- **Variance** $\mathbb{E}\left[\|\widehat{F}_m - F_m\|_0^2\right] \leqslant AD_m/n$
- **Model selection** :

$$\widehat{m} = \arg\min_{m \in I_n} \left\{\gamma_n(\widehat{F}_m) + pen(m)\right\} \quad \text{with} \quad pen(m) = \theta A \frac{D_m}{n}$$

$\hookrightarrow \theta > 1$ necessary to control the risk

Bias-variance decomposition. Let $F_m = \arg\min_{h\in S_m} \|h-F\|_0^2 = \arg\min_{h\in S_m} \mathbb{E}[\gamma_n(h)]$ the $\|\cdot\|_0$-projection of F on S_m.

$$\text{Thus} \quad \langle \underbrace{\widehat{F}_m - F_m}_{\in S_m} , \underbrace{F_m - F}_{\in S_m^\perp} \rangle_0 = 0 \quad \text{therefore}$$

$$\mathbb{E}\left[\|\widehat{F}_m - F\|_0^2\right] = \underbrace{\|F_m - F\|_0^2}_{\substack{\text{bias = distance} \\ \text{between } F \text{ and } S_m}} + \underbrace{\mathbb{E}\left[\|\widehat{F}_m - F_m\|_0^2\right]}_{\substack{\text{variance = cost of} \\ \text{estimating the model}}}$$

- **Bias** $\|F_m - F\|_0^2 = \min_{h\in S_m} \mathbb{E}[\gamma_n(h)] + C$ estimated by $\gamma_n(\widehat{F}_m) = \min_{h\in S_m} \gamma_n(h)$.
- **Variance** $\mathbb{E}\left[\|\widehat{F}_m - F_m\|_0^2\right] \leqslant AD_m/n$
- **Model selection** :

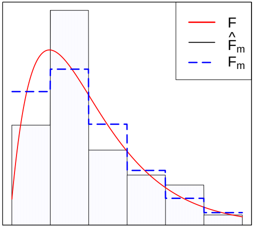
$$\widehat{m} = \arg\min_{m\in I_n} \left\{\gamma_n(\widehat{F}_m) + pen(m)\right\} \quad \text{with} \quad pen(m) = \theta A \frac{D_m}{n}$$

$\hookrightarrow \theta > 1$ necessary to control the risk

- **Model selection estimator** : $\boxed{\widehat{F}_{\widehat{m}}}$

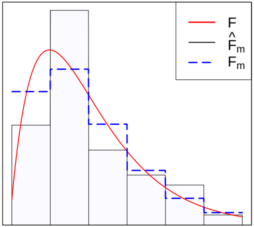
Bias and variance

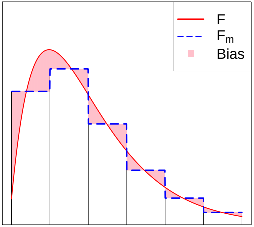
Ex : density estimation from an i.i.d. sample by histogram with 6 bins ($D_m = 6$)



Bias and variance

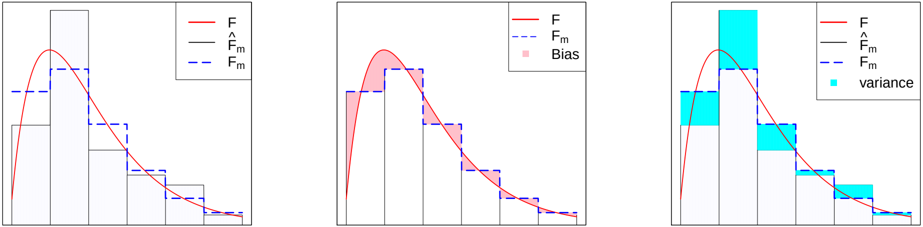
Ex : density estimation from an i.i.d. sample by histogram with 6 bins ($D_m = 6$)





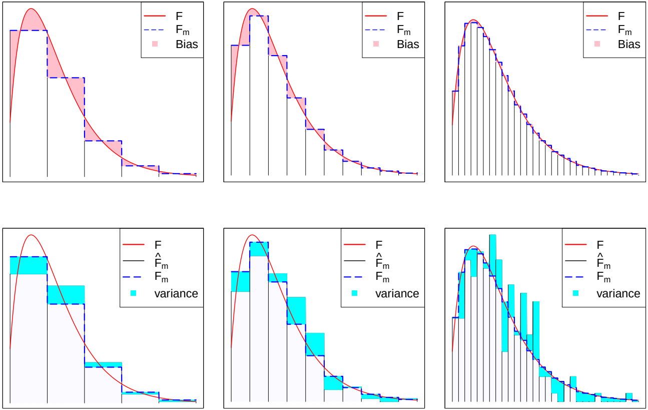
Bias and variance

Ex : density estimation from an i.i.d. sample by histogram with 6 bins ($D_m = 6$)



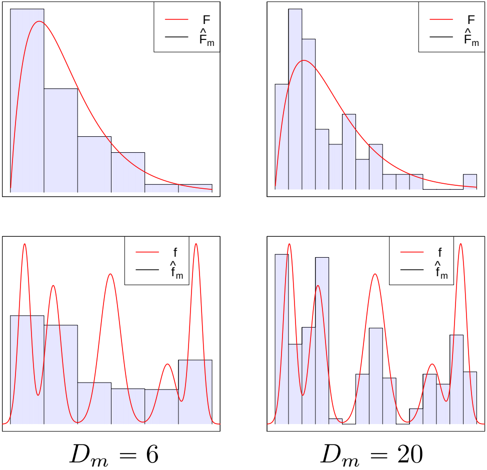
Bias-variance compromise. As the dimension D_m increases

- ◇ Bias $\|F_m - F_m\|_0^2$ decreases
- ◇ Variance $\mathbb{E}\left[\|\widehat{F}_m - F_m\|_0^2\right] \leqslant AD_m/n$ increases



$D_m = 5$ $D_m = 10$ $D_m = 30$

The dimension of the optimal model increases as the regularity of F decreases.



← More smooth F

← Less smooth F

Result : oracle inequality

The risk of the model selection estimator $\widehat{F}_{\widehat{m}}$ satisfies :

$$\mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F\|_0^2\right] \leqslant C_0 \inf_{m \in I_n} \left\{\|F - F_m\|_0^2 + A\frac{D_m}{n}\right\} \quad (1)$$

- Proof based on concentration inequalities [Talagrand]
- Optimality among the collection of estimators
- AD_m/n is only an upper-bound of the variance
- More general optimality : minimax rate.

Rate of convergence over regularity classes

- Consider **classes of regularity** $\mathcal{H}^\beta$ s.t. for suitable approximation spaces S_m :

$$\inf_{h \in S_m} \|h - F\|_0 \leqslant C_0 D_m^{-\beta} \quad \forall F \in \mathcal{H}^\beta$$

Rate of convergence over regularity classes

- Consider **classes of regularity** $\mathcal{H}^\beta$ s.t. for suitable approximation spaces S_m :

$$\inf_{h \in S_m} \|h - F\|_0 \leqslant C_0 D_m^{-\beta} \quad \forall F \in \mathcal{H}^\beta$$

- If $F \in \mathcal{H}^\beta$, the bias/variance sum for a model S_m is upper-bounded by

$$\underbrace{D_m^{-2\beta}}_{\searrow \text{ as } D_m \nearrow} \qquad + \qquad \underbrace{\theta A \frac{D_m}{n}}_{\nearrow \text{ as } D_m \nearrow}$$

Rate of convergence over regularity classes

- Consider **classes of regularity** $\mathcal{H}^\beta$ s.t. for suitable approximation spaces S_m :

$$\inf_{h \in S_m} \|h - F\|_0 \leqslant C_0 D_m^{-\beta} \quad \forall F \in \mathcal{H}^\beta$$

- If $F \in \mathcal{H}^\beta$, the bias/variance sum for a model S_m is upper-bounded by

$$\underbrace{D_m^{-2\beta}}_{\searrow \text{ as } D_m \nearrow} \qquad + \qquad \underbrace{\theta A \frac{D_m}{n}}_{\nearrow \text{ as } D_m \nearrow}$$

- The **bias-variance compromise** is achieved with $D_m \propto n^{1/(2\beta+1)}$ and the rate of convergence = is $n^{-2\beta/(2\beta+1)}$

Minimax rate of convergence

- **Minimax lower bounds.** We prove that $n^{-2\beta/(2\beta+1)}$ is the **minimax rate of convergence** over $\mathcal{H}^\beta$ that is the rate of convergence of the best possible estimator in a given context (regression, current status) based on a $n-$sample for functions in $\mathcal{H}^\beta$:

$$\inf_{\widehat{F}_n} \sup_{F \in \mathcal{H}^\beta} \mathbb{E}\left[\|\widehat{F}_n - F\|_0^2\right] n^{2\beta/(2\beta+1)} \geqslant \kappa_1$$

$\hookrightarrow$ Computing based on maximum coverage of $\mathbb{R}^n$-balls.

Minimax rate of convergence

- **Minimax lower bounds.** We prove that $n^{-2\beta/(2\beta+1)}$ is the **minimax rate of convergence** over $\mathcal{H}^\beta$ that is the rate of convergence of the best possible estimator in a given context (regression, current status) based on a $n-$sample for functions in $\mathcal{H}^\beta$:

$$\inf_{\widehat{F}_n} \sup_{F\in\mathcal{H}^\beta} \mathbb{E}\left[\|\widehat{F}_n - F\|_0^2\right] n^{2\beta/(2\beta+1)} \geqslant \kappa_1$$

$\hookrightarrow$ Computing based on maximum coverage of $\mathbb{R}^n$-balls.

- Therefore, the model selection estimator is minimax

$$\kappa_1 \leqslant \inf_{\widehat{F}_n} \sup_{F\in\mathcal{C}^\beta} \mathbb{E}\left[\|\widehat{F}_n - F\|_0^2\right] n^{2\beta/(2\beta+1)} \leqslant \mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F\|_0^2\right] n^{2\beta/(2\beta+1)} \leqslant \kappa_0$$

Minimax rate of convergence

- **Minimax lower bounds.** We prove that $n^{-2\beta/(2\beta+1)}$ is the **minimax rate of convergence** over $\mathcal{H}^\beta$ that is the rate of convergence of the best possible estimator in a given context (regression, current status) based on a $n-$sample for functions in $\mathcal{H}^\beta$:

$$\inf_{\widehat{F}_n} \sup_{F \in \mathcal{H}^\beta} \mathbb{E}\left[\|\widehat{F}_n - F\|_0^2\right] n^{2\beta/(2\beta+1)} \geqslant \kappa_1$$

$\hookrightarrow$ Computing based on maximum coverage of $\mathbb{R}^n$-balls.

- Therefore, the model selection estimator is minimax

$$\kappa_1 \leqslant \inf_{\widehat{F}_n} \sup_{F \in \mathcal{C}^\beta} \mathbb{E}\left[\|\widehat{F}_n - F\|_0^2\right] n^{2\beta/(2\beta+1)} \leqslant \mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F\|_0^2\right] n^{2\beta/(2\beta+1)} \leqslant \kappa_0$$

- $\widehat{F}_{\widehat{m}}$ is called **adaptive** as it adapts to the unknown regularity of F.

Summary

- Define a collection of finite dimensional linear subspaces of L^2 called **models**
- Compute a **collection of estimators** by minimsation of a contrast over the collection of models
- Estimate the bias-variance sum of each estimator and select the model
- **Oracle inequality** : the risk of the selected model is smaller than the risk of the best model up to a multiplicative constante
- More general optimality : **minimax rate of convergence** over classes of regularity
- **Comment :** choice of the approximation basis used to build the collection of models $\mathcal{M}_m$
  - The nature of the basis affect the bias
  - Bases are associated to regularity classes which allow control of the bias (global/local regularity)
  - Choice may be guided by desired property of the estimator (differentiability, localisation on an interval...)

Reminder of the framework

- Consider $(T_i, Z_i)_{i=1,\dots,n}$ i.i.d with $T_i \in \mathbb{R}_+$, $Z_i \in \mathbb{R}$ and T_i **unobserved**.
- **Observations**

$$(C_i, \Delta_i = 1\!\!1_{T_i \leqslant C_i}, Z_i)_{i=1,\dots,n}$$

  with $(C_i)_{i=1,\dots,n}$ i.i.d. positive and $C_i \perp\!\!\!\perp T_i | Z_i$.
- We want to estimate the **conditional c.d.f.** of T_i

$$F(t,z) = \mathbb{P}[T_i \leqslant t | Z_i = z]$$

  on a compact $A = A_1 \times A_2 \subset \mathbb{R}_+ \times \mathbb{R}$.
- As $\mathbb{E}[\Delta_i | C_i, Z_i] = F(C_i, Z_i)$, we consider the **least square contrast**

$$\gamma_n(h) = \frac{1}{n} \sum_{i=1}^{n} \left(\Delta_i - h(C_i, Z_i)\right)^2, \quad h : A \mapsto \mathbb{R}$$

Collection of models on $A_1 \times A_2$

- Consider two collections of models on A_1 and A_2 :

$$\mathcal{M}_n^{(j)} = \{S_{m_j}^{(j)}, m_j \in I_n^{(j)}\}, \quad j = 1, 2 \quad \text{and}$$

$(\phi_k^{m_1})_{k=1,\dots,D_{m_1}^{(1)}}$ and $(\psi_k^{m_2})_{k=1,\dots,D_{m_2}^{(2)}}$ orthonormal basis of $S_{m_1}^{(1)}$ and $S_{m_2}^{(2)}$.

- Linear subspaces of $L^2(A_1 \times A_2)$ built as **tensor products** of the linear subspaces of $L^2(A_1)$ and $L^2(A_2)$. For every $m = (m_1, m_2)$, let

$$S_m = S_{m_1}^{(1)} \otimes S_{m_2}^{(2)} = \left\{ (t, z) \in A \mapsto \sum_{k=1}^{D_{m_1}} \sum_{\ell=1}^{D_{m_2}} a_{k,\ell}^m \phi_k^{m_1}(t) \psi_\ell^{m_2}(z),\ (a_{k,\ell}^m)_{k,\ell} \in \mathbb{R}^{D_m} \right\}$$

linear subspace of $L^2(A)$ of dimension $D_m = D_{m_1}^{(1)} D_{m_2}^{(2)}$

- Finally, the collection of model on $L^2(A)$ is

$$\mathcal{M}_n = \{S_m, m \in I_n = I_n^{(1)} \times I_n^{(2)}\}$$

Example : Let $A_1 = A_2 = [0, 1]$, and

$$\begin{cases} S_{m_1}^{(1)} = \text{regular histograms with } D_1 \text{ bins on } [0,1] = \text{vect}\left\{ 1\!\!1_{J_k^1}, k = 1, \dots, D_1 \right\} \\ S_{m_2}^{(2)} = \text{regular histograms with } D_2 \text{ bins on } [0,1] = \text{vect}\left\{ 1\!\!1_{J_\ell^2}, \ell = 1, \dots, D_2 \right\} \end{cases}$$

Then

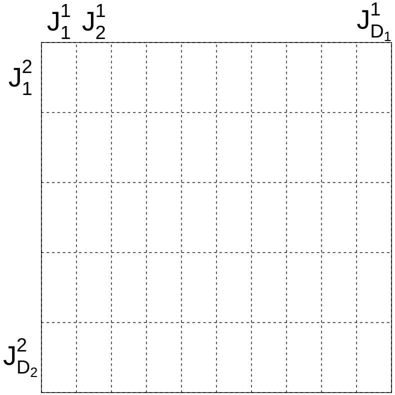
$$S_{m_1}^{(1)} \otimes S_{m_2}^{(2)} = \text{vect}\left\{ 1\!\!1_{J_k^1 \times J_\ell^2}, k = 1, \dots, D_1, \ell = 1, \dots, D_1 \right\}$$

Example : Let $A_1 = A_2 = [0,1]$, and

$$\begin{cases} S^{(1)}_{m_1} = \text{regular histograms with } D_1 \text{ bins on } [0,1] = \text{vect}\left\{ 1\!\!1_{J^1_k}, k = 1, \ldots, D_1 \right\} \\ S^{(2)}_{m_2} = \text{regular histograms with } D_2 \text{ bins on } [0,1] = \text{vect}\left\{ 1\!\!1_{J^2_\ell}, \ell = 1, \ldots, D_2 \right\} \end{cases}$$

Then

$$S^{(1)}_{m_1} \otimes S^{(2)}_{m_2} = \text{vect}\left\{ 1\!\!1_{J^1_k \times J^2_\ell}, k = 1, \ldots, D_1, \ell = 1, \ldots, D_1 \right\}$$

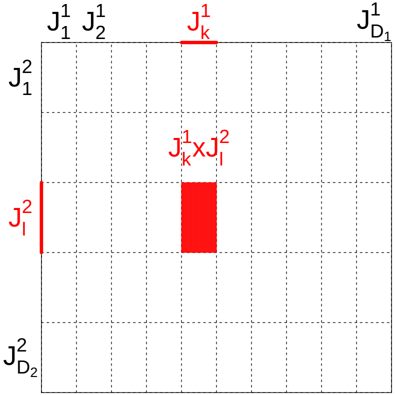


Example : Let $A_1 = A_2 = [0,1]$, and

$$\left\{ \begin{array}{l} S^{(1)}_{m_1} = \text{regular histograms with } D_1 \text{ bins on } [0,1] = \text{vect}\left\{ 1\!\!1_{J^1_k}, k = 1, \ldots, D_1 \right\} \\ S^{(2)}_{m_2} = \text{regular histograms with } D_2 \text{ bins on } [0,1] = \text{vect}\left\{ 1\!\!1_{J^2_\ell}, \ell = 1, \ldots, D_2 \right\} \end{array} \right.$$

Then

$$S^{(1)}_{m_1} \otimes S^{(2)}_{m_2} = \text{vect}\left\{ 1\!\!1_{J^1_k \times J^2_\ell}, k = 1, \ldots, D_1, \ell = 1, \ldots, D_1 \right\}$$



Least square estimators

$$\gamma_n(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\Delta_i - h(C_i, Z_i)\right)^2$$

- For every $S_m \in \mathcal{M}_n$,

$$\widehat{F}_m = \arg\min_{h \in S_m} \gamma_n(h)$$

- $\widehat{F}_m$ is **uniquely defined on the observations design** $(\mathbf{C}, \mathbf{Z})$

[notation : $\mathbf{C} = (C_1, \dots, C_n)$ and $\mathbf{\Delta} = (\Delta_1, \dots, \Delta_n)$]

- **Collection of estimators** $\{\widehat{F}_m, m \in I_n\}$.

We consider two risks to quantify the distance between F and an estimator $\widehat{F}_m$

- **The empirical risk :** we show that

$$\mathbb{E}[\gamma_n(h)|\mathbf{C},\mathbf{Z}] = \|h-F\|_n^2 + cte \quad \text{with} \quad \|h_0\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}(h_0(C_i,Z_i))^2$$

  thus we consider the risk

$$\mathbb{E}\left[\|\widehat{F}_m - F\|_n^2|\mathbf{C},\mathbf{Z}\right]$$

  $\hookrightarrow$ Evaluate the quality of estimation at the observations : naturally arise in least-square

- **The integrated risk**

$$\mathbb{E}\left[\|\widehat{F}_m - F\|^2\right] \quad \text{with} \quad \|.\| \text{ the } L^2\text{-norm}$$

  $\hookrightarrow$ More general control
  $\hookrightarrow$ Requires additional assumption to control the behaviour of the function out of the observations.

We will first state an upper bound for empirical risk, then derive the result for the L^2-risk.

Model selection

- **Bias-variance decomposition** for the empirical risk : let $F_m = \arg\min_{h \in S_m} \|F - h\|_n^2$,

$$\mathbb{E}\left[\|\widehat{F}_m - F\|_n^2 | \mathbf{Z}, \mathbf{C}\right] = \|F_m - F\|_n^2 + \mathbb{E}\left[\|\widehat{F}_m - F_m\|_n^2 | \mathbf{Z}, \mathbf{C}\right]$$

- Bias estimated by $\gamma_n(\widehat{F}_m)$.
- Variance : as Δ_i has Bernoulli distribution

$$\mathbb{E}\left[\|\widehat{F}_m - F_m\|_n^2 | \mathbf{Z}, \mathbf{C}\right] \leqslant \frac{1}{4}\frac{D_m}{n} \quad \text{with} \quad D_m = D_{m_1}^{(1)} D_{m_2}^{(2)}$$

- **Model selection :** let

$$\widehat{m} = \arg\min_{m \in I_n} \left\{\gamma_n(\widehat{F}_m) + pen(m)\right\} \quad \text{with} \quad pen(m) = \frac{\theta}{4}\frac{D_m}{n}, \theta > 1$$

Oracle inequality for the empirical risk

Theorem

Assume that for every $b > 0$, *for* $j = 1, 2$, *there exists* $B_j(b)$ *s.t.*

$$\sum_{m \in I_n^{(j)}} \exp\left(-b\sqrt{D_{m_j}^{(j)}}\right) \leqslant B_j \qquad (\mathbf{H_0})$$

Then, there exists constants C_1 *and* C_2 *such that,*

$$\mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F\|_n^2 | \mathbf{C}, \mathbf{Z}\right] \leqslant C_1 \inf_{m \in I_n} \left\{ \inf_{h \in S_m} \|F - h\|_n^2 + \frac{D_m}{n} \right\} + \frac{C_2}{n}$$

- The model selection estimator realises the bias-variance compromise
- There is no assumptions on the distributions of C and Z.
- The result holds for non-random $\mathbf{Z}$ and/or $\mathbf{C}$.

Integrated risk : additional assumptions

$(\mathbf{H_1})$ (C, Z) has a density and there exists $0 < h_0 \leqslant h_1 < \infty$ s.t.

$$h_0 \leqslant f_{(C,Z)}(t, z) \leqslant h_1, \quad \forall (t, z) \in A$$

$\hookrightarrow$ guarantees sufficiently dense observations on the estimation set A and the equivalence between norms $\| \cdot \|$ and $\| \cdot \|_{f_{(\mathbf{C},\mathbf{Z})}} = \mathbb{E}[\| \cdot \|_n^2]$

$(\mathbf{H_2})$ Restriction of the number of models in the collection $\mathcal{M}_n$ and

$$\max_{m \in \mathcal{M}_n} D_m = \max_{m_1 \in \mathcal{M}_n^{(1)}, m_2 \in \mathcal{M}_n^{(2)}} D_{m_1}^{(1)} D_{m_2}^{(2)} \leqslant \sqrt{n}/\log(n)$$

$(\mathbf{H_3})$ Assumptions related to the nature of the models.
$\hookrightarrow$ satisfied for classic collections (piecewise polynomials, wavelet, trigonometric basis...)

Oracle inequality for the integrated risk

Corollary

Assume that $(\mathbf{H_0})$, $(\mathbf{H_1})$, $(\mathbf{H_2})$ *and* $(\mathbf{H_3})$ *hold, there exists constants* C_1' *and* C_2' *such that,*

$$\mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F\|^2\right] \leqslant C_1' \inf_{m \in I_n} \left\{ \inf_{h \in S_m} \|F - h\|^2 + \frac{D_m}{n} \right\} + \frac{C_2'}{n}$$

- Optimality over the collection of estimator up to a multiplicative constant.
- Optimality in a more general sense ? Minimax bound over classes of regularity

Rate of convergence over classes of regularity

Anisotropic Besov balls $\mathcal{B}^{\beta}_{2,\infty}(L)$ with $\beta = (\beta_1, \beta_2) \in (\mathbb{R}^*_+)^2$: generalisation of the function $\mathcal{C}^{(\beta_1,\beta_2)}$ with square integrable derivative.

Lemma

Assume that the $S^{(j)}_m$ are generated from either :

- *piecewise polynomials*
- *wavelets*
- *trigonometric polynomials*

Then there exists a constant $C_0(L)$ s.t. for all $F \in \mathcal{B}^{\beta}_{2,\infty}(L)$,

$$\inf_{h \in S_m} \|F - h\|^2 \leqslant C_0 \left((D^{(1)}_{m_1})^{-\beta_1} + (D^{(2)}_{m_2})^{-\beta_2} \right)$$

Corollary

Assume that $F \in \mathcal{B}^{\beta}_{2,\infty}(L)$ *with* $\beta_1, \beta_2 \geqslant 1$. *The bias-variance trade-off is obtained with* $\overline{m}$ *s.t.*

$$D^{(1)}_{\overline{m}_1} \propto n^{\beta_2/(\beta_1+\beta_2+2\beta_1\beta_2)} \quad \text{and} \quad D^{(2)}_{\overline{m}_2} \propto n^{\beta_1/(\beta_1+\beta_2+2\beta_1\beta_2)}$$

and the L^2*-risk of the model selection estimator is upper bounded by*

$$\mathbb{E}\left[\|\widehat{F}_{\widehat{m}} - F\|^2\right] \leqslant Cn^{-\overline{\beta}/(\overline{\beta}+1)}$$

with $\overline{\beta} = 2\beta_1\beta_2/(\beta_1 + \beta_2)$ *the harmonic mean.*

Comments

- The dimensions of the optimal model depends on the regularity of the function
- The estimator adapts to different regularities w.r.t. to the two variables.
- Assumption $(\mathbf{H_2})$ on the maximum dimension of models imposes a minimum regularity on F.
- The rate of convergence with respect to the time variable is not $1/n$ like in right-censoring framework.

Minimax rate of convergence

Theorem

Let $\beta \in (1, +\infty)^2$*, assume that* $(\mathbf{A_1^{sup}})$ *holds, then there exists a constant* $c(\beta, L, h_1)$ *s.t.*

$$\inf_{\widehat{F}_n} \sup_{F \in \mathcal{B}^{\beta}_{2,\infty}(L)} \mathbb{E}\left[n^{\overline{\beta}/(\overline{\beta}+1)} \|\widehat{F}_n - F\|^2\right] \geqslant c$$

Comments

- The model selection estimator is minimax over anisotropic Besov balls
- The infimum is taken over all possible estimators : more general result than oracle inequality

Improvement of the estimator $\widehat{F}_{\widehat{m}}$

- Restriction to [0,1]

$$\tilde{F}_m(x,y) = \begin{cases} 0 & \text{if } \widehat{F}_m(x,y) < 0 \\ \widehat{F}_m(x,y) & \text{if } 0 \leqslant \widehat{F}_m(x,y) \leqslant 1 \\ 1 & \text{if } \widehat{F}_m(x,y) > 1 \end{cases}$$

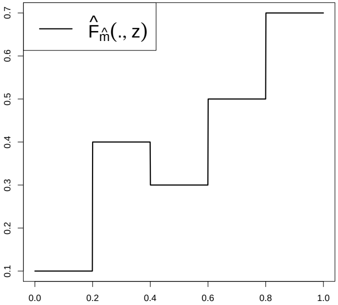
- For each z, $\widehat{F}^*_{\widehat{m}}(\cdot, z)$ increasing rearrangement of $\widehat{F}_{\widehat{m}}(\cdot, z)$

Improvement of the estimator $\widehat{F}_{\widehat{m}}$

- Restriction to [0,1]

$$\tilde{F}_m(x,y) = \begin{cases} 0 & \text{if } \widehat{F}_m(x,y) < 0 \\ \widehat{F}_m(x,y) & \text{if } 0 \leqslant \widehat{F}_m(x,y) \leqslant 1 \\ 1 & \text{if } \widehat{F}_m(x,y) > 1 \end{cases}$$

- For each z, $\widehat{F}^*_{\widehat{m}}(\cdot, z)$ increasing rearrangement of $\widehat{F}_{\widehat{m}}(\cdot, z)$

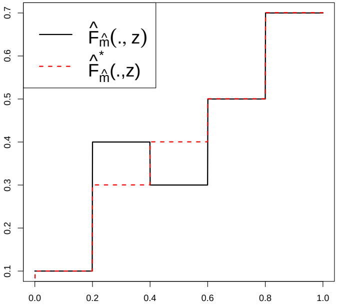


Improvement of the estimator $\widehat{F}_{\widehat{m}}$

- Restriction to [0,1]

$$\tilde{F}_m(x,y) = \begin{cases} 0 & \text{if } \widehat{F}_m(x,y) < 0 \\ \widehat{F}_m(x,y) & \text{if } 0 \leqslant \widehat{F}_m(x,y) \leqslant 1 \\ 1 & \text{if } \widehat{F}_m(x,y) > 1 \end{cases}$$

- For each z, $\widehat{F}^*_{\widehat{m}}(\cdot,z)$ increasing rearrangement of $\widehat{F}_{\widehat{m}}(\cdot,z)$

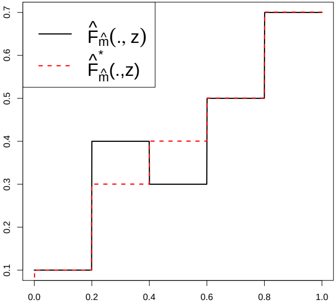


Improvement of the estimator $\widehat{F}_{\widehat{m}}$

- Restriction to [0,1]

$$\tilde{F}_m(x,y) = \begin{cases} 0 & \text{if } \widehat{F}_m(x,y) < 0 \\ \widehat{F}_m(x,y) & \text{if } 0 \leqslant \widehat{F}_m(x,y) \leqslant 1 \\ 1 & \text{if } \widehat{F}_m(x,y) > 1 \end{cases}$$

- For each z, $\widehat{F}^*_{\widehat{m}}(\cdot, z)$ increasing rearrangement of $\widehat{F}_{\widehat{m}}(\cdot, z)$



$\hookrightarrow$ Decreases the risk of the estimator.

1. Current status data

2. Model selection (univariate framework)

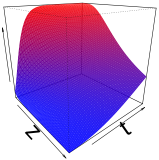
3. Adaptive estimation of the conditional c.d.f from current status data

4. Results on simulated data

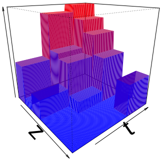
Distribution of the simulated data $(Z_i, T_i, C_i)_{i=1,\dots,n}$

$$\begin{cases} Z \sim \Gamma(k = 1.5, \theta = 2) \\ T = Z + \varepsilon \quad \text{with} \quad \varepsilon \sim \Gamma(k = 3, \theta = 2) \\ C = Z + \varepsilon' \quad \text{with} \quad \varepsilon' \sim \Gamma(k = 3, \theta = 2) \end{cases}$$

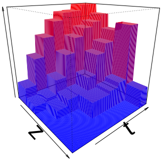
Model selection estimator of F with histogram models



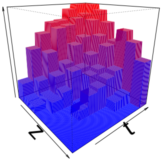
True F



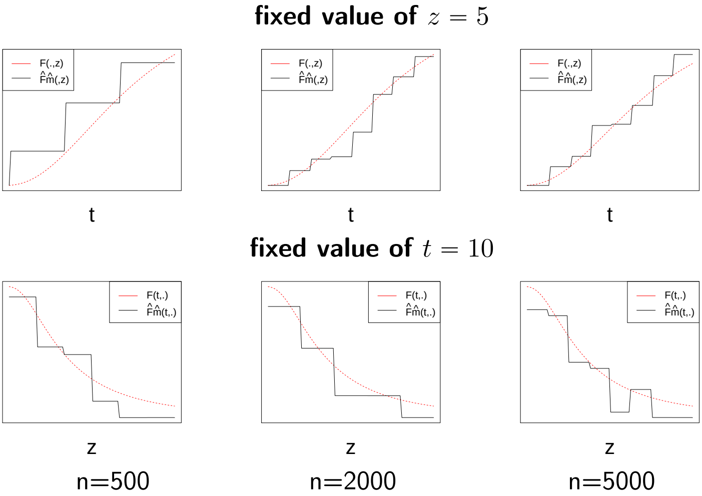
n=500



n=2000



n=5000



- Accurate estimation **require large sample size** due to
  - current status data : low informative
  - bi-dimensional setting without restrictive assumption : dimension curse.

Simulations (2)

- In right censoring, the **censoring rate**, defined as the proportion of unobserved times of event, impacts the quality of estimation.
  $\hookrightarrow$ Mean of the censoring rate depends on distribution of C and T.

Simulations (2)

- In right censoring, the **censoring rate**, defined as the proportion of unobserved times of event, impacts the quality of estimation.
  $\hookrightarrow$ Mean of the censoring rate depends on distribution of C and T.
- In current status dat, impact of distance between distributions of C and T.

  Heuristic : for a given sample size n and $Z = z$,

  - $\diamond$ F is more accurately estimated if observations $(C_i)_i$ are concentrated on area where F varies the most
  - $\diamond$ High variations of $F \Leftrightarrow$ high density of T
  - $\diamond$ Thus, estimation should improve as distance between densities f_C and f_T decreases.

Simulations (2)

- In right censoring, the **censoring rate**, defined as the proportion of unobserved times of event, impacts the quality of estimation.
  $\hookrightarrow$ Mean of the censoring rate depends on distribution of C and T.
- In current status dat, impact of distance between distributions of C and T.

  Heuristic : for a given sample size n and $Z = z$,

  - $\diamond$ F is more accurately estimated if observations $(C_i)_i$ are concentrated on area where F varies the most
  - $\diamond$ High variations of $F \Leftrightarrow$ high density of T
  - $\diamond$ Thus, estimation should improve as distance between densities f_C and f_T decreases.

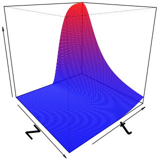
- **Simulations**

$$\begin{cases} Z \sim \Gamma(k = 1.5, \theta = 2) \\ T = a + Z \times \varepsilon \quad \text{with} \quad \varepsilon \sim \Gamma(k = 3, \theta = 2) \\ C = Z \times \varepsilon' \quad \text{with} \quad \varepsilon' \sim \Gamma(k = 3, \theta = 2) \end{cases}$$

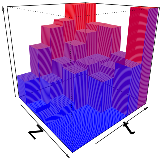
and the parameter a tunes the distance between $f_{T|Z}$ and $f_{C|Z}$.

$$dist(a) = \|f_C - f_T\|_{L^1}$$

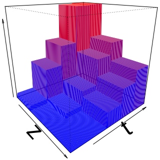
a	0	2	5	10
dist(a)	0	0.63	1.12	1.54



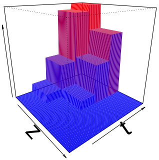
True F



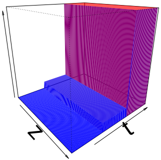
$a = 0$



$a = 2$



$a = 5$



$a = 10$

Summary

- Framework :

$$\begin{cases} T_i \text{ unobserved time of interest} \\ C_i \text{ observation time} \\ \Delta_i = 1\!\!1_{T_i \leqslant C_i} \text{ current status at time } C_i \\ Z_i \in \mathbb{R} \text{ covariate} \\ C_i \perp T_i | Z_i \end{cases}$$

- Least square contrast based on $\mathbb{E}[\Delta_i | C_i = c, Z_i = z] = F(c, z)$ with F the conditional c.d.f. of T_i.
- Model selection estimator
  - ◇ non-parametric estimation in finite dimensional spaces of functions called models.
  - ◇ Data driven-criterion to select a model by estimation of the bias-variance sum
- Oracle inequalities : the selected model realises the bias-variance trade-off
- Minimax optimality
- Require large sample size

Conclusion and perspectives

- Time and covariate treated the same : unusual
  $\hookrightarrow$ Minimax optimality validate the method
- Method valid for non-random observation times :
  $\hookrightarrow$ the distribution of C is not involved in the estimator (contrary to inverse probability weighted method e.g.)
  $\hookrightarrow$ for the control empirical risk control, no assumption on the covariate and observation time repartition
- Extension to covariates of dimension p : require very large sample size (limitation to uni-variate model of dimension $D_{m_j}^{(j)} \leqslant n^{1/2p}$)
- Generalisation to probabilistic discriminant classifier.
- Isotonic regression
- Simulations suggest impact of the distance between densities of T and C.