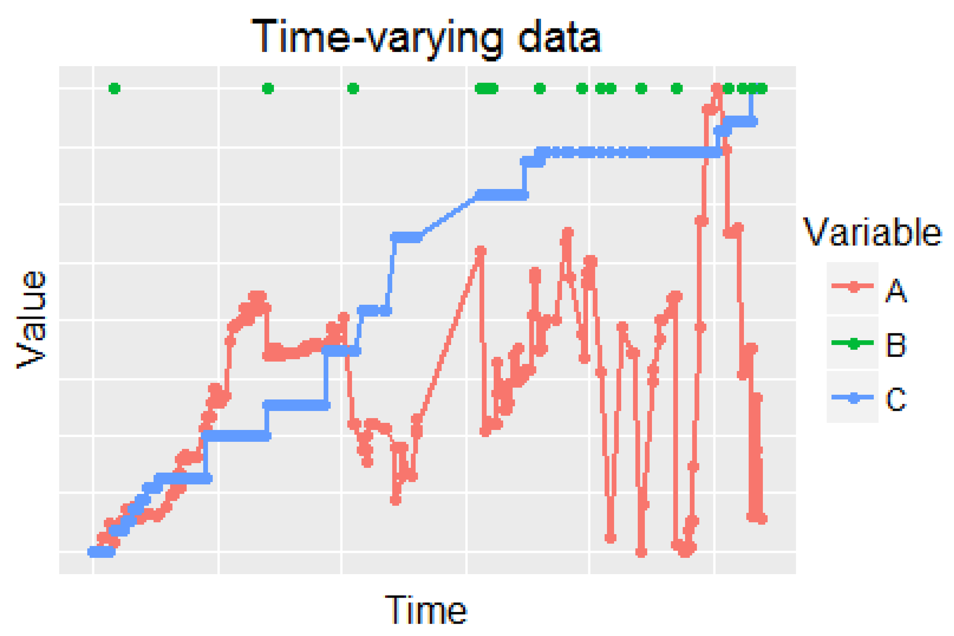
Stochastic Gradient Descent on Cox Model
Time varying data & coefficients

Thibault Allart, Agathe Guilloux

Introduction

- Model covariates influence on time-to-event data.
- **High number of individuals.**
- Many covariates : model has to select good ones.
- Time varying data.

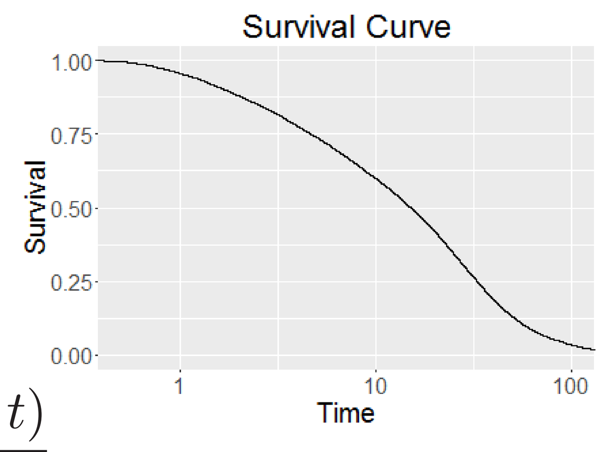


- Individual characteristics change over time
- Functional value only known at some points of time.
- Observation times differ for individuals & covariates.

Survival Analysis

Survival function :

$$S(t) = \mathbb{P}\left(T \geq t | X(s), s \leq t\right)$$



Hazard function :

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}\left(t \leq T \leq t + \Delta t | T \geq t, X(s), s \leq t\right)}{\Delta t}$$

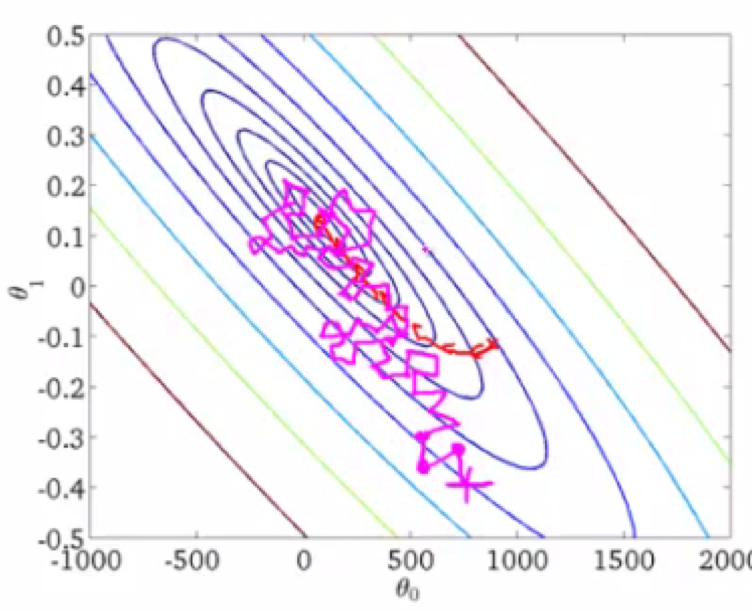
$\lambda(t)$ is the instantaneous risk that an event occurred at time t, given covariates and knowing that the event did not occur before.

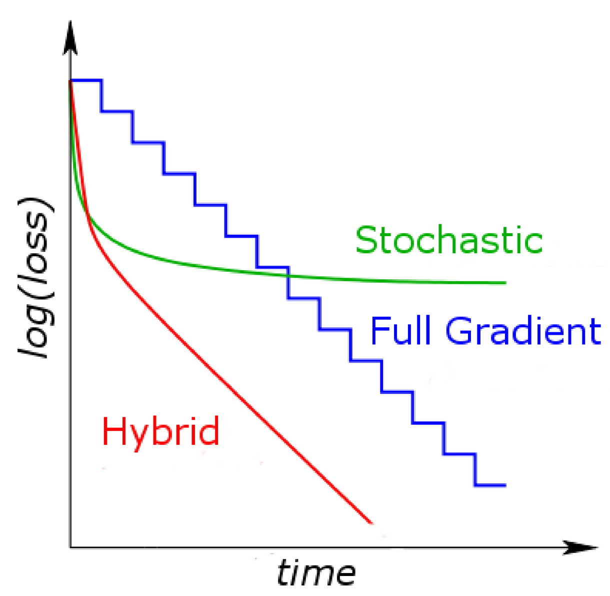
We consider the Cox model [1] where covariates X and coefficients β depend on time.

$$\lambda^\star(t|X(t)) = \exp\left(X(t)\beta^\star(t)\right)$$

Stochastic Gradient Descent

- Converges faster than full gradient algorithm to a weak solution.
- Can be used for online learning.





Best optimization strategy is to start with some SGD steps and complete with a full gradient descent algorithm.

Problem with Cox proportional likelihood

Cox proportional likelihood :

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

Individual gradient of minus log-likelihood :

$$\Delta f_i(\theta) = -x_i + \sum_{j \in R_i} \frac{x_j \exp(x_j^T \theta)}{\sum_{k \in R_i} \exp(x_k^T \theta)}$$

- Each SGD step cost $O(np)$ operations.
- It only costs $O(p)$ for regression and logistic regression.

Method

Using counting process notations, minus log-likelihood is given by :

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

see [2] for details.
Integral approximation can be done using numerical algorithm. This allows considering splines approximation for coefficients and/or data.

Piecewise constant function

To speed up the evaluation process
we consider piecewise constant functions :

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

This gives us pL coefficients to estimate.

$$\ell_n(\beta) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{l=1}^{L} \left(X_{i,l}\beta_l N_{i,l}(I_l) - \exp\left(X_{i,l}\beta_l\right) \int_{I_l} Y_i(t) ds \right).$$

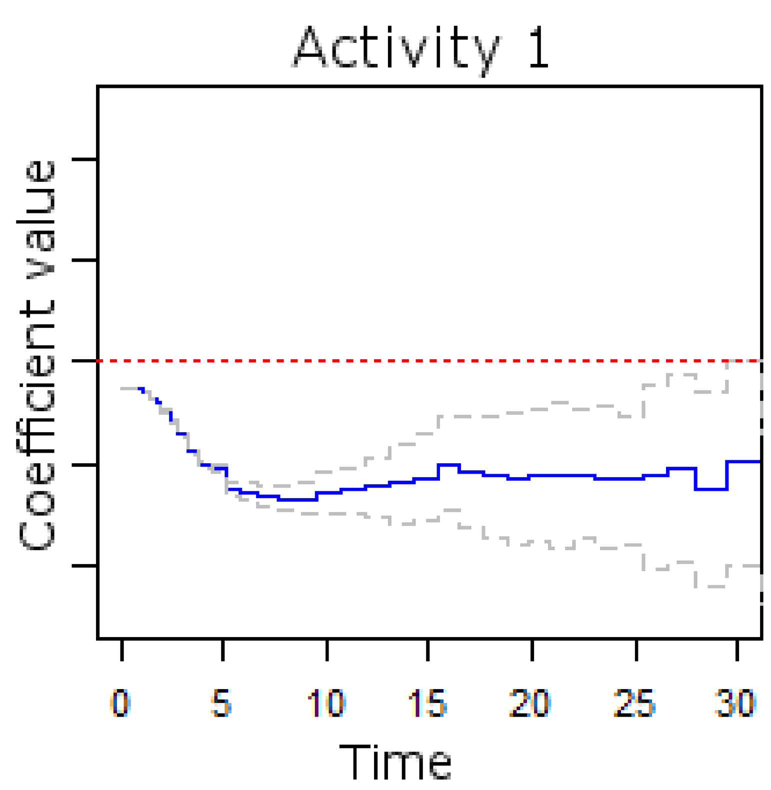
To avoid the curse of dimensionality we introduce a penalty which combines Lasso and Total Variation.

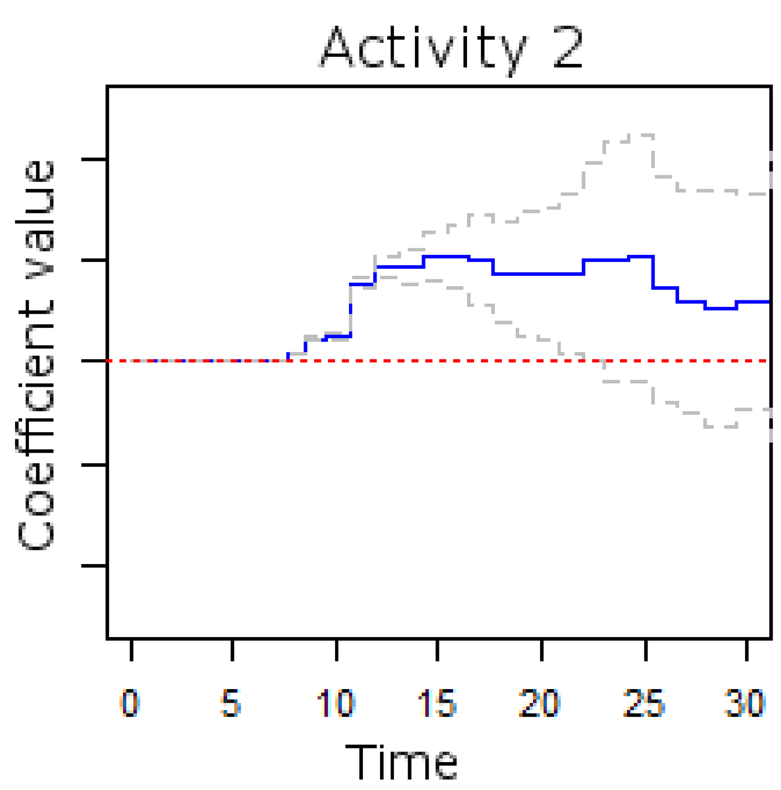
$$\|\beta\|_{\mathrm{gTV},\hat{\gamma}} = \lambda \sum_{j=1}^{p} \left(\hat{\gamma}_{j,1} |\beta_{j,1}| + \sum_{l=2}^{L} \hat{\gamma}_{j,l} |\beta_{j,l} - \beta_{j,l-1}| \right)$$

- L and $\gamma_{j,l}$ have theoretical values.
- λ is set using cross-validation

Application

Model is applied in video games industry to model design influence on player retention.





- Player doing a lot of activity 1 tends to play longer than others.
- Activity 2 has no impact on player retention in the first hours.
- After 10 hours players who practice Activity 2 have a higher probability to stop playing than others.

R package

- Full C++ code interfaced with R[3] via RCpp[4]
- Deal with data files bigger than RAM.

References

[1] David, C. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society*

[2] Martinussen, T. and Scheike, T. H. (2002). Efficient Estimation of Fixed and Time-Varying Covariate Effects in Multiplicative Intensity Models. *Scandinavian Journal of Statistics*

[3] R Core Team (2016). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing.*

[4] Eddelbuettel, D. and Francois, R (2011). Rcpp: Seamless R and C++ Integration. *Journal of Statistical Software*