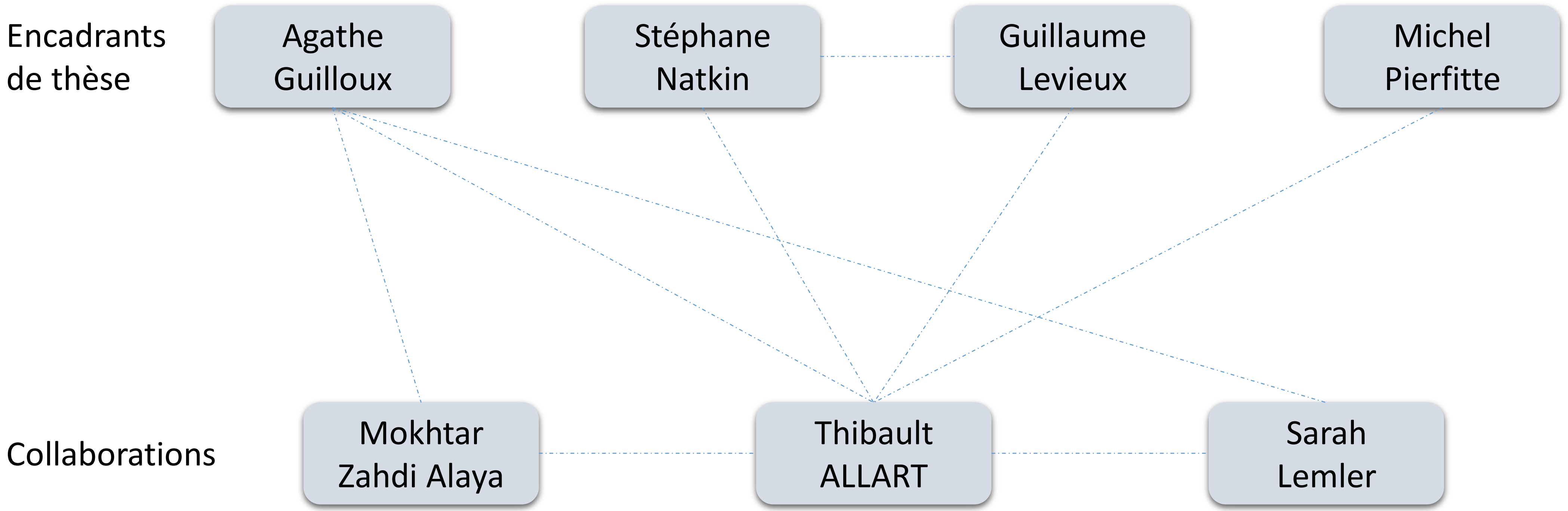
Analyse de données longitudinales : Application aux jeux vidéo

Young Statisticians and Probabilists

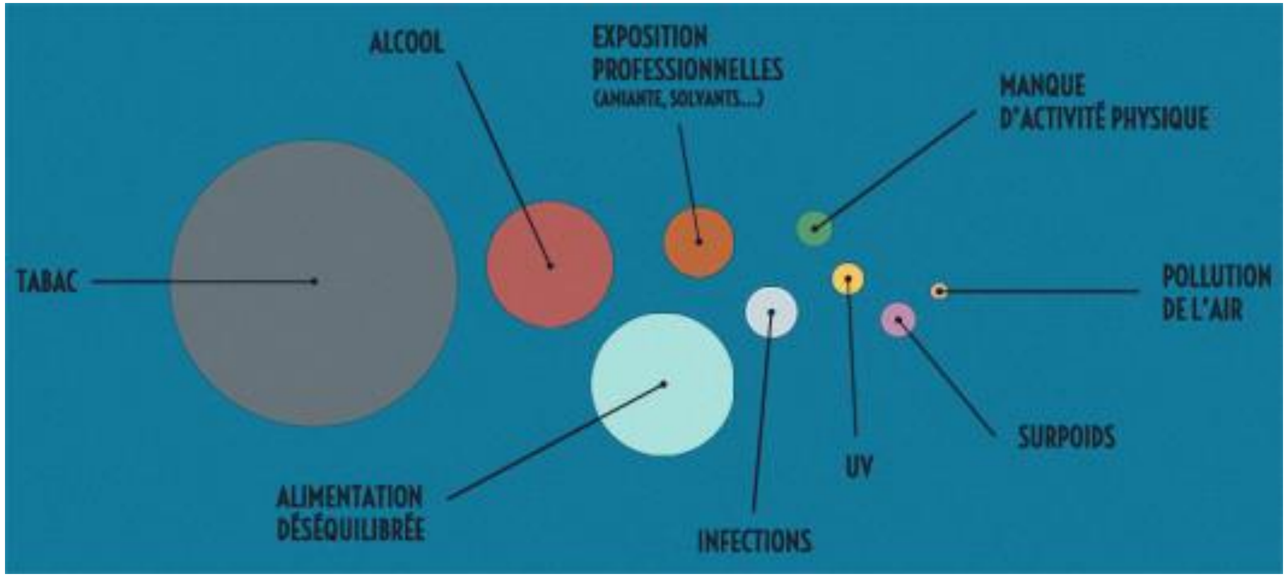
Thibault ALLART



Encadrants de thèse
Agathe Guilloux
Stéphane Natkin
Guillaume Levieux
Michel Pierfitte
Collaborations
Mokhtar Zahdi Alaya
Thibault ALLART
Sarah Lemler

On s’intéresse à l’influence de certains facteurs sur le temps d’apparition d’un évènement.

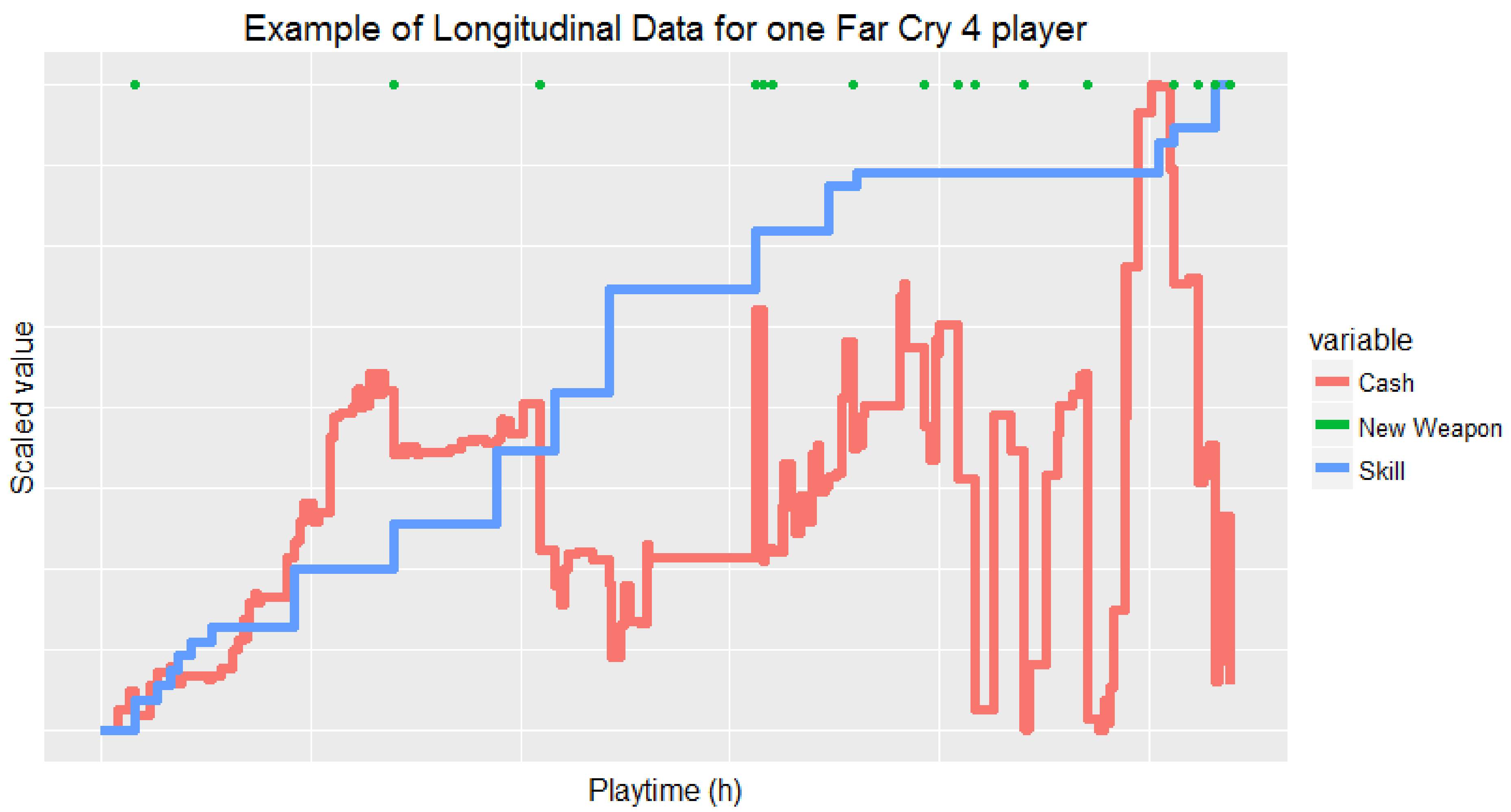
Poids des différents facteurs de risque de cancer



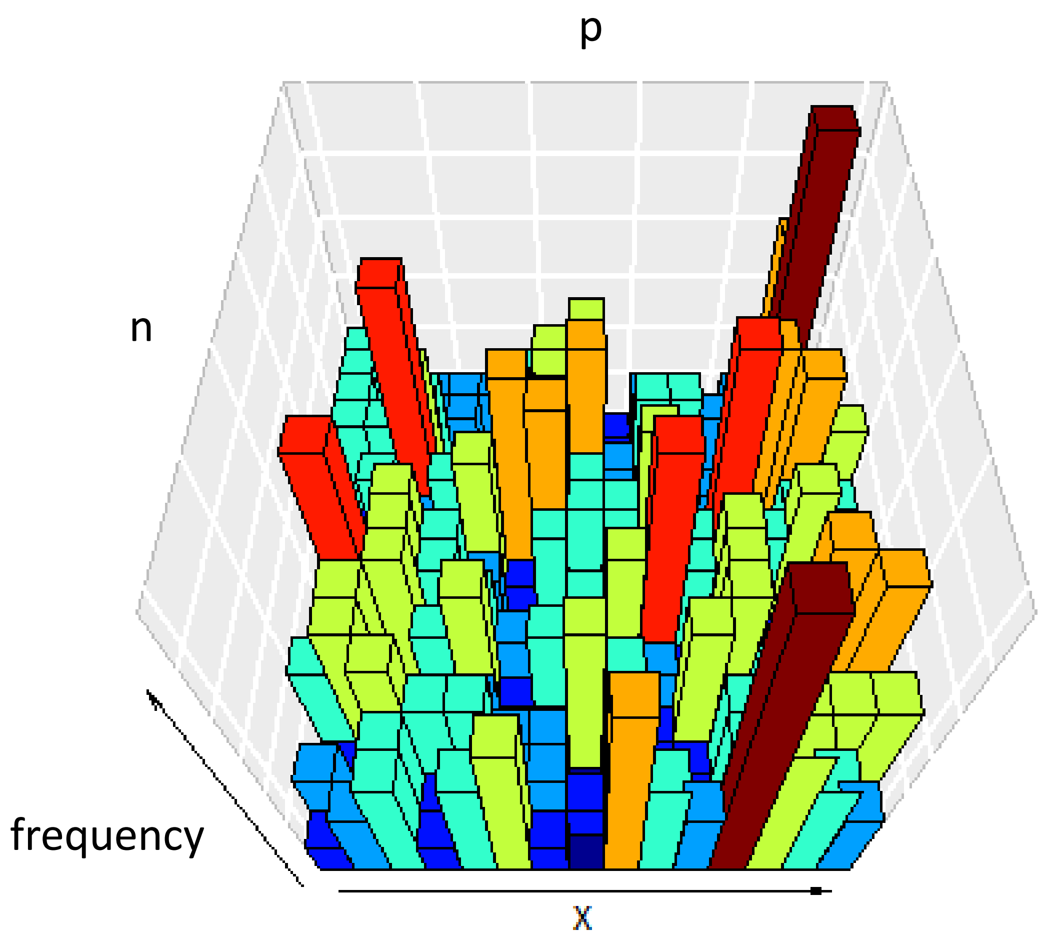
Source : institut national du cancer

Notre exposition change au cours du temps

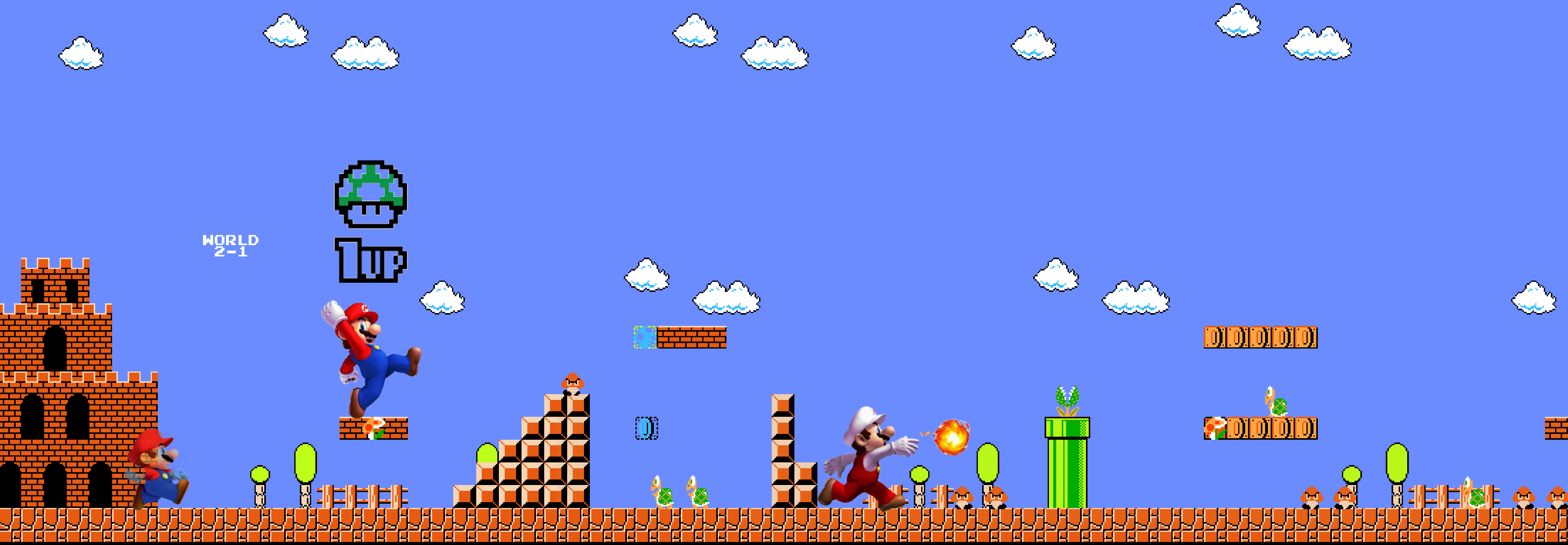
Longitudinal data



Example of Longitudinal Data for one Far Cry 4 player
Scaled value
Playtime (h)
variable
Cash
New Weapon
Skill



Observation frequency depends on individuals **and covariates**



Playtime

Why do we need Longitudinal data ?

Player 1

	Retention
Fire Mario	Correlated
Water	Correlated

Quits before the end

P1 Stop

P2 Stop

2h

30h

Playtime

Player 2

End of the game

Player 1

Quits before the end

	Retention	
	Aggregated Data	Time varying data
Fire Mario	Correlated	Not correlated
Water	Correlated	Correlated

P1 Stop

2h

P2 Stop

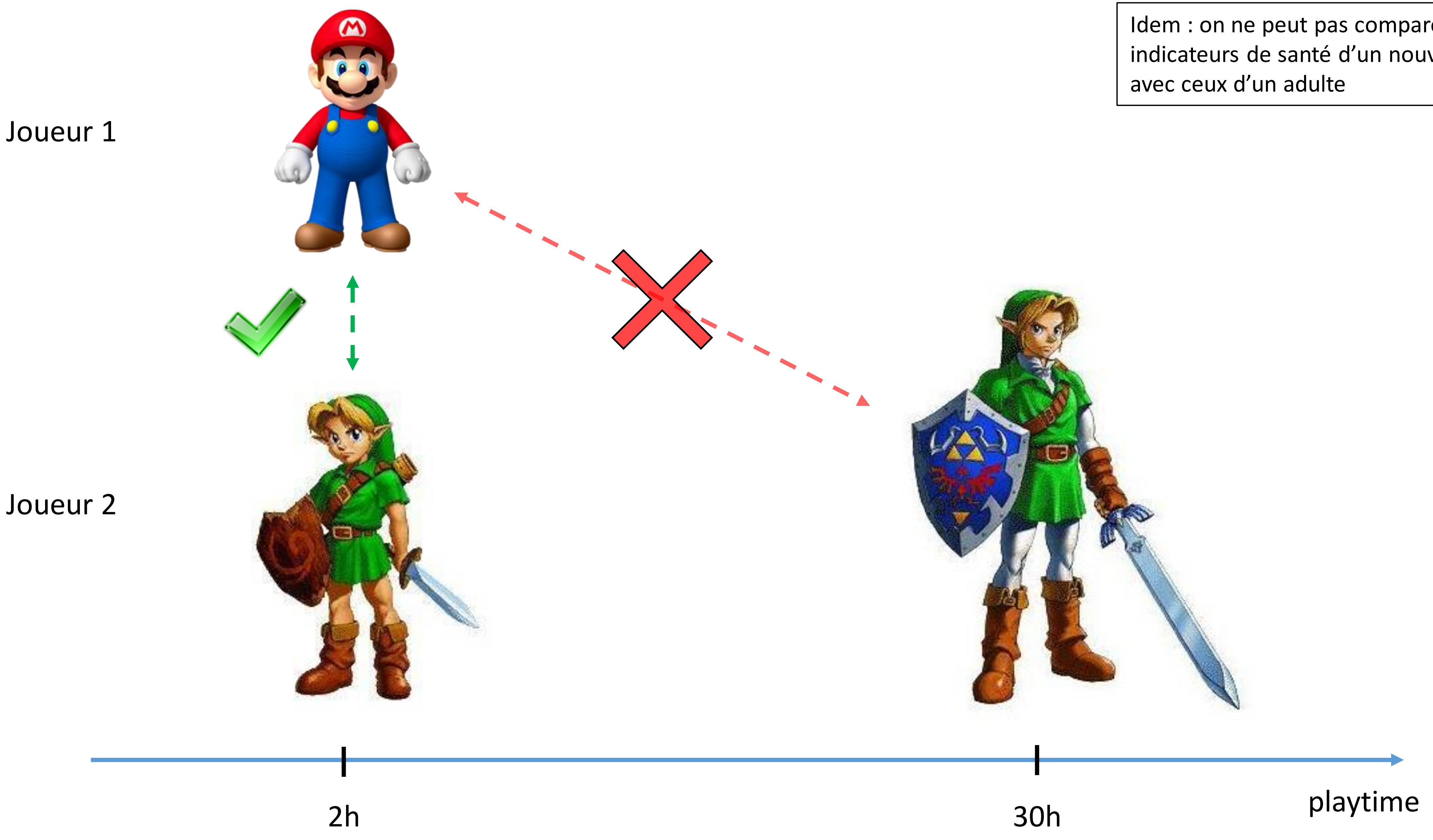
30h

Playtime

Player 2

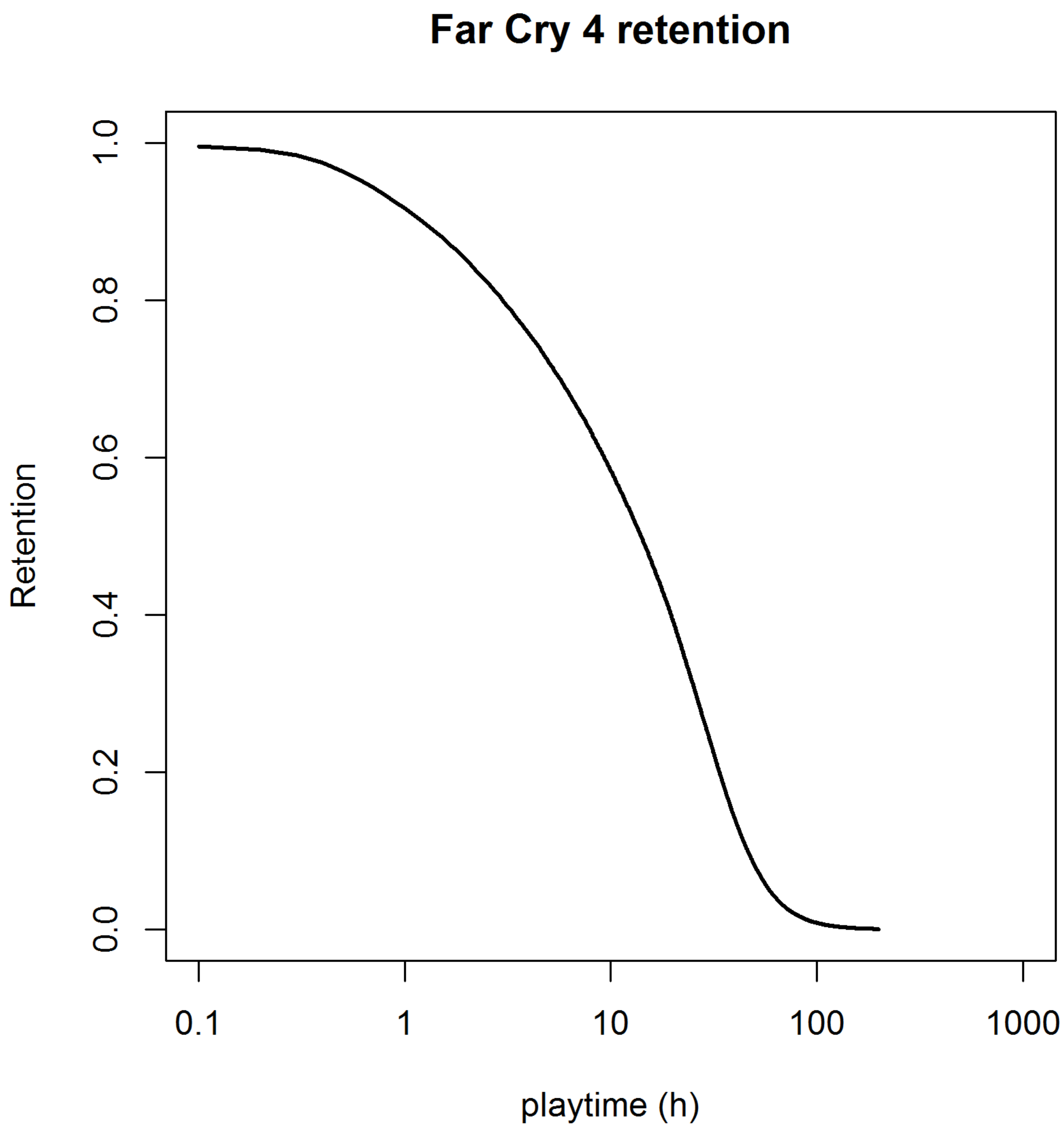
End of the game

Idem : on ne peut pas comparer les indicateurs de santé d’un nouveau né avec ceux d’un adulte



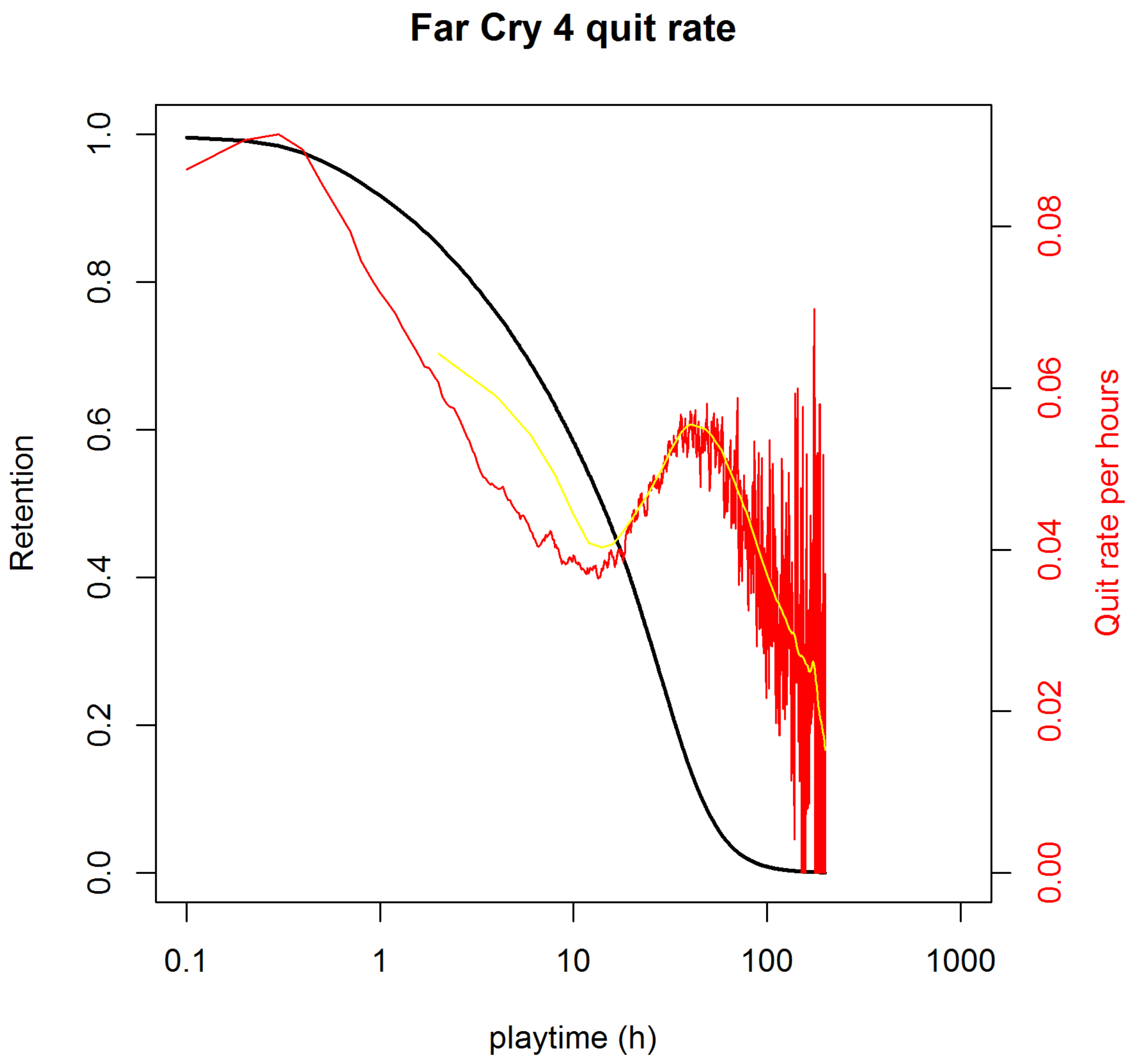
Modélisation

Analyse de survie



Soit T la variable aléatoire positive associée au temps de jeu (playtime) des joueurs.

$$S(t) = \mathbb{P}(\mathrm{T} > t)$$



Soit T la variable aléatoire positive associée au temps de jeu (playtime) des joueurs.

$$S(t) = \mathbb{P}(\mathrm{T} > t)$$

Taux de risqué instantané :

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

On cherche à modéliser $\lambda(t)$

Time-varying Cox proportional hazard model

Baseline

Effect of player behavior

$$\lambda(t|X_i(t)) = \lambda_0(t) \exp\left(\sum_{j=1}^{p} X_{i,j}(t)\beta_j(t)\right)$$

Quit rate

Player behavior

Comment estimer les coefficients du modèle ?

Estimation

Maximum de vraisemblance

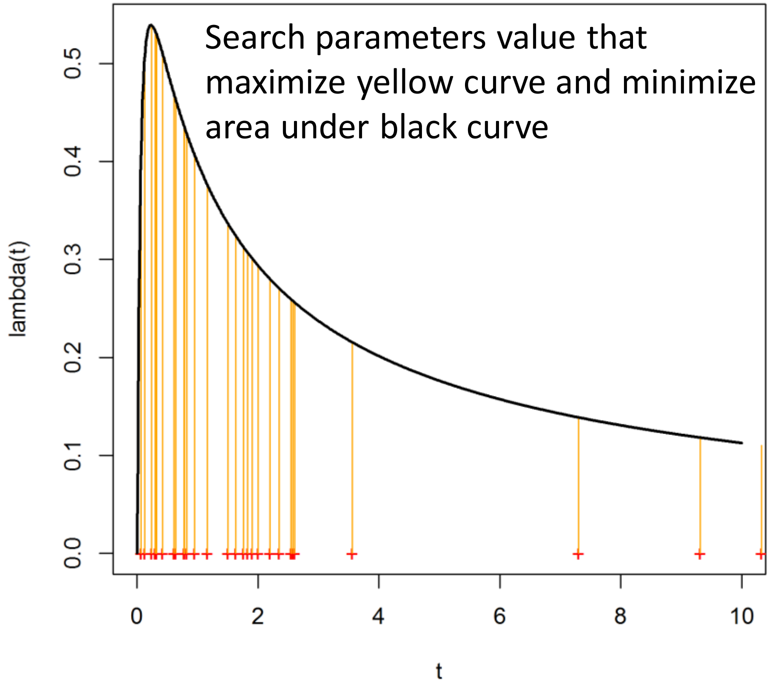
Likelihood

Soit t_i le temps de réalisation de l'évènement pour l'individu i et τ le temps de fin d'observation, alors la vraisemblance se décompose comme suit :

Probabilité que les évènements aient eu lieu au moment où on les a observés, conditionnellement au passé du processus

Probabilité qu'il n'y ait pas d'évènement aux autres temps, conditionnellement au passé du processus

$$\mathcal{L}_n(\beta) = \boxed{\left\{ \prod_{t_i \leq \tau} \lambda(t_i) \right\}} \boxed{S(\tau)}$$



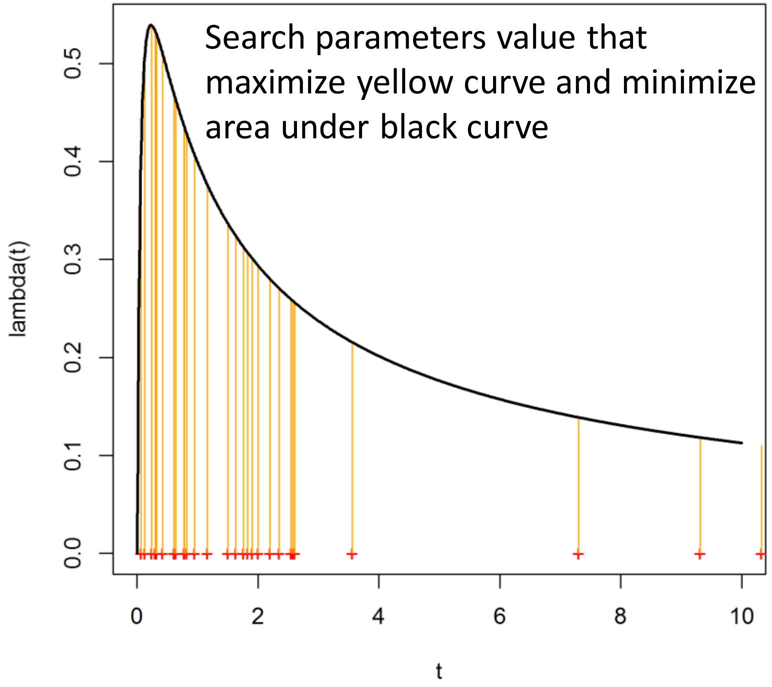
Likelihood

Soit τ_i le temps de réalisation de l'évènement pour l'individu i, alors la vraisemblance se décompose comme suit :

Probabilité que les évènements aient eu lieu au moment où on les a observés, conditionnellement au passé du processus

Probabilité qu'il n'y ait pas d'évènement aux autres temps, conditionnellement au passé du processus

$$\mathcal{L}_n(\beta) = \boxed{\left\{ \prod_{t_i \leq \tau} \lambda(t_i) \right\}} \boxed{\exp\left(- \int_0^\tau \lambda(s) ds \right)}$$



Likelihood

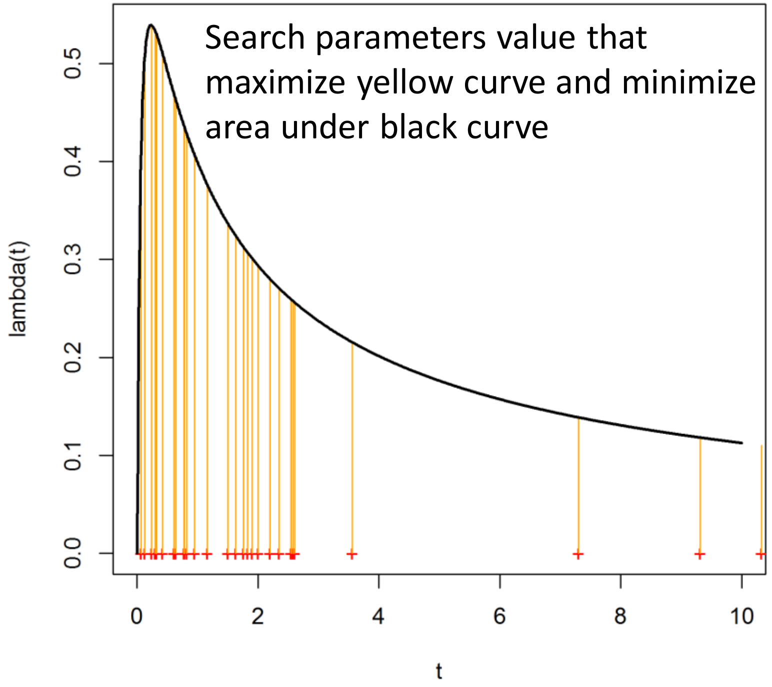
Soit τ_i le temps de réalisation de l'évènement pour l'individu i, alors la vraisemblance se décompose comme suit :

Probabilité que les évènements aient eu lieu au moment où on les a observés, conditionnellement au passé du processus

Probabilité qu'il n'y ait pas d'évènement aux autres temps, conditionnellement au passé du processus

$$\mathcal{L}_n(\beta) = \boxed{\left\{ \prod_{t_i \leq \tau} \lambda(t_i) \right\}} \boxed{\exp\left(- \int_0^\tau \lambda(s) ds \right)}$$

$$\mathcal{L}_n(\beta) = \left\{ \prod_{t_i \leq \tau} \exp\left(X^T(t_i)\beta(t_i) \right) \right\} \cdot \exp\left(- \int_0^\tau Y(s) \exp\left(X^T(s)\beta(s) \right) ds \right)$$



Likelihood

Soit τ_i le temps de réalisation de l'évènement pour l'individu i, alors la vraisemblance se décompose comme suit :

Probabilité que les évènements aient eu lieu au moment où on les a observés, conditionnellement au passé du processus

Probabilité qu'il n'y ait pas d'évènement aux autres temps, conditionnellement au passé du processus

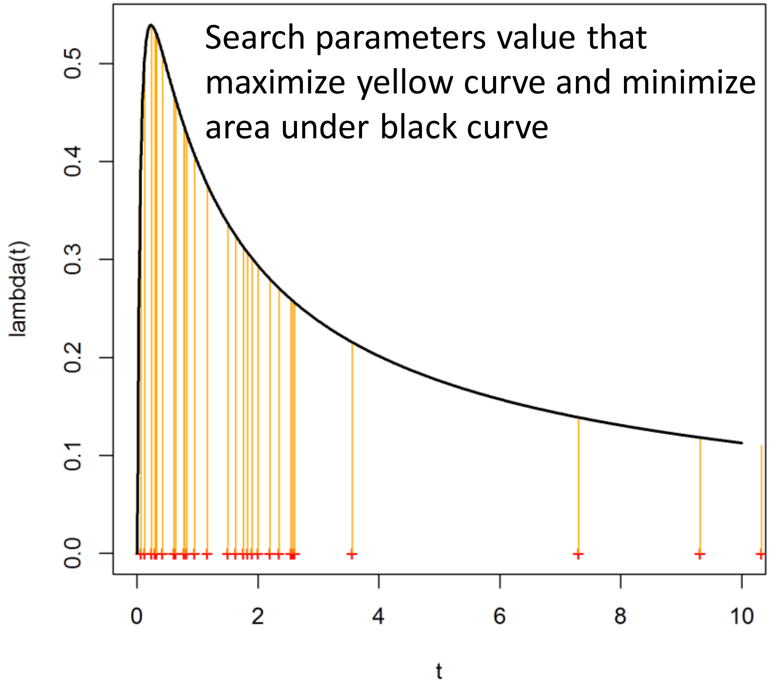
$$\mathcal{L}_n(\beta) = \left\{ \prod_{t_i \leq \tau} \lambda(t_i) \right\} \exp\left(- \int_0^\tau \lambda(s) ds \right)$$

$$\mathcal{L}_n(\beta) = \left\{ \prod_{t_i \leq \tau} \exp\left(X^T(t_i)\beta(t_i) \right) \right\} \cdot \exp\left(- \int_0^\tau Y(s) \exp\left(X^T(s)\beta(s) \right) ds \right)$$

La log vraisemblance s'écrit :

$$\ell_n(\beta) = \left\{ \sum_{t_i \leq \tau} X^T(t_i)\beta(t_i) \right\} - \int_0^\tau Y(s) \exp\left(X^T(s)\beta(s) \right) ds$$

$$\boxed{\ell_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \int_0^\tau X_i^T(t)\beta(t) dN_i(t) - \int_0^\tau Y_i(t) \exp\left(X_i^T(t)\beta(t) \right) dt \right\}}$$



- Integrale
- Quelle forme pour les $\beta(t)$?

Coefficients constants par morceaux

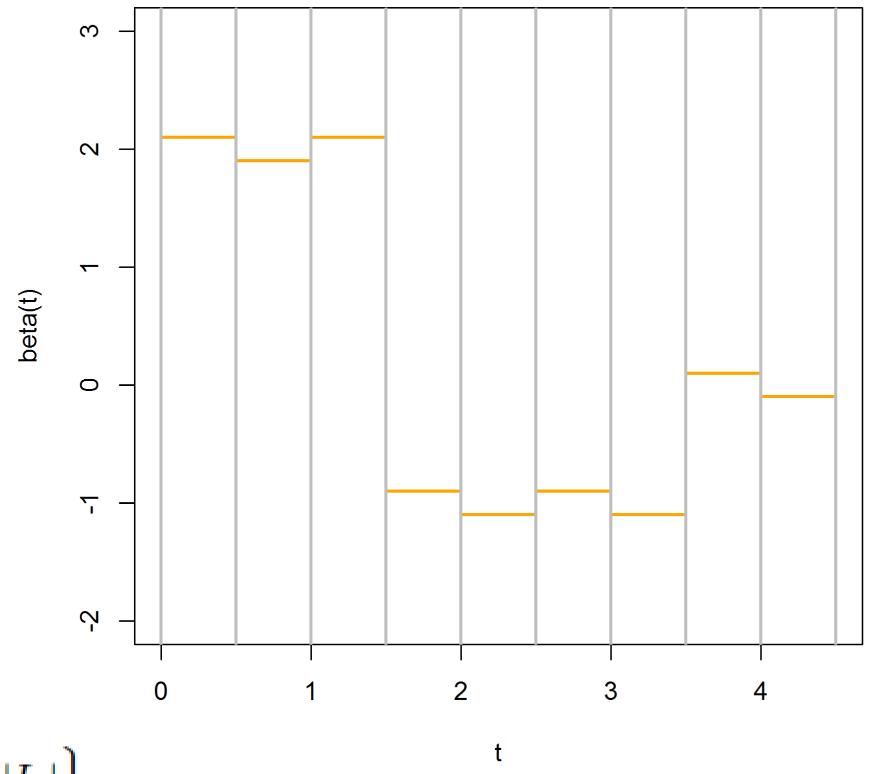
Soit $(I_l)_{l\in\{0,L\}}$ une partition de $[0,\tau]$.

$$\beta_j(t) = \sum_{l=1}^{L} \beta_{j,l} \mathbb{1}_{(I_l)}(t)$$

Et en utilisant le fait que les $X_{i,j}(t)$ sont constants sur de petits intervalles.

$$\ell_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left\{\int_0^\tau X_i^T(t)\beta(t)dN_i(t) - \int_0^\tau Y_i(t)\exp\left(X_i^T(t)\beta(t)\right)dt\right\}$$

$$\ell_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{L}\left\{\sum_{j=1}^{p}\sum_{s} X_{i,s}^{j}\beta_{j,l}N_i(I_s) - \sum_{s} Y_i(I_s)\exp\left(\sum_{j=1}^{p} X_{i,s}^{j}\beta_{j,l}\right)|I_s|\right\}$$



Pénalité

Pour éviter l'explosion du nombre de dimensions, on voudrait pénaliser le nombre de coefficients (ou le nombre de sauts).

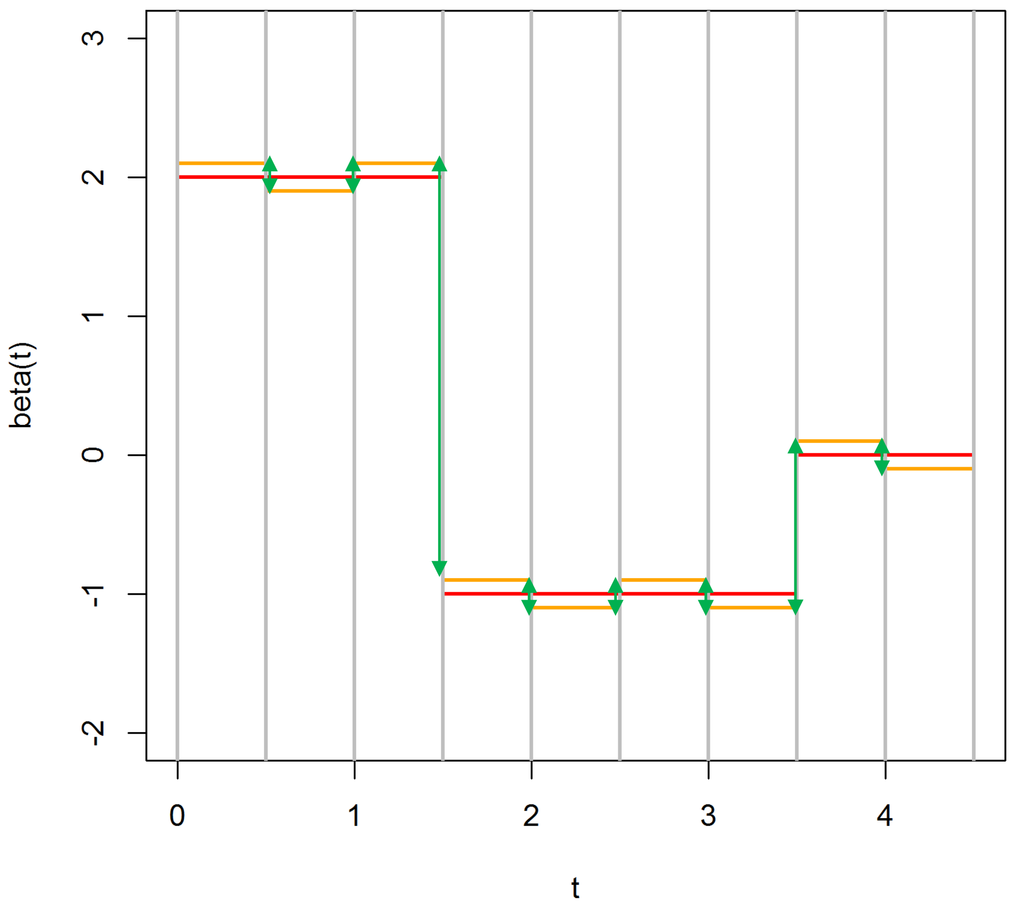
$$\sum_{j=2}^{p} (\beta_{j+1} \neq \beta_j)$$

Cette norme n'étant pas convexe, on utilise sa relaxation convexe (Total variation ou Fused Lasso).
On pénalise la hauteur de sauts.

$$\|\beta\|_{\mathrm{TV}} = \sum_{l=2}^{L} |\beta_{j,l} - \beta_{j,l-1}|$$

Version multivariée (group-TV):

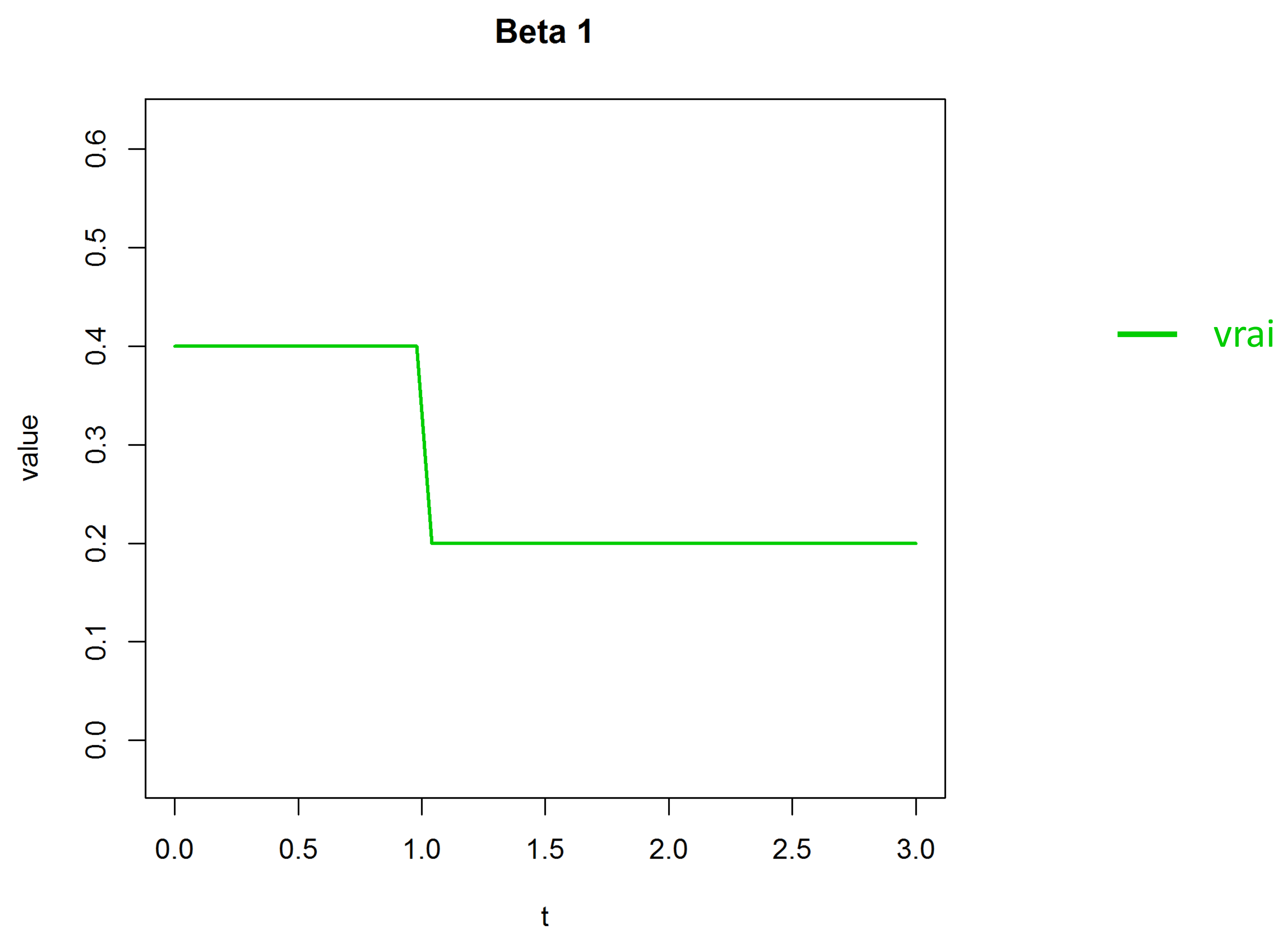
$$\|\beta\|_{\mathrm{TV}} = \sum_{j=1}^{p} \left(|\beta_{j,1}| + \sum_{l=2}^{L} |\beta_{j,l} - \beta_{j,l-1}| \right)$$

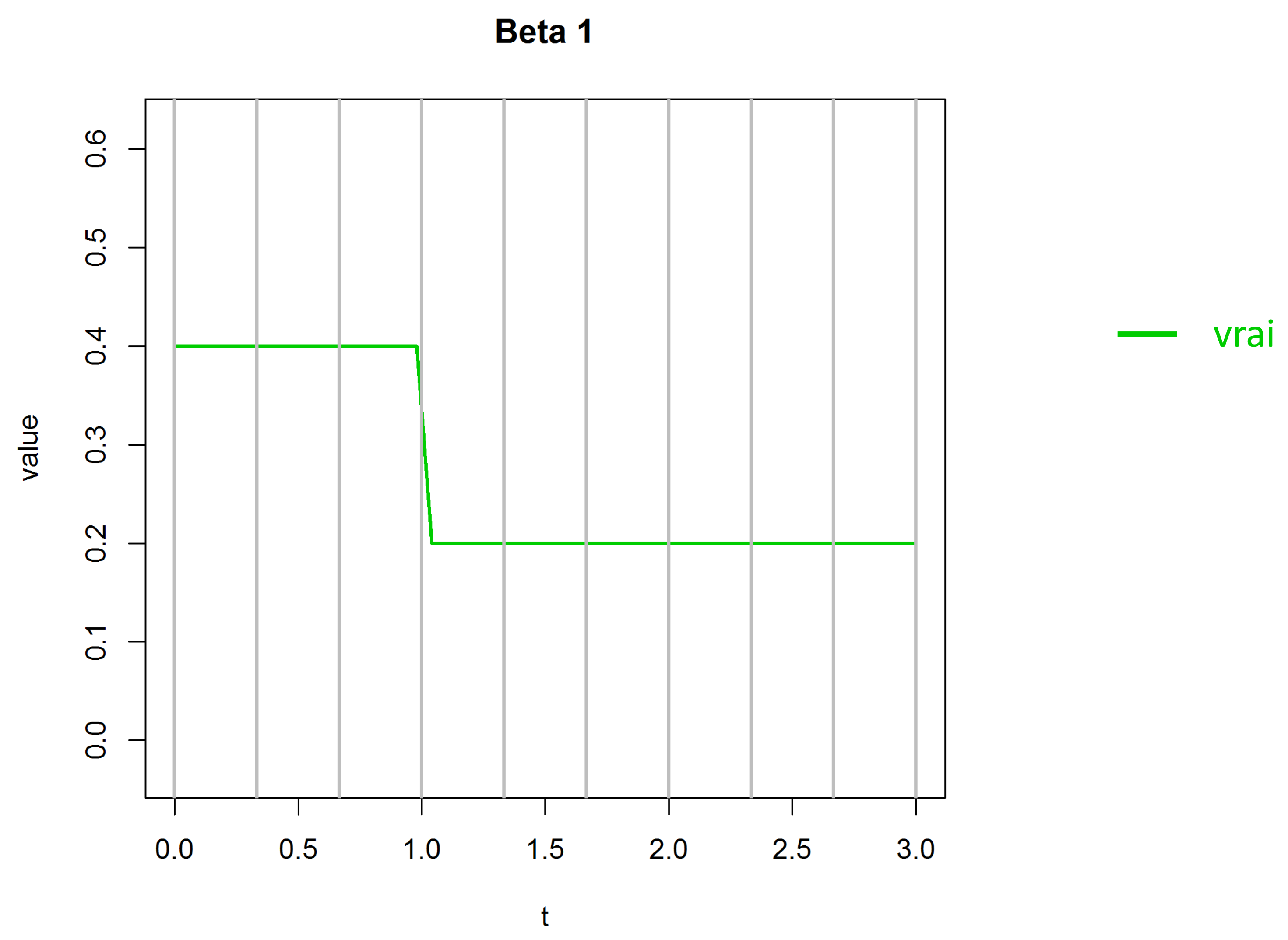


$$\hat{\beta}^{\mathrm{TV}} \in \underset{\beta \in \mathbb{R}^L}{\operatorname{argmin}} \left\{ \hat{R}(\beta) + \lambda \|\beta\|_{\mathrm{TV}} \right\}, \lambda > 0$$

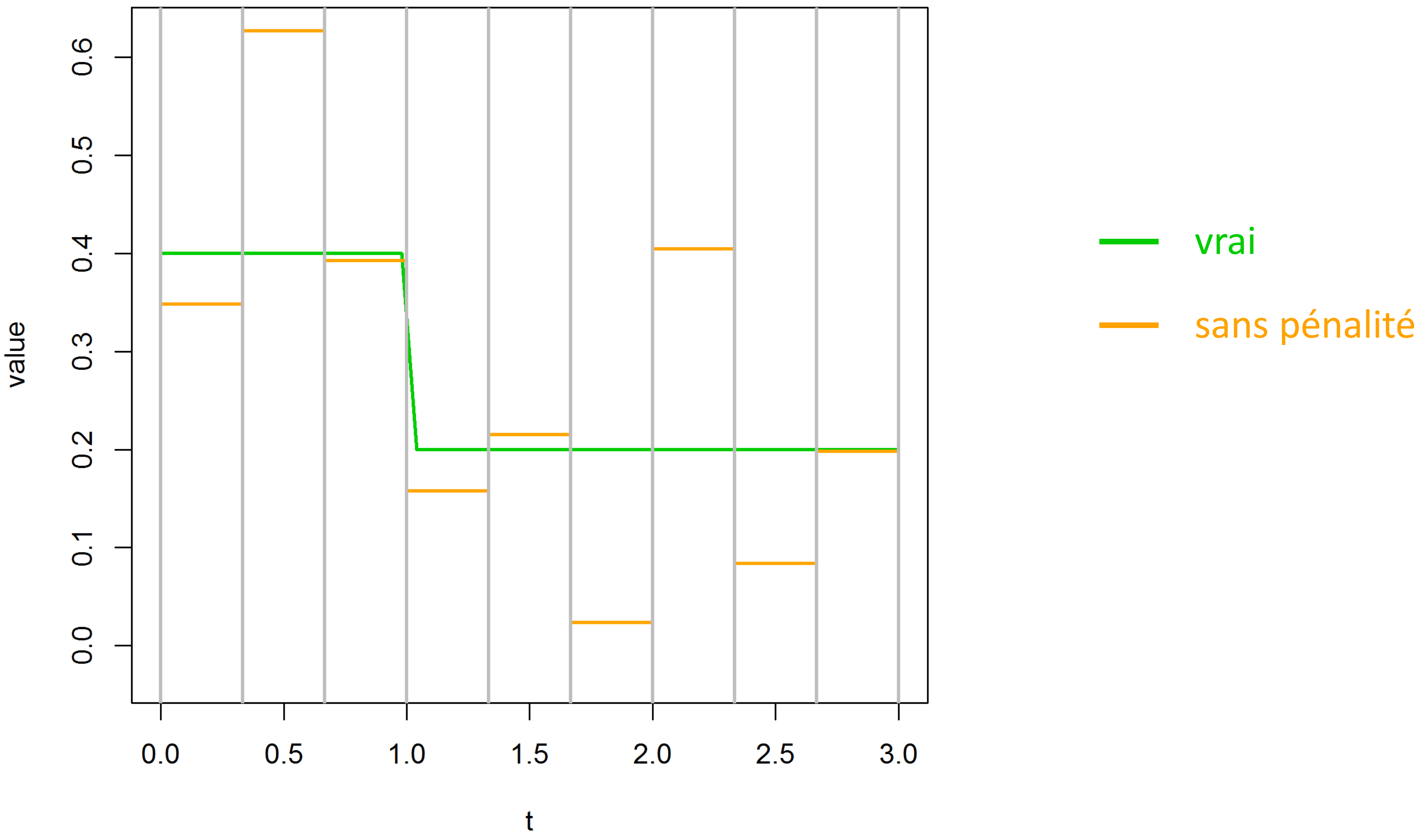
Simulation

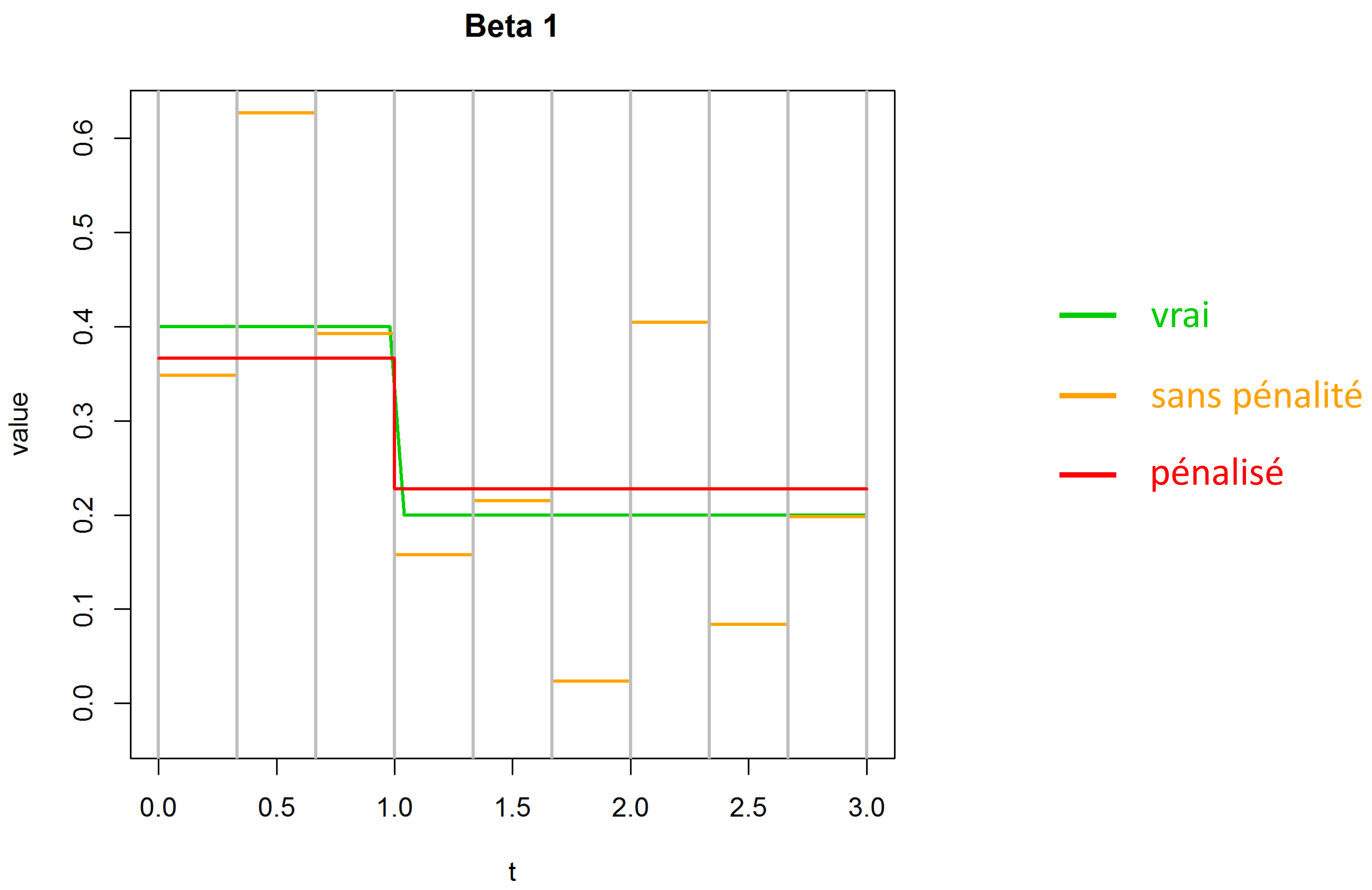
Thinning


Beta 1
value
0.0
0.1
0.2
0.3
0.4
0.5
0.6
t
0.0
0.5
1.0
1.5
2.0
2.5
3.0
vrai

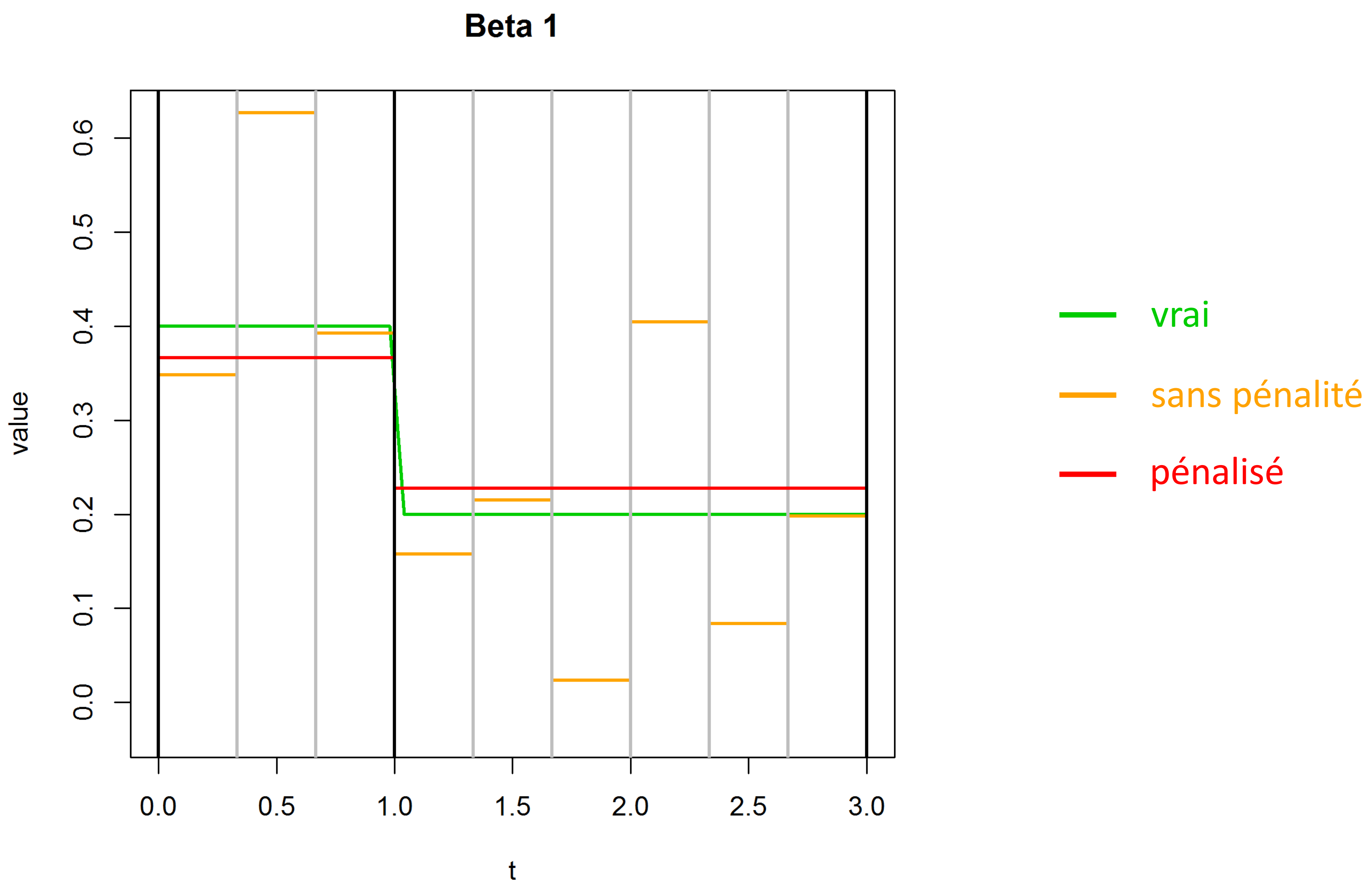

Beta 1
value
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
t
vrai

Beta 1

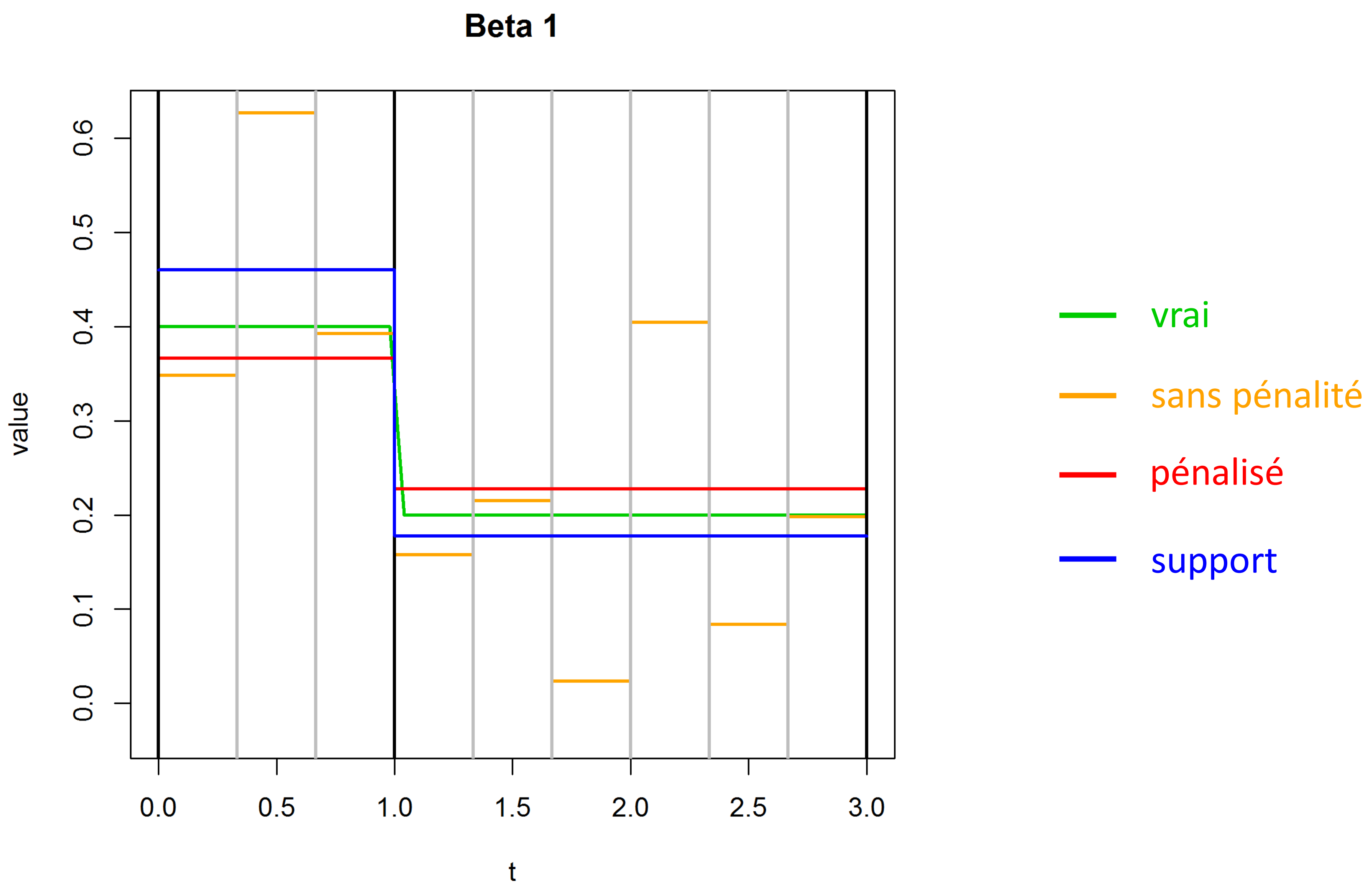

value
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
t
vrai
sans pénalité



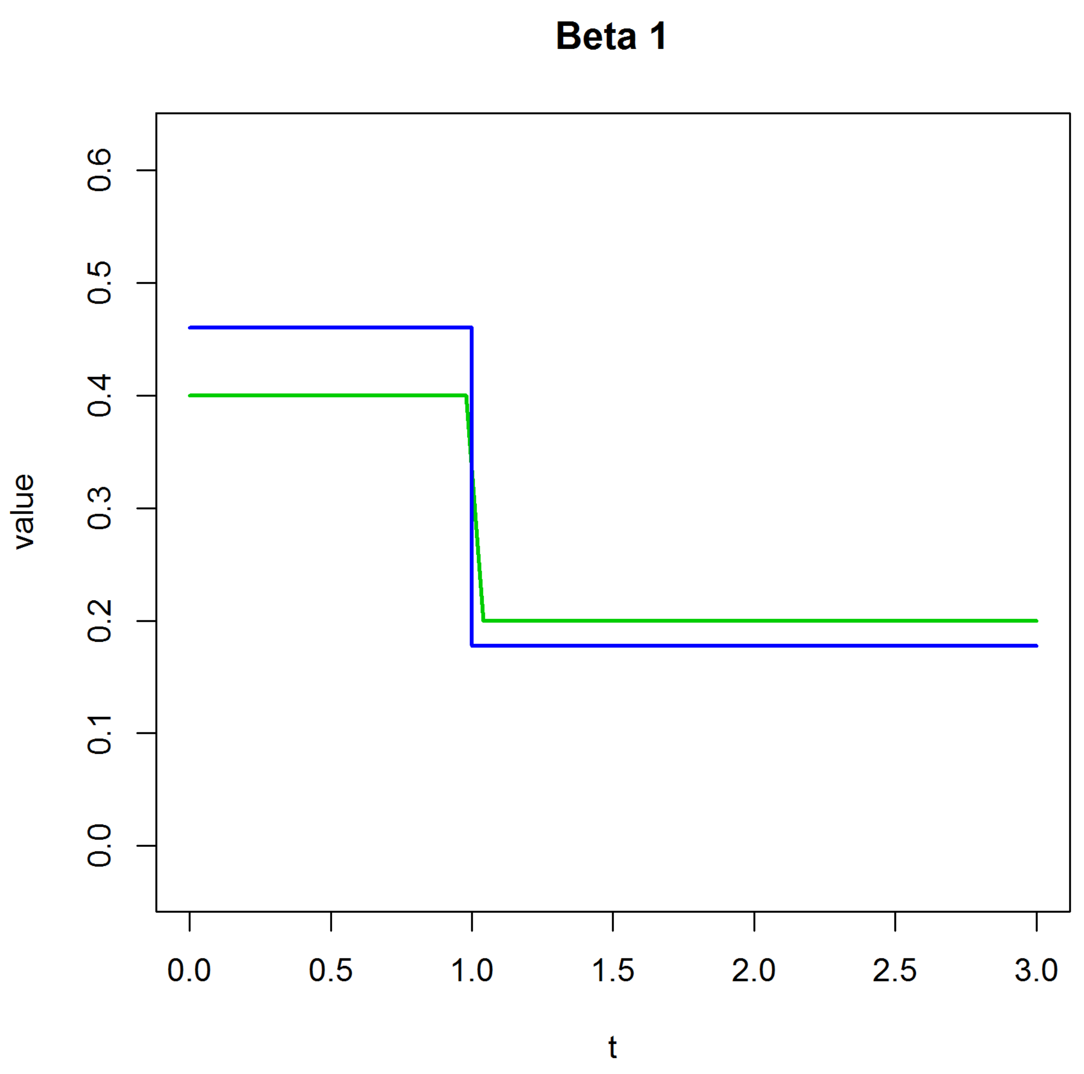
Beta 1
value
0.0
0.1
0.2
0.3
0.4
0.5
0.6
t
0.0
0.5
1.0
1.5
2.0
2.5
3.0
vrai
sans pénalité
pénalisé



Beta 1
value
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
t
vrai
sans pénalité
pénalisé



Beta 1
value
t
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
vrai
sans pénalité
pénalisé
support


Beta 1
value
0.0
0.1
0.2
0.3
0.4
0.5
0.6
t
0.0
0.5
1.0
1.5
2.0
2.5
3.0

timereg

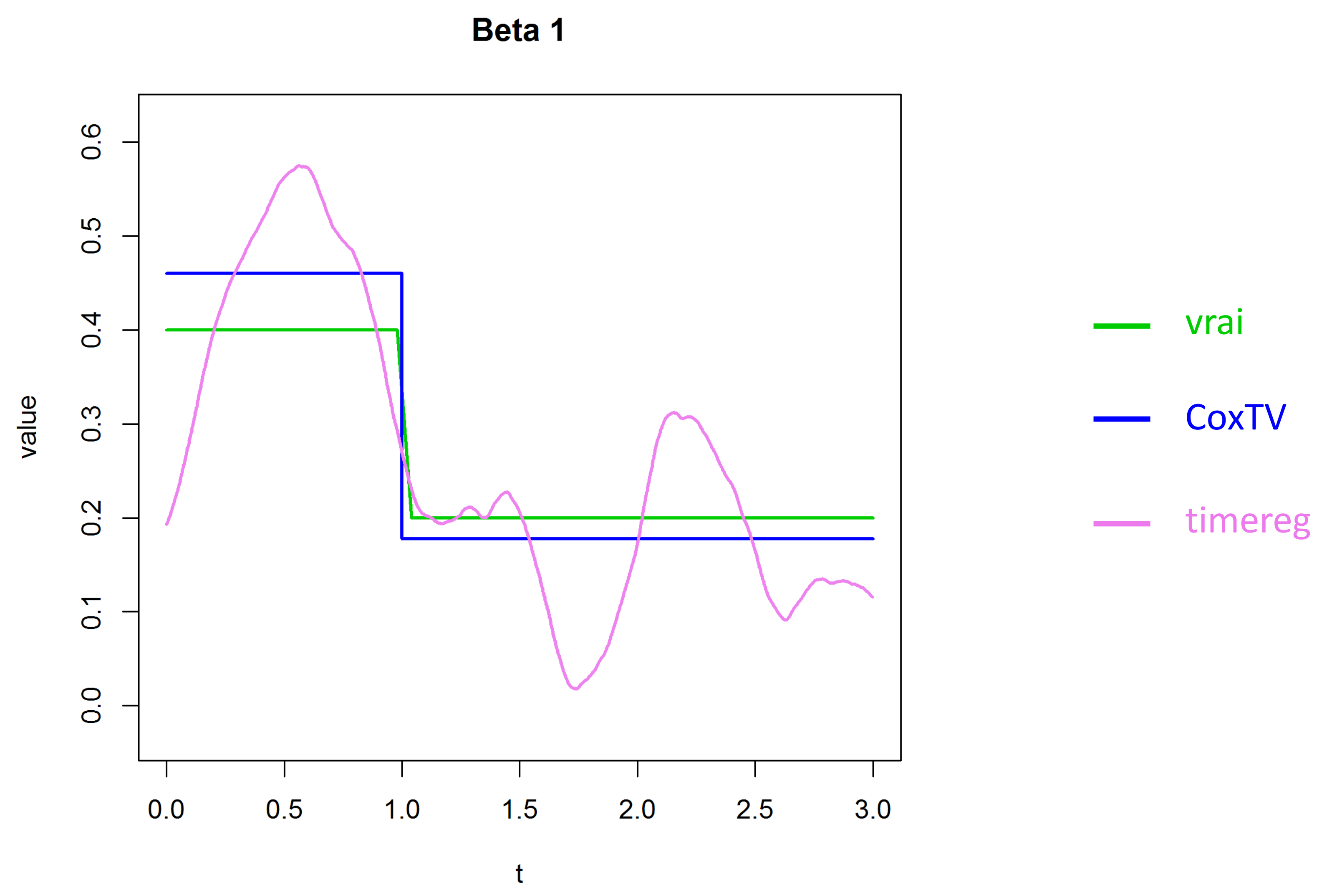
Il existe déjà un package R pour les modèles de survie dépendant du temps.

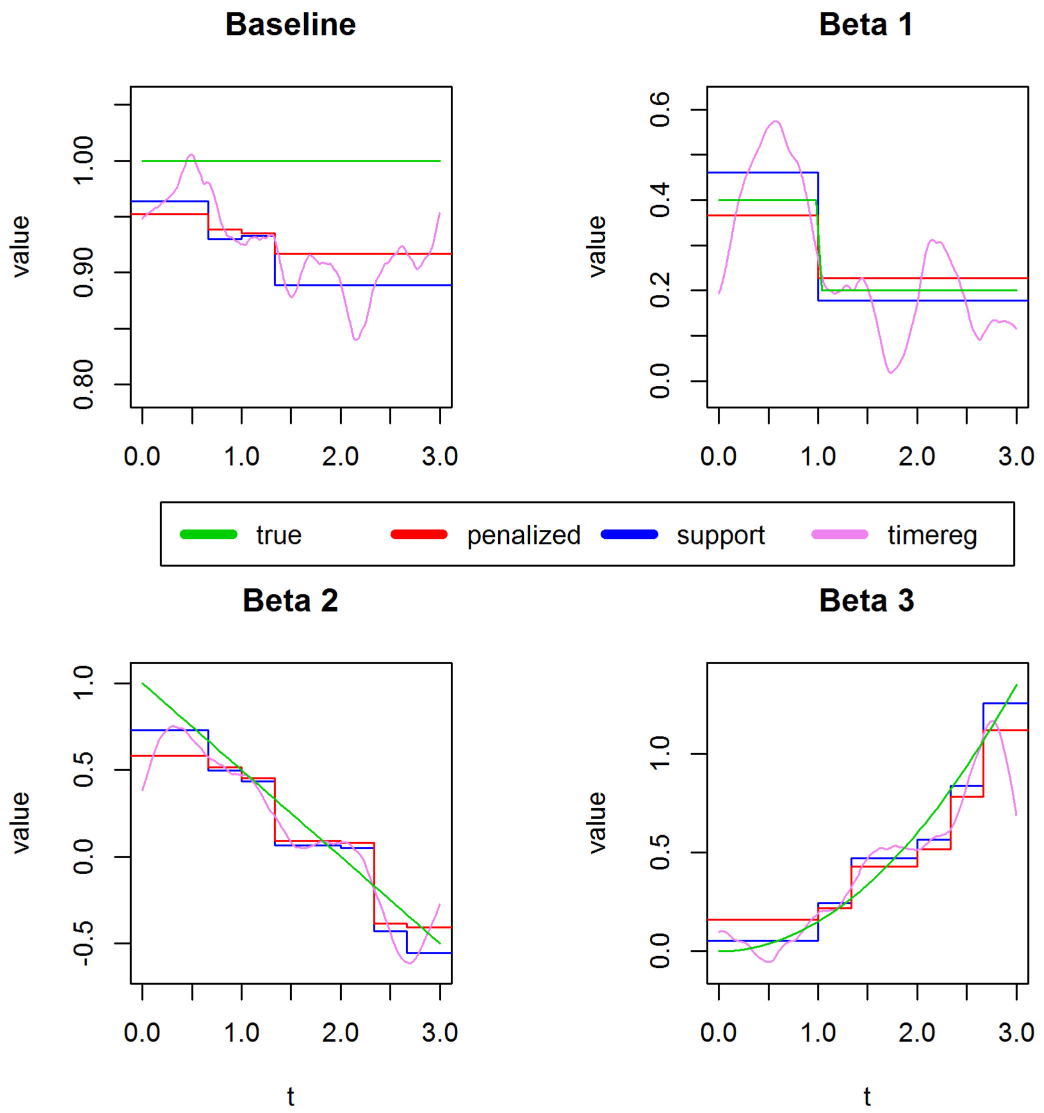
timereg: Flexible Regression Models for Survival Data

Programs for Martinussen and Scheike (2006), 'Dynamic Regression Models for Survival Data'

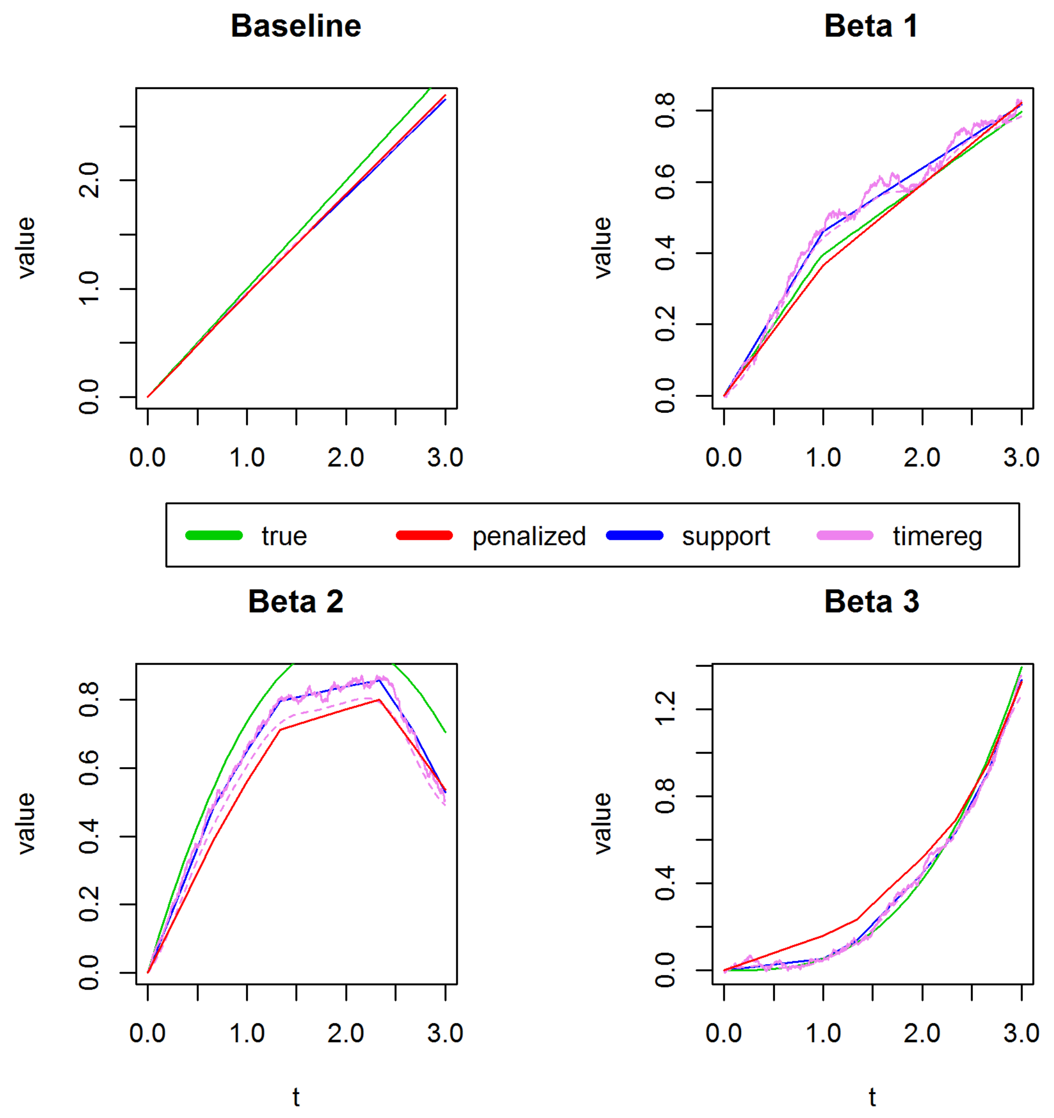
Différences :

- Estimation basée sur la fonction de hasard **cumulée**
- Retourne les coefficients cumulés, il faut ajouter un estimateur à noyaux pour obtenir les coefficients.
- L'optimisation repose sur une inversion matricielle et des itérés de lissage par noyaux

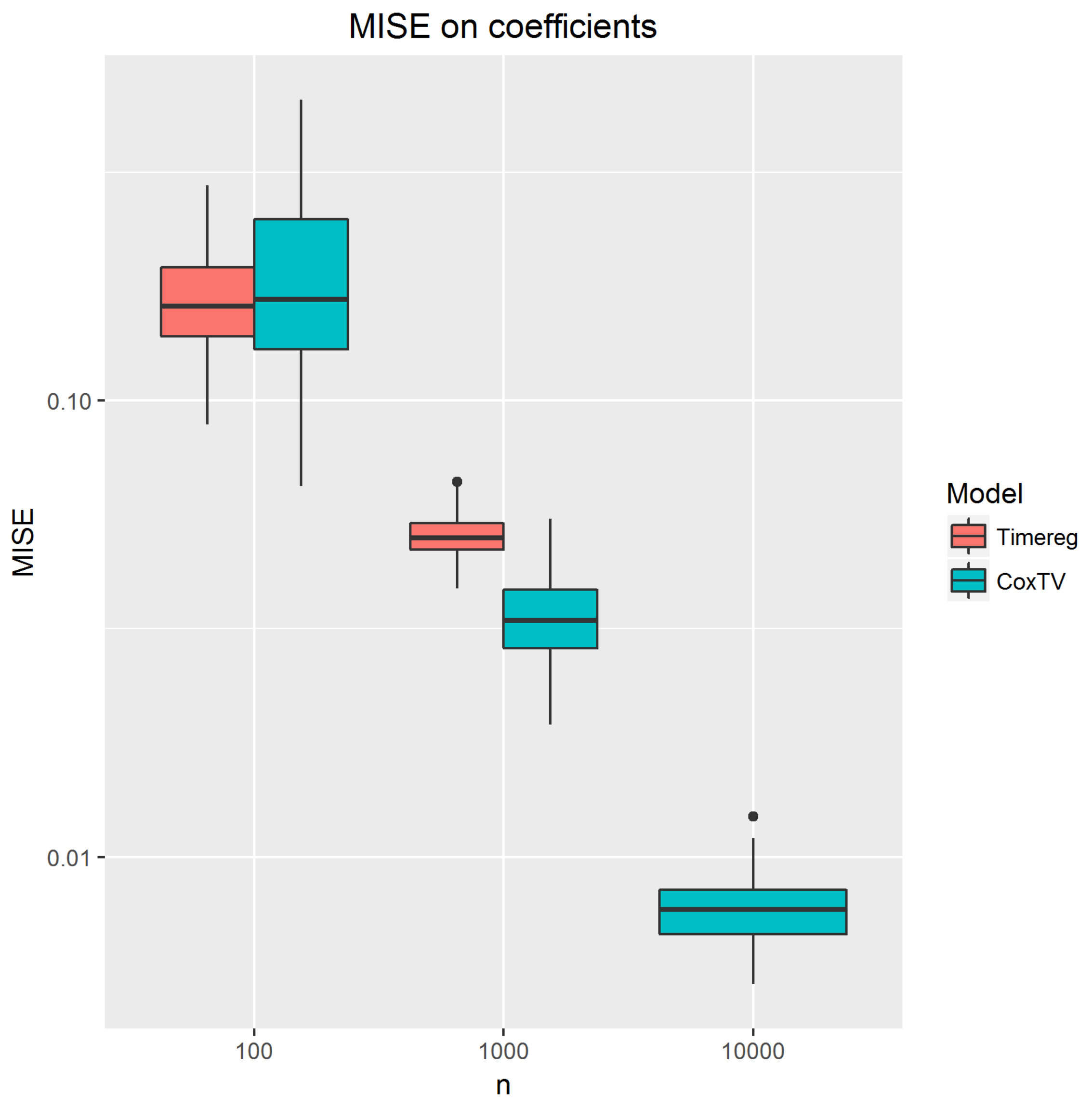

Beta 1
value
0.0
0.1
0.2
0.3
0.4
0.5
0.6
t
0.0
0.5
1.0
1.5
2.0
2.5
3.0
vrai
CoxTV
timereg

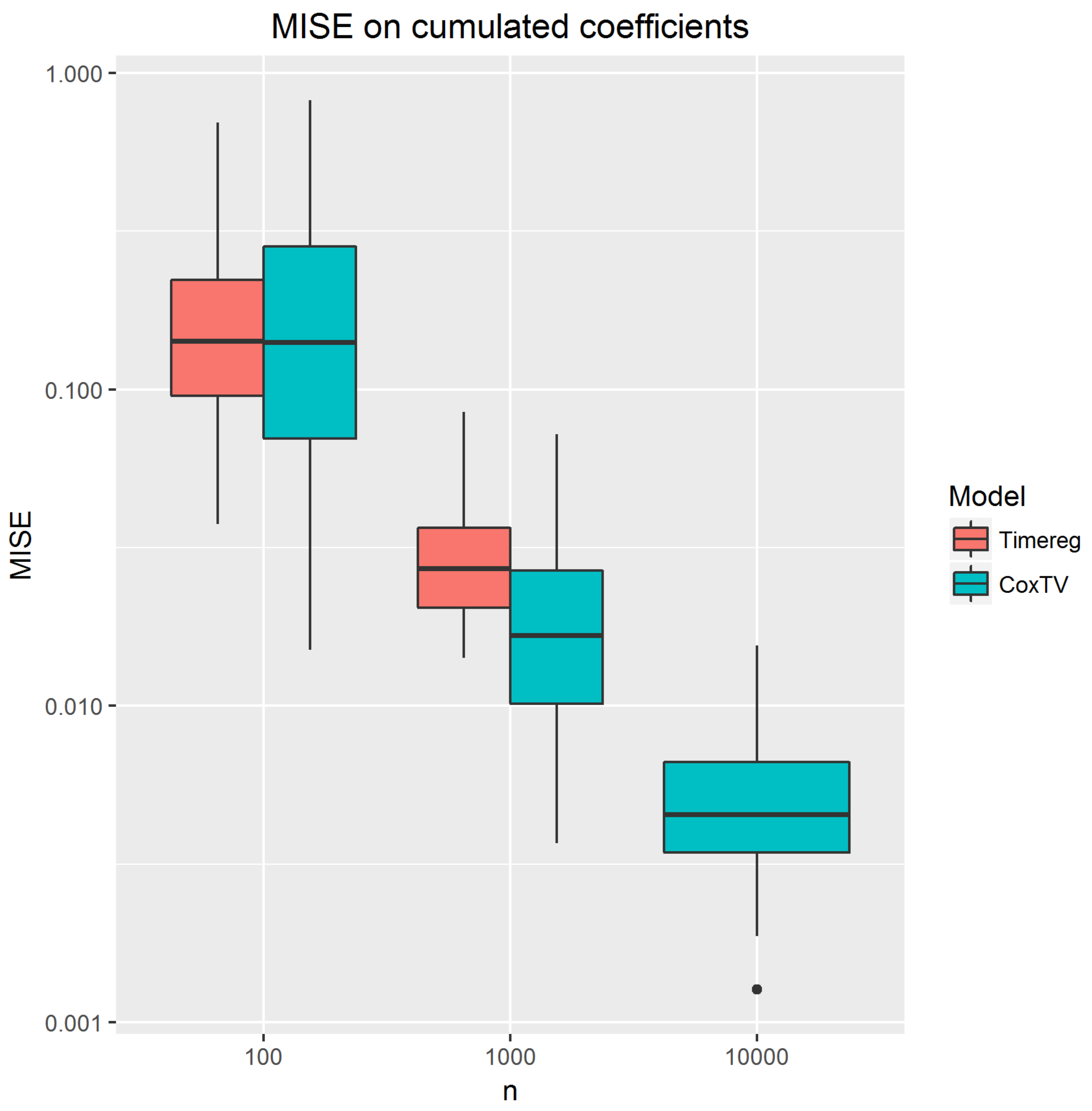


Baseline
Beta 1
Beta 2
Beta 3
value
t
0.80
0.90
1.00
0.0
0.2
0.4
0.6
-0.5
0.0
0.5
1.0
0.0
1.0
2.0
3.0
true
penalized
support
timereg



Baseline
Beta 1
true
penalized
support
timereg
Beta 2
Beta 3
value
t
0.0
1.0
2.0
3.0
0.2
0.4
0.6
0.8
1.2

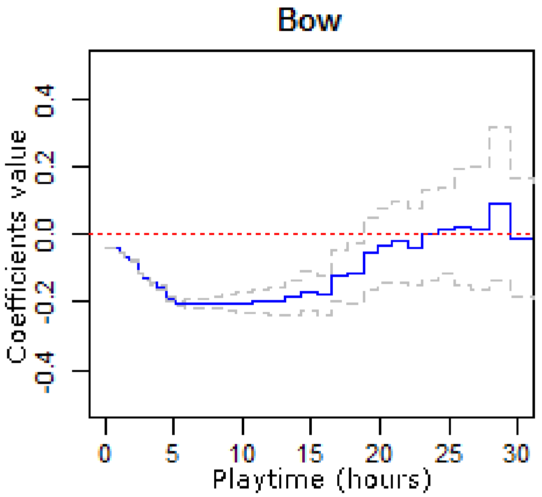


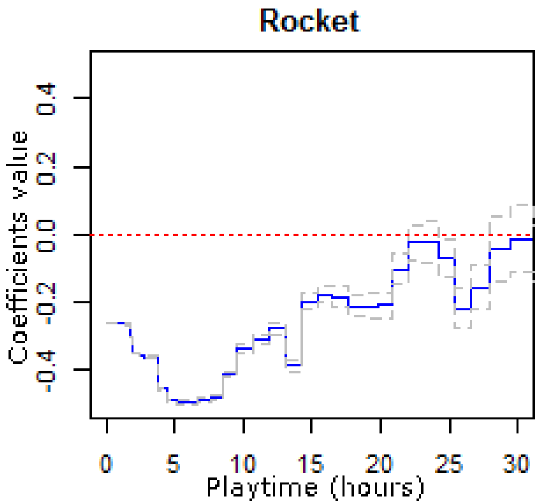


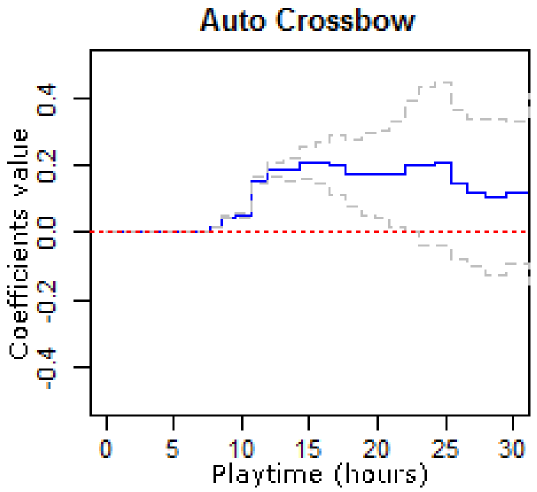
100 répétitions Monte-Carlo

Application

Les facteurs de rétention dans les jeux vidéo

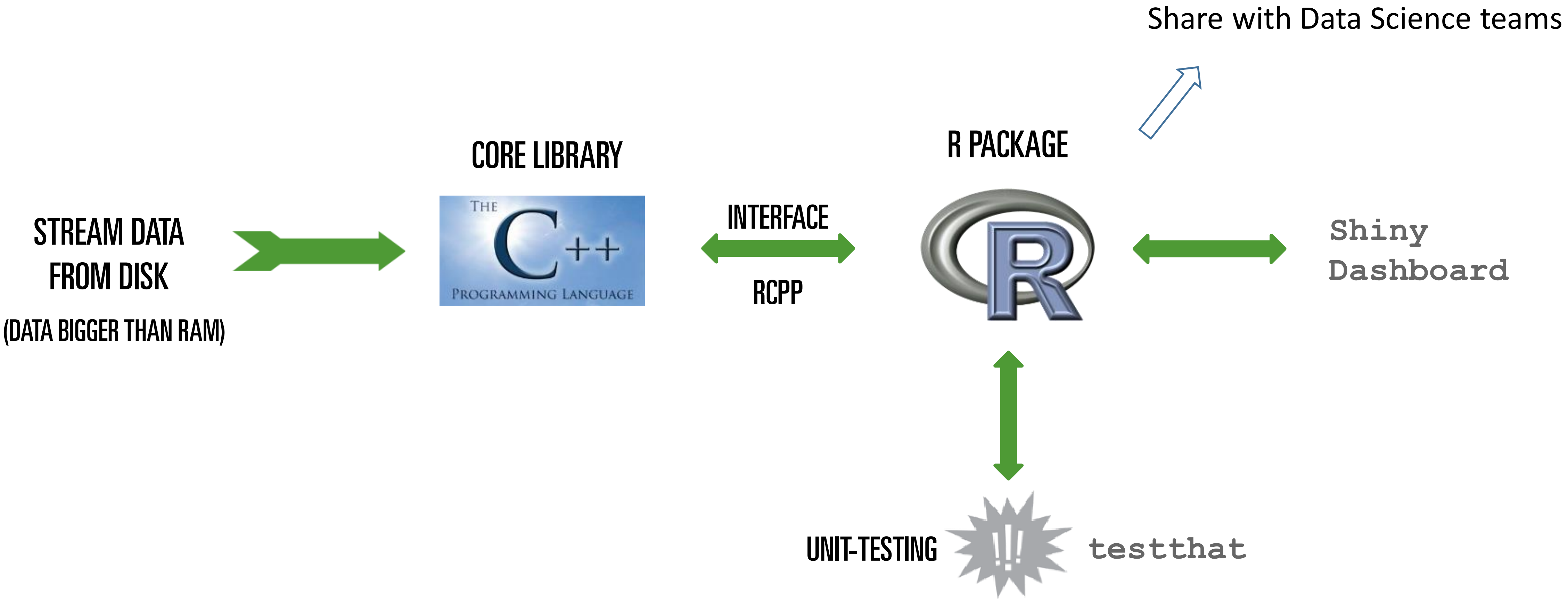

Bow
Coefficients value
Playtime (hours)
0.4
0.2
0.0
-0.2
-0.4
0
5
10
15
20
25
30


Rocket
Coefficients value
Playtime (hours)
0.4
0.2
0.0
-0.2
-0.4
0
5
10
15
20
25
30


Auto Crossbow
Coefficients value
Playtime (hours)
0.4
0.2
0.0
-0.2
-0.4
0
5
10
15
20
25
30

Merci !

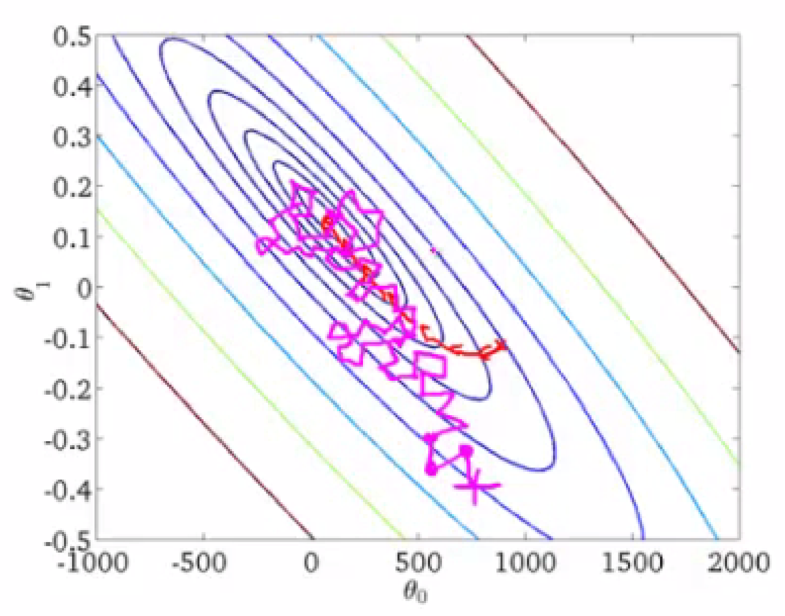
Bonus

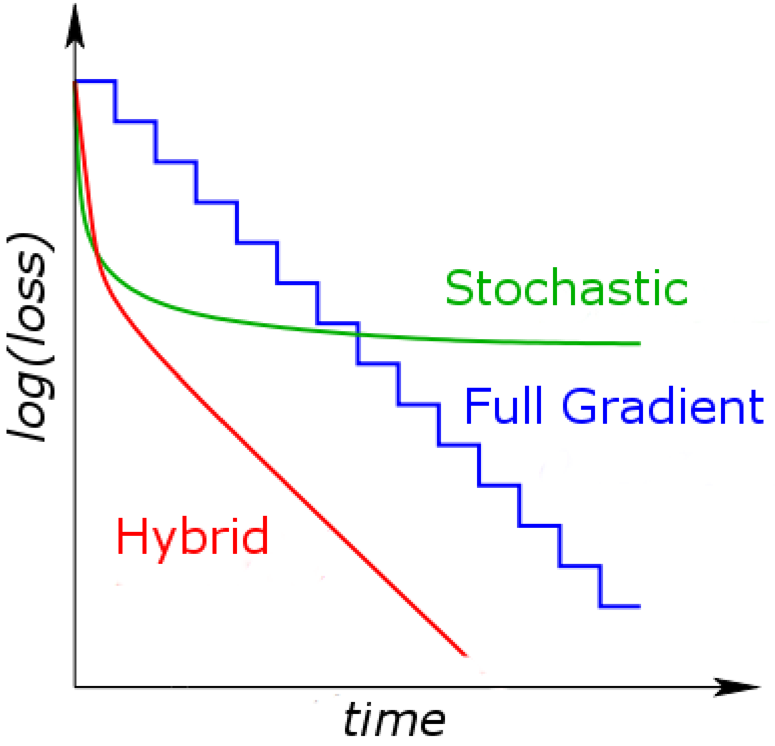

Share with Data Science teams
CORE LIBRARY
THE C++ PROGRAMMING LANGUAGE
R PACKAGE
STREAM DATA
FROM DISK
(DATA BIGGER THAN RAM)
INTERFACE
RCPP
Shiny
Dashboard
UNIT-TESTING
testthat

Stochastic Gradient Descent

$$\beta \leftarrow \beta - \mu \underbrace{\nabla_{\boldsymbol{i},\beta}}$$

Gradient compute on one observation





Cox proportional likelihood

Cost function :

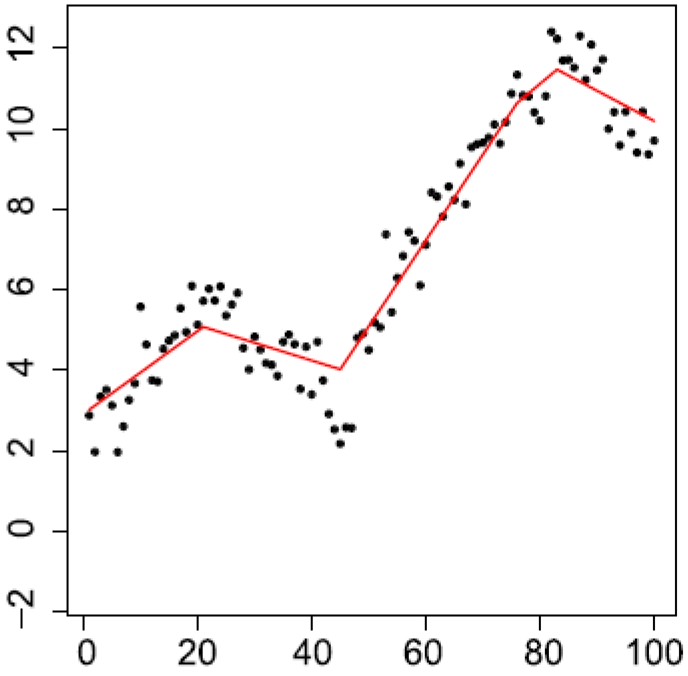
$$L(\beta) = \frac{1}{n} \prod_{i \in D} \frac{\exp(x_i^T \beta)}{\sum_{j \in R_i} \exp(x_j^T \beta)}$$

Individual gradients:

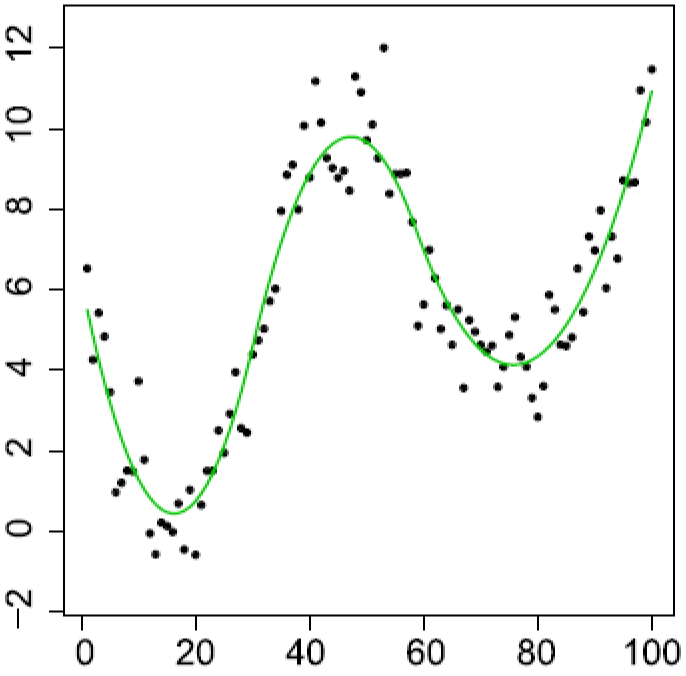
$$\Delta f_i(\beta) = -x_i + \sum_{j \in R_i} \frac{x_j \exp(x_j^T \beta)}{\sum_{k \in R_i} \exp(x_k^T \beta)} \qquad \sim o(np)$$

It is $o(p)$ for regression or logistic regression.

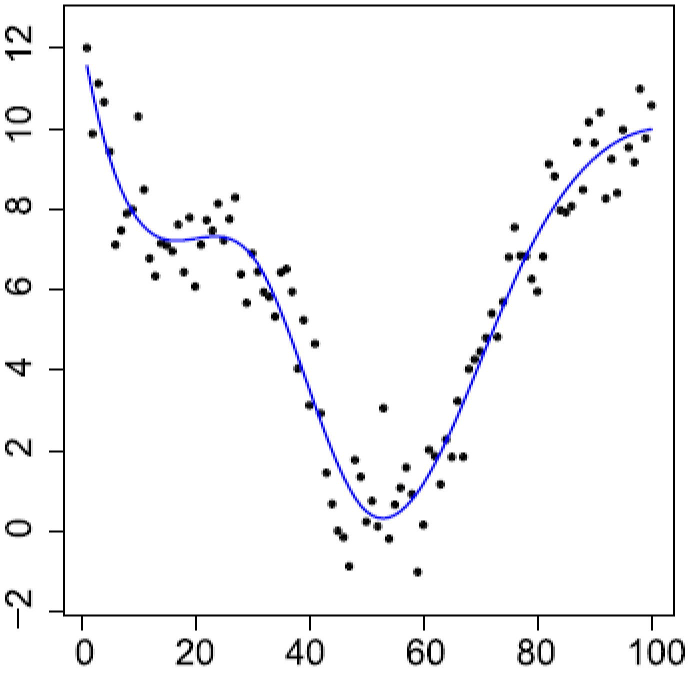
Computing individual gradient costs almost the same as computing full gradient !



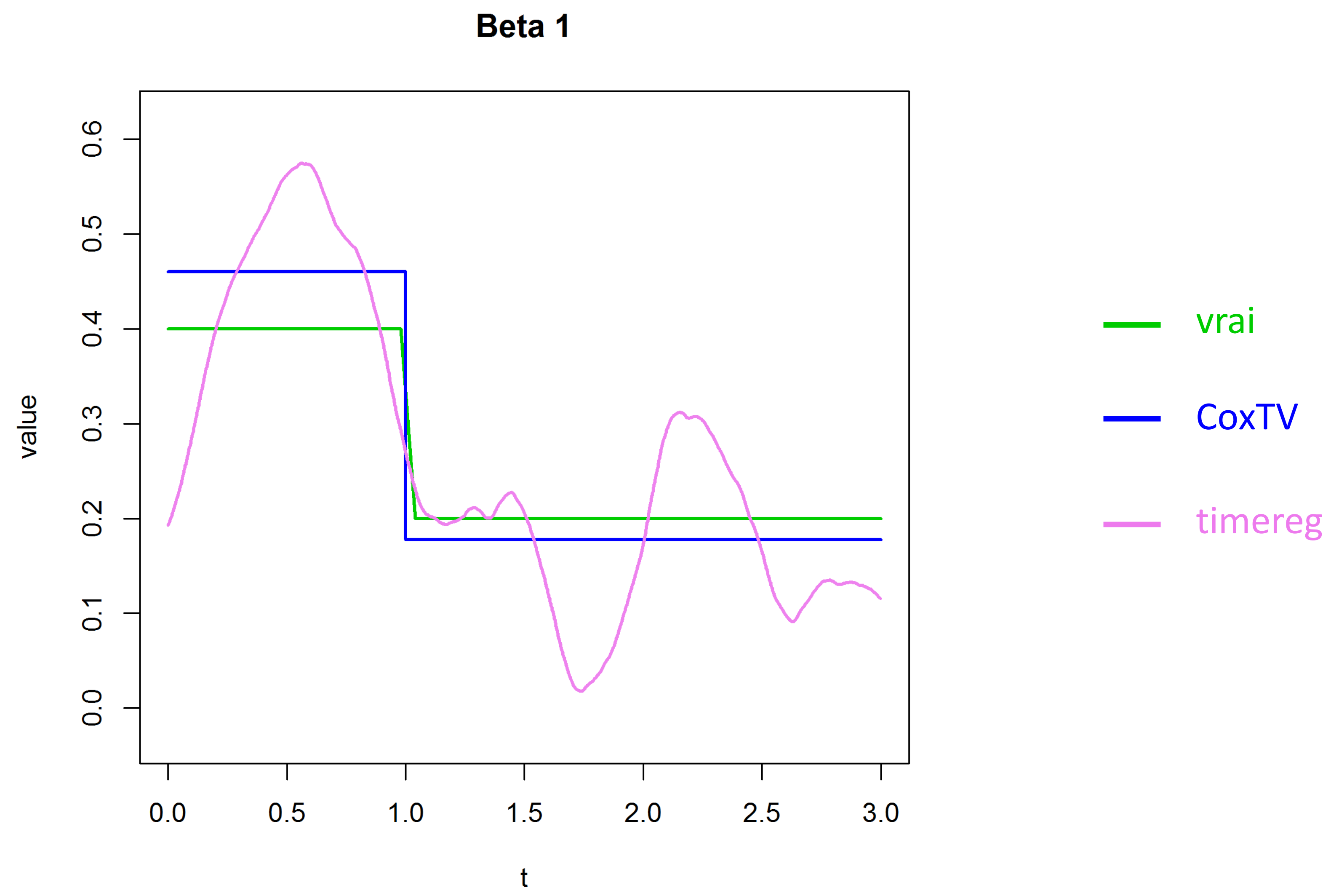
(a) Linear



(b) Quadratic



(c) Cubic



N = 1 000

N_event ≈ 3 000

Nb_obs ≈ 54 000

Nous introduisons la pénalité $(\ell_1 + \ell_1)$-variation totale avec poids défini par

$$\|\beta\|_{\mathrm{gTV},\hat{\gamma}} = \sum_{j=1}^{p} \Big(\hat{\gamma}_{j,1}|\beta_{j,1}| + \sum_{l=2}^{L} \hat{\gamma}_{j,l}|\beta_{j,l} - \beta_{j,l-1}|\Big)$$

pour tout $\beta \in \mathbb{R}^{p\times L}$ avec $\hat{\gamma} = (\hat{\gamma}_{1,\cdot}^\top, \dots, \hat{\gamma}_{p,\cdot}^\top)^\top$, tel que $\hat{\gamma}_{j,\cdot} \in \mathbb{R}_+^L$ pour tout $j = 1, \dots, p$, donné par

$$\hat{\gamma}_{j,l} \approx \sqrt{\frac{L\log(pL)}{n}\hat{V}_{j,l}}, \text{ avec } \hat{V}_{j,l} = \frac{1}{n}\sum_{i=1}^{n}\int_{\bigcup_{u=l}^{L} I_u} (X_i^j(t))^2 dN_i(t).$$

$$\hat{\beta}^{\mathrm{M}} = \mathrm{argmin}_{\beta\in\mathbb{R}^{p\times L}} \Big\{\ell_n^{\mathrm{M}}(\beta) + \|\beta\|_{\mathrm{gTV},\hat{\gamma}}\Big\}.$$

divergence de Kullback empirique associée au modèle de Cox (2)

$$\begin{aligned} K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left(\log\lambda_\star^{\mathrm{M}}(t, X_i(t)) - \log\lambda_\beta^{\mathrm{M}}(t, X_i(t))\right)\lambda_\star^{\mathrm{M}}(t, X_i(t))Y_i(t)dt \\ - \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left(\lambda_\star^{\mathrm{M}}(t, X_i(t)) - \lambda_\beta^{\mathrm{M}}(t, X_i(t))\right)Y_i(t)dt. \end{aligned}$$

Théorème 2. *Pour $x > 0$ fixé, l'estimateur $\hat{\lambda}^{\mathrm{M}}$ defini dans (4), vérifie avec une probabilité supérieur à $1 - C_{\mathrm{M}}\mathrm{e}^{-x}$ ($C_{\mathrm{M}} > 0$),*

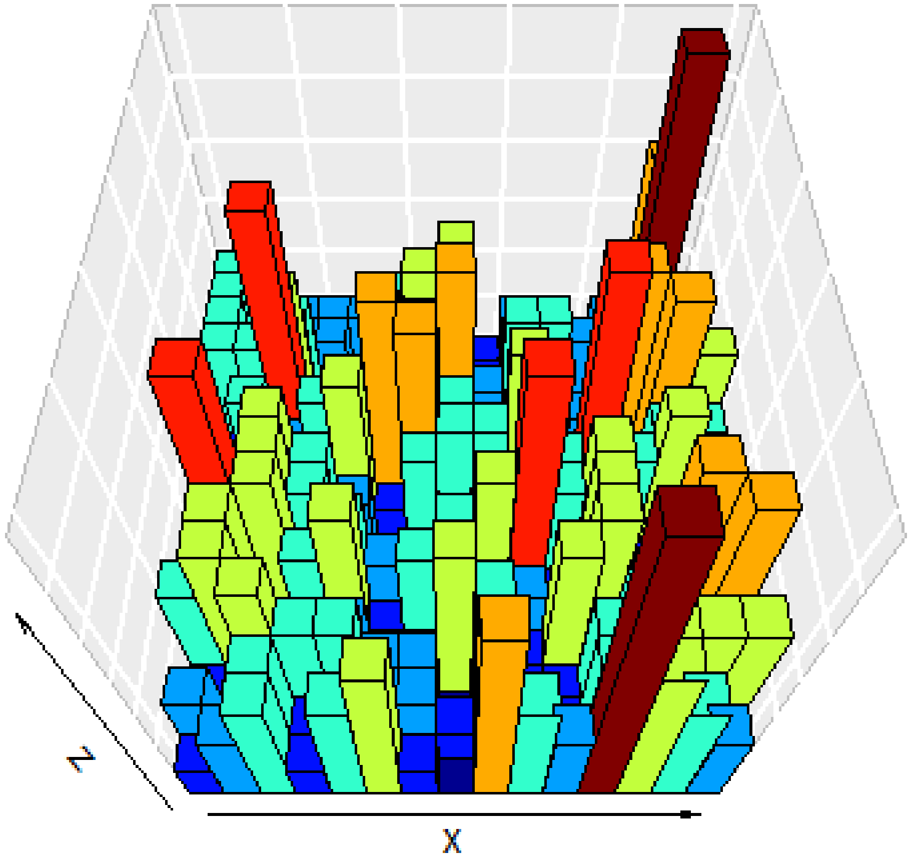
$$K_n(\lambda_\star^{\mathrm{M}}, \hat{\lambda}^{\mathrm{M}}) \leq \inf_{\beta \in \mathbb{R}^{p \times L}} \Big(K_n(\lambda_\star^{\mathrm{M}}, \lambda_\beta^{\mathrm{M}}) + 2\|\beta\|_{\mathrm{gTV},\hat{\gamma}} \Big). \tag{6}$$

Ces théorèmes admettent l'interprétation suivante : les risques théoriques de nos estimateurs sont bornés par les meilleurs risques atteignables sur l'ensemble des modèles basés sur les histogrammes (Λ^{A} et Λ^{M}), plus un terme satisfaisant

$$\|\beta\|_{\mathrm{gTV},\hat{\gamma}} \asymp \|\beta\|_{\mathrm{gTV}} \max_{j=1,\ldots,p} \max_{l=1,\ldots,L} \sqrt{\frac{L \log pL}{n} \hat{V}_{j,\ell}}. \tag{7}$$

pour tout $\beta \in \mathbb{R}^{p \times L}$. La norme $\|\cdot\|_{\mathrm{gTV}}$ représente la $(\ell_1 + \ell_1)$-variation totale sans poids (*i.e.* $\hat{\gamma}_{j,l} = 1$). Le terme dominant dans (7) est d'ordre $\|\beta\|_{\mathrm{gTV}} \big(L \log(pL)/n\big)^{1/2}$ (à un facteur de $\log\log$ près).

$n * p * t_{i,j}$



$n * p * t_i$

$t_i = \sum temps\ observation \geq \max_j t_{i,j}$

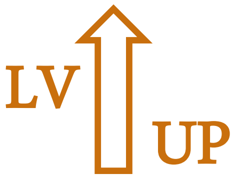
Id	event	start	stop	x_1	…	x_p
1001	0	0	1.2	1	…	-5
1001	0	1.2	2.0	3	…	-5
1001	0	2.0	5.3	5	…	-15
1001	1	5.3	5.8	8	…	22
1002	0	0	0.2	2	…	-1

1. Increase frequency for all covariates
2. Lose information

Or losing and transforming information if we discretize

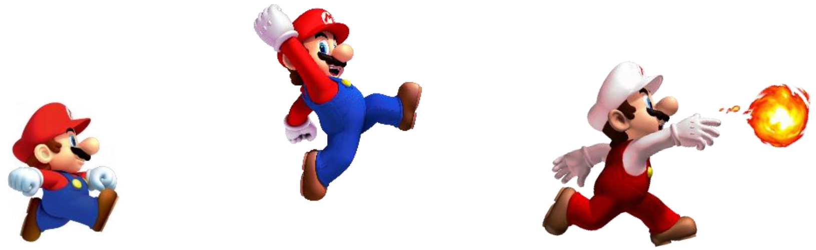
II Data

- We can track every change in the game



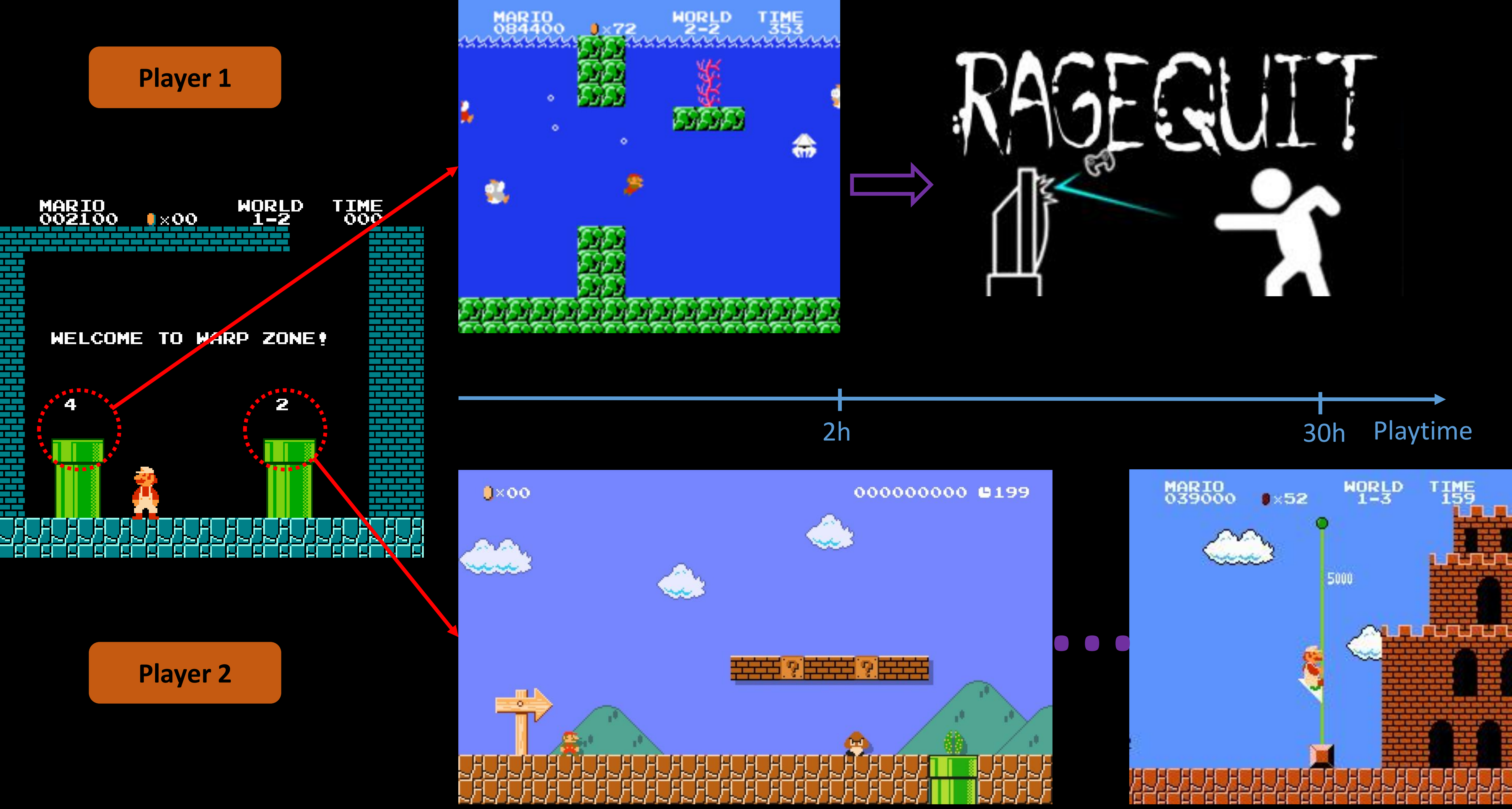
Levelling

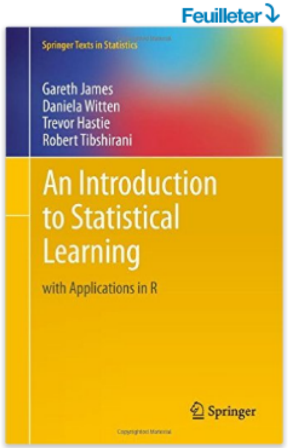
Weapon usage



Datasets size :

millions of players $*$ thousands of observations per variable $*$ many variables


Player 1
MARIO 002100
×00
WORLD 1-2
TIME 000
WELCOME TO WARP ZONE!
4
2
Player 2
MARIO 084400
×72
WORLD 2-2
TIME 353
RAGEQUIT
2h
30h
Playtime
×00
000000000
199
MARIO 039000
×52
WORLD 1-3
TIME 159
5000



An Introduction to Statistical Learning: With Applications in R (Anglais) R

de Gareth James (Auteur), Trevor Hastie (Auteur), Robert Tibshirani (Auteur), Daniela Witten (Auteur)

★★★★★ 128 Commentaires sur Amazon.com | Soyez la première personne à écrire un commentair

#1 Meilleure vente dans Mathematical & Statistical Software

Voir les formats et éditions

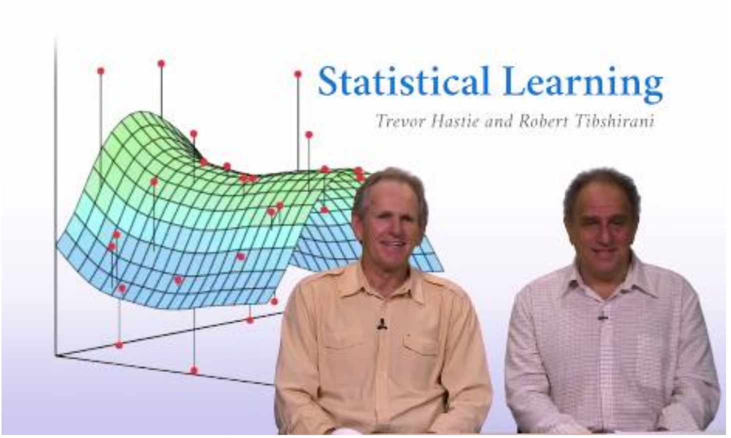
Relié
EUR 51,43

7 d'occasion à partir de EUR 51,63
17 neufs à partir de EUR 48,76

Voulez-vous le faire livrer aujourd'hui, mardi 10 jan.? Commandez-le dans les **3 h et 15 mins** et choisissez

Note: Cet article est éligible à la livraison **en points de collecte.** Détails

An Introduction to Statistical Learning provides an accessible overview of the field of statistical learning, an essential toolset for making sense of the vast and complex data sets that have emerged in fields ranging from biology to finance to marketing to astrophysics in the past twenty years. This book

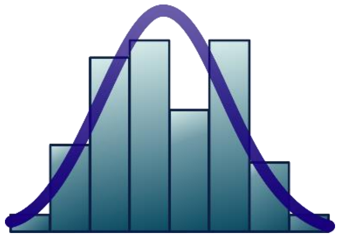


★★★★★ **wonderful but watch the movie** 14 février 2014

Par I Teach Typing - **Publié sur Amazon.com**

Format: Relié | **Achat vérifié**

This is a wonderful book written by luminaries in the field. While it is not for casual consumption, it is a relatively approachable review of the state of the art for people who do not have the hardcore math needed for The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics). This book is the text for the free Winter 2014 MOOC run out of Stanford called StatLearning (sorry Amazon will not allow me to include the website). Search for the class and you can watch Drs. Hastie and Tibshirani teach the material in this book.



Playtime

Lifetime

Survival Analysis

Repeated Events