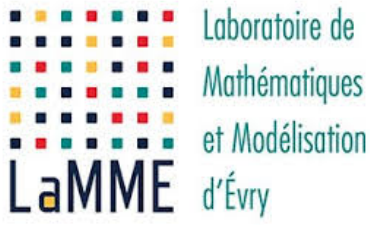
Grandes données de santé : méthodes et challenges

Agathe Guilloux
Professeure au LaMME - Université d'Évry - Paris Saclay



4 examples / 4 dimensions (1)

Clinic: Colo-rectal cancer

- Relapse free survival after tumor resection for colo-rectal cancer patients
- genomic and survival data for $\sim$500 patients

High-dimensional covariates

$$(n, p) \rightarrow (n, p)$$

4 examples / 4 dimensions (2)

Marketing: Monetization for free-to-play games

- Times of monetization for players until their giving-ups
- Several hours of game-play history for $\sim$1MM players

Large number of observations (individuals), time-dependent covariates

$$(n, p) \to \left(n, p, D\right)$$

4 examples / 4 dimensions (3)

Clinic : HEGP-APHP data warehouse

- Adverse events for patients in ARTEMIS cohort (hypertension)
- 25 years of medical history for $\sim$ 30000 patients

Large everything....

$$(n, p) \to \left(n, p, D, K\right)$$

4 examples / 4 dimensions (4)

Public health: SNIRAM data

- Adverse events in SNIRAM database
- 3 years history for $\sim$ 60 MM beneficiaries

Huge everything....

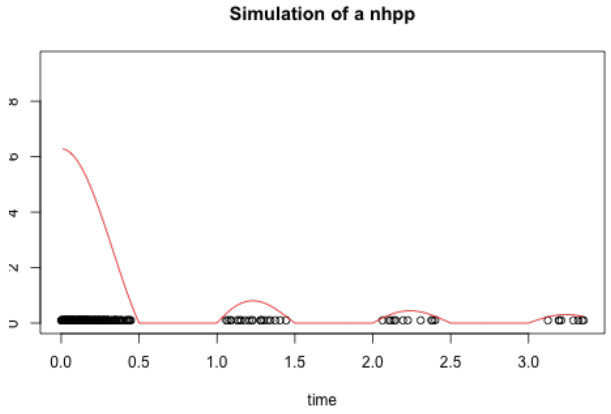
$$(n, p) \to \big(n, p, D, K\big)$$

Contents

Section: Time(s) to event(s) data, censoring and models

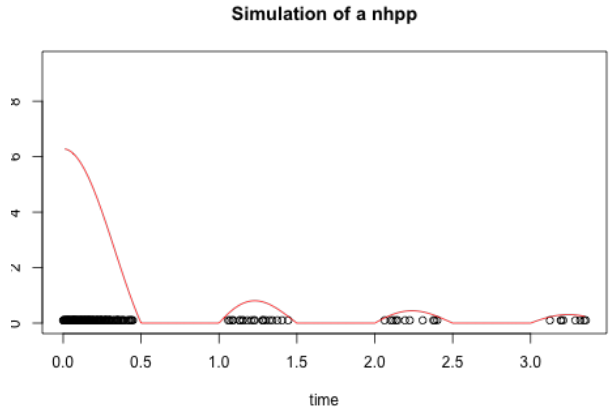
Time(s) to event(s) data

What are we observing ?



Time(s) to event(s) data

What are we observing ?

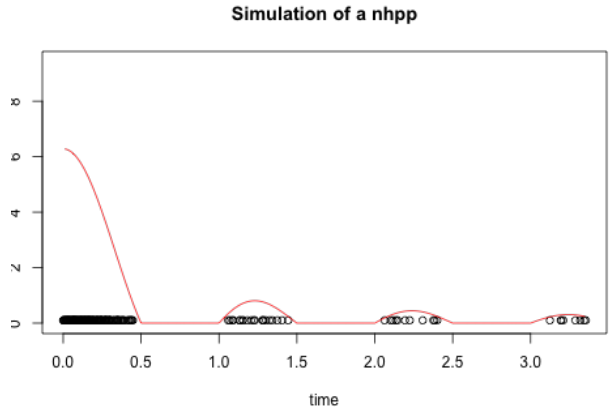


The higher the intensity, the more points we observe :

$$\lambda^\star(t) = \text{infinitesimal } \mathbb{P}(\text{event} \in [t, t + dt])$$

Time(s) to event(s) data

What are we observing ?



The higher the intensity, the more points we observe :

$$\lambda^\star(t) = \text{infinitesimal } \mathbb{P}(\text{event} \in [t, t + dt])$$

Construct a **counting process** $N^\star$ defined as

$$N^\star(t) = \text{number of observed events in } [0, t],$$

we'll say that $N^\star$ has intensity $\lambda^\star$.

One special case: at most one event

One event

Let T be a time of interest and construct

$$N^\star(t) = I(T \leq t)$$

Clinic: Colo-rectal cancer (with the CdR Saint Antoine - INSERM/UPMC)

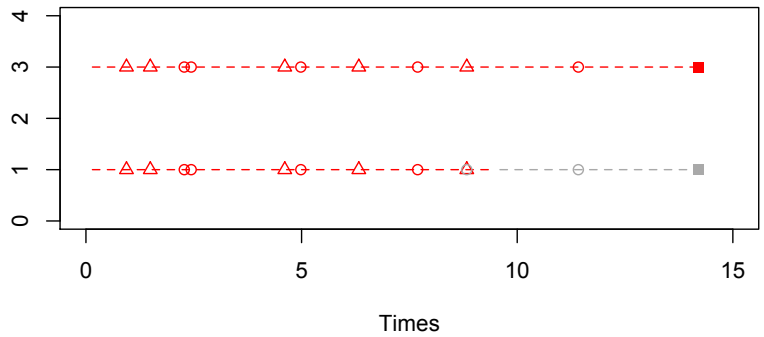
T is the time of recurrence of cancer

In this case

$$\lambda^\star(t) = \frac{f(t)}{1 - F(t)} = \text{infinitesimal } \mathbb{P}(T \in [t, t + dt] \mid T \geq t)$$

Censoring

We observe N only until a censoring C occurs.



Marketing: Monetization for free-to-play games

Times of monetization for players until their **giving-ups**

One special case: at most one event and censoring

Right censoring

Let

- T be a time of interest
- C a censoring time independent of T

We observe

$$T^C = T \wedge C \text{ and } \delta = \mathbb{1}(T \leq C).$$

In terms of counting processes, this is equivalent to observing

$$N(t) = \mathbb{1}(T^C \leq t, \delta = 1) \text{ and } Y(t) = \mathbb{1}(T^C \geq t).$$

Section: Covariates

Covariates

Time independent covariates

Clinic: Colo-rectal cancer
Genomic data

Public health: SNIRAM data
The gender of the beneficiary

Time dependent covariates

Monetization for games
The weapons the player owns

Public health: SNIRAM data
The drugs the beneficiary is on

We'll see that this implies that $n \to D \times n$.

Models for the intensity

We want the intensity to depend on the covariates at time t

$$\lambda^\star(t) \to \lambda^\star(t, X(t)).$$

The Cox model [Cox72]

$$\lambda^*(t) = \alpha^\star(t) \exp(X(t)\beta^\star)$$

Stanford Heart Transplant data [KP11]

Survival of patients on the waiting list for the Stanford heart transplant program.

- accept.dt: acceptance into program
- fustat: dead or alive
- surgery: prior bypass surgery
- age: age (in years)
- futime: follow-up time
- wait.time: time before transplant
- transplant: transplant indicator

```
##    accept.dt fustat surgery      age futime wait.time transplant
## 1 1967-11-15      1       0 30.84463     49        NA          0
## 2 1968-01-02      1       0 51.83573      5        NA          0
## 3 1968-01-06      1       0 54.29706     15         0          1
## 4 1968-03-28      1       0 40.26283     38        35          1
## 5 1968-05-10      1       0 20.78576     17        NA          0
## 6 1968-06-13      1       0 54.59548      2        NA          0
```

Stanford Heart Transplant data [KP11]

Survival of patients on the waiting list for the Stanford heart transplant program.

- accept.dt: acceptance into program
- fustat: dead or alive $\rightarrow \delta$
- surgery: prior bypass surgery
- age: age (in years)
- futime: follow-up time $\rightarrow T^C$
- wait.time: time before transplant
- transplant: transplant indicator

```
##    accept.dt fustat surgery      age futime wait.time transplant
## 1 1967-11-15      1       0 30.84463     49        NA          0
## 2 1968-01-02      1       0 51.83573      5        NA          0
## 3 1968-01-06      1       0 54.29706     15         0          1
## 4 1968-03-28      1       0 40.26283     38        35          1
## 5 1968-05-10      1       0 20.78576     17        NA          0
## 6 1968-06-13      1       0 54.59548      2        NA          0
```

Stanford Heart Transplant data [KP11]

Survival of patients on the waiting list for the Stanford heart transplant program.

- accept.dt: acceptance into program $\rightarrow$ **time independent covariate**
- fustat: dead or alive $\rightarrow \delta$
- surgery: prior bypass surgery $\rightarrow$ **time independent covariate**
- age: age (in years) $\rightarrow$ **time independent covariate**
- futime: follow-up time $\rightarrow T^C$
- wait.time: time before transplant
- transplant: transplant indicator

```
##    accept.dt fustat surgery      age futime wait.time transplant
## 1 1967-11-15      1       0 30.84463     49        NA          0
## 2 1968-01-02      1       0 51.83573      5        NA          0
## 3 1968-01-06      1       0 54.29706     15         0          1
## 4 1968-03-28      1       0 40.26283     38        35          1
## 5 1968-05-10      1       0 20.78576     17        NA          0
## 6 1968-06-13      1       0 54.59548      2        NA          0
```

Stanford Heart Transplant data [KP11]

Survival of patients on the waiting list for the Stanford heart transplant program.

- accept.dt: acceptance into program $\rightarrow$ **time independent covariate**
- fustat: dead or alive $\rightarrow \delta$
- surgery: prior bypass surgery $\rightarrow$ **time independent covariate**
- age: age (in years) $\rightarrow$ **time independent covariate**
- futime: follow-up time $\rightarrow T^C$
- wait.time: time before transplant $\rightarrow$ **time dependent covariate**
- transplant: transplant indicator $\rightarrow$ **time dependent covariate**

```
##    accept.dt fustat surgery      age futime wait.time transplant
## 1 1967-11-15      1       0 30.84463     49        NA          0
## 2 1968-01-02      1       0 51.83573      5        NA          0
## 3 1968-01-06      1       0 54.29706     15         0          1
## 4 1968-03-28      1       0 40.26283     38        35          1
## 5 1968-05-10      1       0 20.78576     17        NA          0
## 6 1968-06-13      1       0 54.59548      2        NA          0
```

Stanford Heart Transplant data [KP11] (2)

On transforme la variable `accept.dt` en gardant seulement l'année.

```
jasa = dplyr::mutate(jasa, accept.yr = year(ymd(accept.dt)))
jasa = dplyr::select(jasa,   fustat, surgery,age,futime,
                wait.time,transplant,accept.yr)

head(jasa)
```

```
##   fustat surgery      age futime wait.time transplant accept.yr
## 1      1       0 30.84463     49        NA          0      1967
## 2      1       0 51.83573      5        NA          0      1968
## 3      1       0 54.29706     15         0          1      1968
## 4      1       0 40.26283     38        35          1      1968
## 5      1       0 20.78576     17        NA          0      1968
## 6      1       0 54.59548      2        NA          0      1968
```

Example with time independent covariates

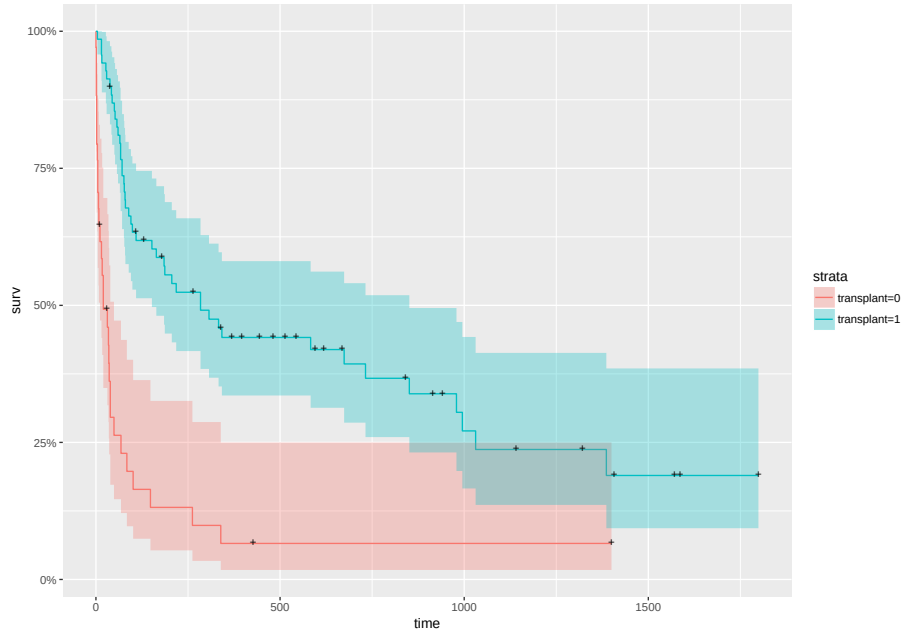
```
coxph(Surv(futime,fustat) ~ accept.yr + surgery + age, data = jasa)
```

```
## Call:
## coxph(formula = Surv(futime, fustat) ~ accept.yr + surgery +
##     age, data = jasa)
##
##               coef exp(coef) se(coef)     z     p
## accept.yr -0.1320    0.8764   0.0681 -1.94 0.053
## surgery   -0.6427    0.5259   0.3673 -1.75 0.080
## age        0.0276    1.0280   0.0134  2.06 0.039
##
## Likelihood ratio test=14.5  on 3 df, p=0.00226
## n= 103, number of events= 75
```

Example with time dependent covariates: false model

- transplant: transplant indicator $\rightarrow$ **time dependent covariate**

```
autoplot(survfit(Surv(futime,fustat) ~transplant , data = jasa))
```



"The key rule for time dependent covariates in a Cox model is simple and essentially the same as that for gambling: *you cannot look into the future*." [TCA17]

Example with time dependent covariates: false model (2)

```
coxph(Surv(futime,fustat) ~ surgery + transplant + age , data = jasa)
```

```
## Call:
## coxph(formula = Surv(futime, fustat) ~ surgery + transplant +
##     age, data = jasa)
##
##                coef exp(coef) se(coef)     z       p
## surgery    -0.4190    0.6577   0.3712 -1.13    0.26
## transplant -1.7171    0.1796   0.2785 -6.16 7.1e-10
## age         0.0589    1.0607   0.0150  3.91 9.1e-05
##
## Likelihood ratio test=45.9  on 3 df, p=6.11e-10
## n= 103, number of events= 75
```

A new format for time dependent covariates: start-stop

```
##     id start stop event transplant        age      year surgery
##     1     0   49     1          0 -17.155373 0.1232033       0
##     2     0    5     1          0   3.835729 0.2546201       0
##     3     0   15     1          1   6.297057 0.2655715       0
##     4     0   35     0          0  -7.737166 0.4900753       0
##     4    35   38     1          1  -7.737166 0.4900753       0
##     5     0   17     1          0 -27.214237 0.6078029       0
```

Notice that for individual 4, we have

- with the old format

```
##   fustat          age futime wait.time transplant
## 4      1     40.26283     38        35          1
```

- with the new format

```
##     id start stop event transplant
##     4     0   35     0          0
##     4    35   38     1          1
```

A new format for time dependent covariates: start-stop (2)

- False model

```
## coxph(formula = Surv(futime, fustat) ~ surgery + transplant +
##     age, data = jasa)
##
##                coef exp(coef) se(coef)     z       p
## surgery    -0.4190    0.6577   0.3712 -1.13    0.26
## transplant -1.7171    0.1796   0.2785 -6.16 7.1e-10
## age         0.0589    1.0607   0.0150  3.91 9.1e-05\end{itemize}
```

- Start-stop model

```
## coxph(formula = Surv(start, stop, event) ~ age + surgery +
##     transplant, data = jasa1)
##
##                coef exp(coef) se(coef)     z     p
## age           0.0306    1.0310   0.0139  2.20 0.028
## surgery      -0.7733    0.4615   0.3597 -2.15 0.032
## transplant    0.0141    1.0142   0.3082  0.05 0.964
```

Section: Estimation

The data

We observe for $i = 1, \ldots, n$ i.i.d.

$$\Big(X_i(s) Y_i(s), N_i(s), Y_i(s), s \leq \tau \Big)$$

and we want to learn the influence of X on $t \mapsto \lambda^*(t, X(t))$.

The likelihood

$$- \frac{1}{n} \sum_{i=1}^{n} \{ \int_{[0,\tau]} \log(\lambda(t, X_i(t))) dN_i(t) - \int_{[0,\tau]} Y_i(t) \lambda(t, X_i(t)) dt \}$$

$$= - \frac{1}{n} \sum_{i=1}^{n} \{ \sum_{T_{i,k}} \delta_{i,k} \log(\lambda(t, X_i(T_{i,k}))) - \int_{[0,\tau]} Y_i(t) \lambda(t, X_i(t)) dt \}$$

To ease the notation, I'll consider that each individual has a most one event

$$- \frac{1}{n} \sum_{i=1}^{n} \{ \delta_i \log(\lambda(t, X_i(T_i^C))) - \int_{[0,\tau]} Y_i(t) \lambda(t, X_i(t)) dt \}$$

Partial likelihood

In the Cox model[Cox72],

$$\lambda^*(t) = \alpha^\star(t) \exp(X(t)\beta^\star),$$

we can estimate $\beta^\star$ only with the partial likelihood (that's what `coxph` does)

$$\begin{aligned}
\ell_n^P(\beta) &= -\frac{1}{n}\sum_{i=1}^{n} \delta_i \log \frac{\exp(X_i(T_i^C)\beta)}{\frac{1}{n}\sum_{j:T_j^C \geq T_i^C} \exp(X_j(T_i)\beta)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \delta_i \Big\{ - X_i(T_i^C)\beta + \log\Big(\sum_{j:T_j^C \geq T_i^C} \exp(X_j(T_i^C)\beta)\Big)\Big\}.
\end{aligned}$$

Large p

When p grows, one can consider to add a lasso penalty:

$$\ell_n^P(\beta) + \gamma \sum_{j=1}^{p} |\beta_j|$$

or an elastic-net penalty

$$\ell_n^P(\beta) + \gamma\big(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^{p} |\beta_j|^2\big).$$

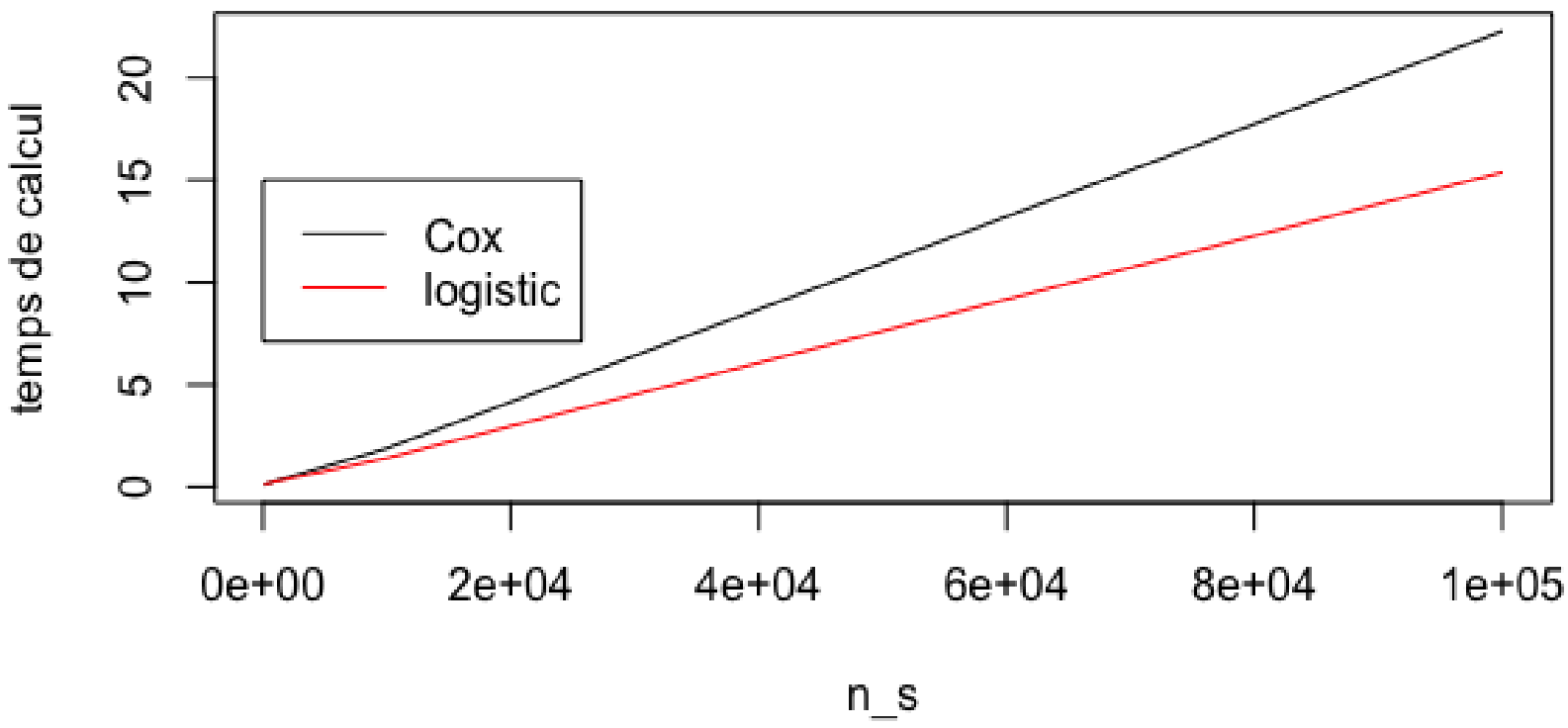
```
data("nki70")
model_matrix = model.matrix( ~ as.factor(Grade)  + . - Grade - 1
                              , data = nki70[3:77])

X = model_matrix[,-1]

elasticnet_solution = cv.glmnet(X,Surv(nki70$time, nki70$event),
        family = "cox" , alpha = 0.5,
        penalty.factor = c(rep(0,6),rep(1,70)))

coef(elasticnet_solution)
```

Few additional remarks

- For the selection,
  - see also“bo-lasso” ([Bac08]) and “stability selection” ([MB10])
  - multi-tests procedures ([DvdL08]) are developed for the Cox model in the package `multtest`
- In very large dimension, a ”screening” is mandatory
  - “Sure Independence Screening” ([FL08]) for Cox model in [ZL12]
  - others “screenings” in `ahaz`
  - or [TBF+12] and [EVR10], adapted to the Cox model in [TBF+12].
- They are other methods in high dimension: PCA, PLS (cf. [WT10]), Survival forest (cf. [IKG+10, IKCM11])

Section: Algorithmic challenges (n)

Timings: logistic regression vs Cox regression



Stochastic gradient descent

When n grows, stochastic algorithms are usually considered. For the logistic regression :

$$\ell_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-Y_i X_i \beta))$$

- Pick $i \sim \mathcal{U}[n]$
- update $\beta^{(t)}$

$$\beta^{(t+1)} = \beta^{(t)} - \eta_t \nabla_\beta \Big(\log(1 + \exp(-Y_i X_i \beta) \Big).$$

Problem : the partial losses are not adequate

- Partial likelihood

$$\begin{aligned}\ell_n^P(\beta) &= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\exp(X_i(T_i)\beta)}{\frac{1}{n}\sum_{j:T_j\geq T_i}\exp(X_j(T_i)\beta)}\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\ell_n^P(\beta)\right)_{(i)}\\ &= \frac{1}{n}\sum_{i=1}^{n}\Big\{-X_i(T_i)\beta+\log\Big(\sum_{j:T_j\geq T_i}\exp(X_j(T_i)\beta)\Big)\Big\}.\end{aligned}$$

- update $\beta^{(t)}$

$$\beta^{(t+1)} = \beta^{(t)} - \eta_t\nabla_\beta\Big(-X_i(T_i)\beta+\log\big(\sum_{j:T_j\geq T_i}\exp(X_j(T_i)\beta)\big)\Big).$$

1-step procedures and Poisson regression (1)

Likelihood

$$\sum_{i=1}^{n}\{\int_{[0,\tau]}\log(\lambda(t,X_i(t)\beta(t)))dN_i(t)-\int_{[0,\tau]}Y_i(t)\lambda(t,X_i(t)\beta(t))dt\}$$

$$=\sum_{i=1}^{n}\sum_{d=1}^{D}\{\int_{I_d}\log(\lambda(t,X_i(t)\beta(t)))dN_i(t)-\int_{I_d}Y_i(t)\lambda(t,X_i(t)\beta(t))dt\}$$

$$=\ldots$$

1-step procedures and Poisson regression (2)

$$\sum_{i=1}^{n}\sum_{d=1}^{D}\{\int_{I_d}\log(\alpha(t)\exp(X_i(t)\beta(t)))dN_i(t) - \int_{I_d} Y_i(t)\alpha(t)\exp(X_i(t)\beta(t))dt\}$$

Assuming that $X_i(t)$ is constant on small interval, and with sieves proposals, we get

$$\sum_{i=1}^{n}\sum_{d=1}^{D}\log(\alpha(I_d)\exp(X_i(I_d)\beta(I_d)))N_i(I_d) - |Y_i(I_d)|\alpha(I_d)\exp(X_i(I_d)\beta(I_d))$$

Poisson regression with $n \times D$ observations and $p \times D$ covariates ! But $n \times D$ and $p \times D$ are (very) large.

Example : simulated data

J'ai reproduit des données de santé de type pharmacovigilance. Elles contiennent des observations pour

- 10000 individus
- sur 5 ans avec une mesure par semaine.

On s'intéresse à la **survenue d'une pathologie**. On suspecte qu'un traitement augmente ce risque, proportionnellement à la dose cumulée reçue par l'individu. Chaque individu /bf commence (à un moment aléatoire sur les 5 ans) le traitement.
Les autres variables mesurées sont le sexe de l'individu et une variable continue (qui dans la simulation n'a pas d'influence sur le risque).
Pour chaque individu et chaque semaine soit $10000 \times 5 \times 52 = 2600000$ lignes de données.
Seuls 26 individus ont développé (au moins une fois) la pathologie.

Example : simulated data (2)

- l'identifiant de l'individu (`ind`)
- le numéro de la semaine (`week`)
- la dose cumulée (`cum_dose`)
- le sexe (`sex`)
- la variable continue (`var`)
- le nombre d'évènements survenus dans la semaine (`nevent`)
- une variable d'identification de la ligne (`id`)

```
##   ind week cum_dose sex        var nevent id
## 1   1    1        0   0  2.2712990      0  1
## 2   1    2        0   0 -0.4266551      0  2
## 3   1    3        0   0 -0.4064611      0  3
## 4   1    4        0   0 -0.7696531      0  4
## 5   1    5        0   0  1.7527771      0  5
## 6   1    6        0   0  1.1024576      0  6
```

Cases only study

Ordered statistics property of Poisson processes, see [Kin93]

In a Poisson regression model,

- when conditioning by the number of adverse events for each individuals with a least an event
- the number of events for each individual in each interval has a multinomial distribution.

This is known as **SCCS** (self-controlled case series), see http://statistics.open.ac.uk/sccs/r.htm.

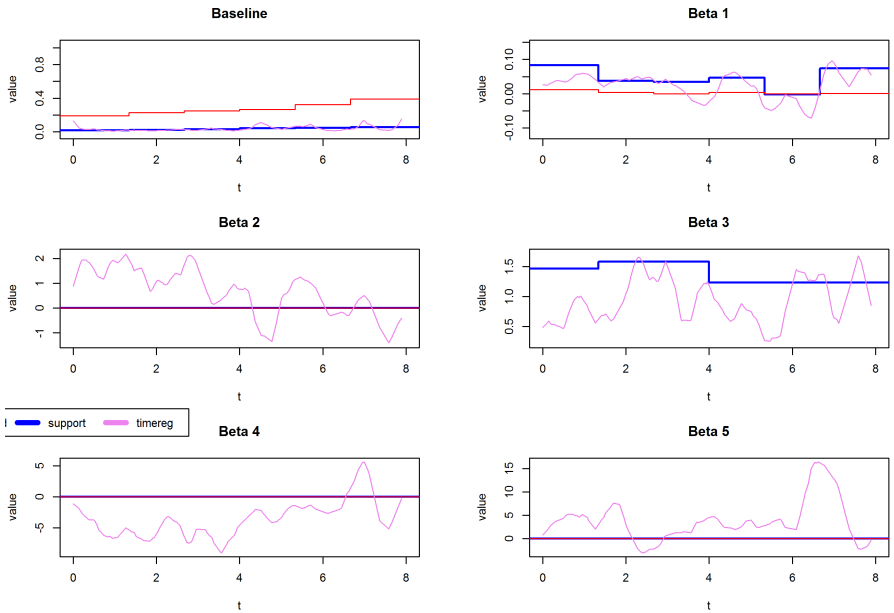
Section: Other challenges

Constant $\beta^\star$? Really ???

We considered that

$$\lambda^*(t) = \lambda^*(t, X(t)\beta^*).$$

But for long history (several years of data), we'd like $\beta^*(t)$!
This is possible but, in this case, $p \to D \times p$.



The K dimension

Clinic : HEGP-APHP data warehouse

- Adverse events for patients in ARTEMIS cohort (hypertension)
- 25 years of medical history for $\sim$ 30000 patients

Large everything....

$$(n, p) \to \big(n, p, D, K\big)$$

References I


F. Bach, *Bolasso: Model consistent lasso estimation through the bootstrap*, ICML, 2008.


David R. Cox, *Regression models and life-tables*, J. Roy. Statist. Soc. Ser. B **34** (1972), 187–220, With discussion by F. Downton, Richard Peto, D. J. Bartholomew, D. V. Lindley, P. W. Glassborow, D. E. Barton, Susannah Howard, B. Benjamin, John J. Gart, L. D. Meshalkin, A. R. Kagan, M. Zelen, R. E. Barlow, Jack Kalbfleisch, R. L. Prentice and Norman Breslow, and a reply by D. R. Cox. MR MR0341758 (49 #6504)


Sandrine Dudoit and Mark J. van der Laan, *Multiple testing procedures with applications to genomics*, Springer Series in Statistics, Springer, New York, 2008. MR 2373771 (2009j:62004)


Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani, *Safe feature elimination in sparse supervised learning*, Tech. Report UC/EECS-2010-126, EECS Dept., University of California at Berkeley, September 2010.


Jianqing Fan and Jinchi Lv, *Sure independence screening for ultrahigh dimensional feature space*, J. R. Stat. Soc. Ser. B Stat. Methodol. **70** (2008), no. 5, 849–911. MR 2530322

References II


Hemant Ishwaran, Udaya B. Kogalur, Xi Chen, and Andy J. Minn, *Random survival forests for high-dimensional data*, Stat. Anal. Data Min. **4** (2011), no. 1, 115–132. MR 2814504 (2012g:62253)


Hemant Ishwaran, Udaya B. Kogalur, Eiran Z. Gorodeski, Andy J. Minn, and Michael S. Lauer, *High-dimensional variable selection for survival data*, J. Amer. Statist. Assoc. **105** (2010), no. 489, 205–217. MR 2757200


John Frank Charles Kingman, *Poisson processes*, Wiley Online Library, 1993.


John D Kalbfleisch and Ross L Prentice, *The statistical analysis of failure time data*, vol. 360, John Wiley & Sons, 2011.


Nicolai Meinshausen and Peter Bühlmann, *Stability selection*, J. R. Stat. Soc. Ser. B Stat. Methodol. **72** (2010), no. 4, 417–473. MR 2758523


Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J. Tibshirani, *Strong rules for discarding predictors in lasso-type problems*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **74** (2012), no. 2, 245–266. MR 2899862


Terry Therneau, Cindy Crowson, and Elizabeth Atkinson, *Using time dependent covariates and time dependent coefficients in the cox model*, Survival Vignettes (2017).

References III


D.M. Witten and R. Tibshirani, *Survival analysis with high-dimensional covariates*, Statistical methods in medical research **19** (2010), no. 1, 29.


Sihai Dave Zhao and Yi Li, *Principled sure independence screening for Cox models with ultra-high-dimensional covariates*, J. Multivariate Anal. **105** (2012), 397–411. MR 2877525