Learning from longitudinal observational data: results and challenges

Agathe Guilloux

Laboratoire de Statistique Théorique et Appliquée - Université Pierre et Marie Curie

23 novembre 2015

5 examples / 4 dimensions (1)

Clinic : Colorectal cancer

- Relapse free survival after tumor resection for colorectal cancer patients
- genomic and survival data for ∼500 patients

High-dimensional covariates

$$(n, p) \to \big(n, \mathbf{\mathit{p}}\big)$$

5 examples / 4 dimensions (4)

Clinic : HEGP-APHP data warehouse

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

Large everything....

$$(n, p) \rightarrow \Big(n, p, D, K\Big)$$

Why do we need more research ? (1)

The best dataset on statins outcomes and side effects remains unused

24 July 2015
stephen black
data scientist
biggleswade, bedford

The BMJ is right to call for the release of more clinical trial data on statins. But perhaps it should also push to use the huge volume of data the NHS already collects.

The NHS in England issues nearly 70m prescriptions for statins every year for perhaps 5 million people. The NHS could, if the motivation and will existed, use the joined-up medical records of those patients to explore many of the issues unresolved in the available clinical trial data. True, GPs don't use proper randomisation to allocate the drugs to their patients (though they could as Ben Goldacre has proposed) but their preferences differ leading to large differences in the mix of statins prescribed in different GP practices. This, in principle, provides an enormous dataset to explore the possible benefits and side effects.

Such a non-randomised dataset would suffer from more confounding than a well-designed clinical trial but the enormous scale and the availability of comprehensive data about the patients might provide information never sought in clinical trial designs.

Moreover, since the NHS knows not just the information recorded by GPs about their patients but also whether patients are admitted to hospitals with the sort of conditions statins are designed to prevent, we might also be able to estimate not just the impact on blood lipids or the incidence of side effects but on the actual primary outcomes.

Perhaps the BMJ should also be campaigning for an NHS that routinely uses the data it collects about patients to be used for the benefit of those patients instead of sitting unused because of overblown fears about confidentiality.

Competing interests: No competing interests

Why do we need more research ? (2)

- High-dimensional, time-dependent covariates $p \times D$
  $\rightarrow$ which sparsity ?
- Large n (and p and D)
  $\rightarrow$ which algorithms ?
- Large K (multi-tasks learning)
  $\rightarrow$ which algorithms/which sparsity ?

The role of statistics

- Statistical modelling and inference
  - To develop “raisonnable” models
  - To develop algorithms to estimate in these models
  - For hypothesis generation : detection of possible effects on the risk
  - (and prediction)
- Mathematical statistics
  - To control the errors

A. Guilloux

Introduction
The data

Data and estimation criterions
The data
Losses
Models

Estimation in high dimension (p)
Risks bound for β^*
Risks bound for α^*

Temporal sparsity (D)
Too simple ...
Discrete time
Continuous time

Algorithmical challenges (n)
MCMC
Poisson regression

Other challenges

(Marked) points processes on $[0, T]$

Modelisation and assumptions

We observe $N^* = (N^{*,1}, \dots, N^{*,K})$ a K-variate counting process on $[0, T]$ and

- T is a terminal event, which prevents from observing N^* past it
- two components have (p.s.) no simultaneous jumps
- N^* has a càglàd K-variate intensity

$$\lambda^* \mathbb{1}(t \leq T) = (\lambda^{*,1} \mathbb{1}(t \leq T), \dots, \lambda^{*,K} \mathbb{1}(t \leq T))$$

Clinic : Colorectal cancer

- $K = 1$
- $N^*(t) = \mathbb{1}(T \leq t)$

Covariate processes

Covariates

Together with N^*, we observe a p-variate bounded predictable process of covariates $\{X(t), 0 \leq t \leq T\}$ which influences the intensity

$$\forall k, \forall t \leq T, \lambda^{*,k}(t, X(t)) = \mathbb{P}(dN^{*,k}(t) = 1 | N^{*,k}(s), s \leq t, X(t))$$

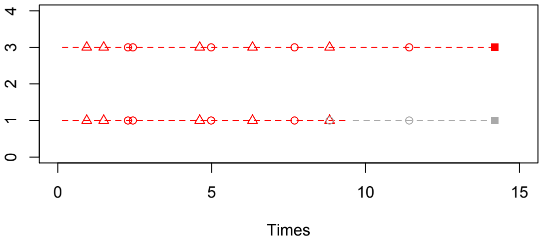
so that for all $t \leq T$

$$N^{*,k}(t) = \int_{[0,t]} \lambda^{*,k}(s, X(s)) ds + M^{*,k}(t)$$

where M^* is a vector of orthogonal martingales in

$$\mathcal{M}_2\Big(\big(\mathcal{F}_t = \sigma(N^*(s), \mathbb{1}(s \leq T), X(s), s \leq t)\big)\Big).$$

Independent filtering



Censoring is the simplest problem of observation

When the end of study (or lost of follow up) appends at time $C \leq T$, we observe for $t \leq T$

$$N(t) = \int_{[0,t]} \mathbb{1}(s \leq C)\, dN^*(s)$$

Independent filtering

Key assumption : independent filtering

More generally, we observe

$$N(t) = \int_{[0,t]} C(s) dN^*(s)$$

where C is a predictable (and observable) filtering process for $(\mathcal{G}_t) \supseteq (\mathcal{F}_t)$.
We assume that M defined as

$$M(t) = N(t) - \int_{[0,t]} C(s) \mathbb{1}(s \leq T) \lambda^*(s) ds = N(t) - \int_{[0,t]} Y(s) \lambda^*(s) ds$$

is a vector of orthogonal martingales in $\mathcal{M}_2\big((\mathcal{G}_t)\big)$.

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 each $t \mapsto \lambda^{*,k}(t, X(t))$.

"Signal + error" representation

$$
\left(\begin{array}{c|c|c} dN_1^1(t) & \dots & dN_1^K(t) \\ \dots & \dots & \dots \\ dN_n^1(t) & \dots & dN_n^K(t) \end{array}\right)
$$

$$
= \underbrace{Y(t)\left(\begin{array}{c|c|c} \lambda^{*,1}(t, X_1(t)) & \dots & \lambda^{*,K}(t, X_1(t)) \\ \dots & \dots & \dots \\ \lambda^{*,1}(t, X_n(t)) & \dots & \lambda^{*,K}(t, X_n(t)) \end{array}\right)}_{\text{"signal"}} dt
$$

$$
+ \underbrace{\left(\begin{array}{c|c|c} dM_1^1(t) & \dots & dM_1^K(t) \\ \dots & \dots & \dots \\ dM_n^1(t) & \dots & dM_n^K(t) \end{array}\right)}_{\text{"error"}}.
$$

Two losses

Least squares

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left\| N_i^k(\tau) - \int_{[0,\tau]} Y_i(t)\lambda^k(t, X_i(t))\,dt \right\|^2$$

Likelihood

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

From now on $K = 1$

Non-parametric models

Non-parametric estimation

Antoniadis, Grégoire and Nason (1999), Brunel and Comte (2005), Reynaud-Bouret (2006)

Non-parametric estimation with covariates - Comte Gaïffas G. (2010)

When estimating λ^* as a function of t and $X(t)$ with $X(t) \in \mathbb{R}^p$, the rate of convergence is of order $n^{-2\bar{\gamma}/(p+1+2\bar{\gamma})}$ ($\bar{\gamma}$ is the anisotropic regularity of λ^*).

Semiparametric models

Two famous models

- the Cox model (1972)

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

- the Aalen model (1980)

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

Cox model : Huang et al. (2013) and with S. Lemler and ML Taupin (2014a,b), Aalen model : with S. Gaïffas (2012)

Cox partial likelihood

$$\ell_n^P(\beta) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\log\frac{\exp(X(t)\beta)}{S_n(\beta,t)}\,dN_i(t)$$

with $S_n(\beta,t) = \dfrac{1}{n}\sum_{i=1}^{n} Y_i(t)\exp(X(t)\beta)$.

Estimation of β^*

Assume that $\|\beta^*\|_0 << p$ and define $\hat{\beta} = \underset{\beta \in \mathcal{B}(0,R)}{\arg\min} \{\ell_n^P(\beta) + \gamma|\beta|_1\}$.

Proposition [Huang et al. (2013)/G, Lemler, Taupin (2015a)]

Under mild assumptions, with large probability :

$$|\hat{\beta} - \beta^*|_2^2 \sim \|\beta^*\|_0 \frac{\log(pn^k)}{n}.$$

We have $\lambda^*(t, X(t)) = \alpha^*(t) \exp(X(t)\beta^*)$.

Kernel estimation of α^*

Define

$$\hat{\alpha}_h^{\hat{\boldsymbol{\beta}}}(t) = \frac{1}{nh}\sum_{i=1}^{n}\int_0^{\tau} K\left(\frac{t-u}{h}\right)\frac{\mathbb{1}_{\{\bar{Y}(u)>0\}}}{S_n(u,\hat{\boldsymbol{\beta}})}\mathrm{d}N_i(u),$$

choose the bandwidth via the Goldenshluger and Lepski method.

Theorem [G, Lemler, Taupin (2015b)]

Under classical assumptions,

$$\mathbb{E}[||\hat{\alpha}_{\hat{h}^{\hat{\beta}}}^{\hat{\beta}} - \alpha^*||_2^2] \leq \frac{1}{n^{(2\gamma)/(2\gamma+1)}} + \|\beta^*\|_0 \frac{\log np}{n}$$

Asymptotic results (in small dimension) : Ramlau-Hansen (1983b) and Grégoire (1993).

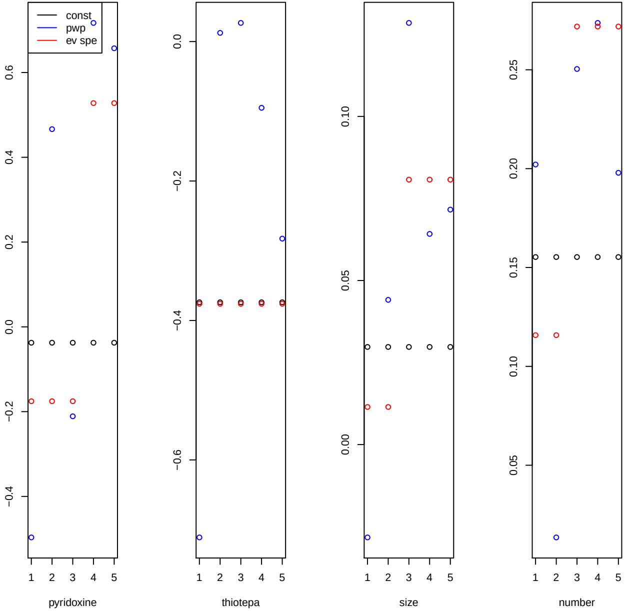
We have

$$dN(t) = \lambda^*(t)dt + dM(t)$$

and we considered that

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

We want $\beta^*(t)$! but not entirely non-parametrically.


const
pwp
ev spe
pyridoxine
thiotepa
size
number

Estimation (Bouaziz&G. 2014)

Define

$$\hat{\beta}_{\mathsf{TV}} = \arg \min_{\theta \in \mathbb{R}^{p*B}} \left\{ \ell_n^P(\beta) + \sum_{j=1}^{p} \sum_{b=1}^{B} |\beta^j(b) - \beta^j(b-1)| \right\}$$

Asymptotic results.

Algorithm

There exists a $p * B$ block-diagonal lower triangular matrix T such that

$$\beta = T\theta$$

and

$$\hat{\theta}_{\mathsf{TV}} = \arg \min_{\theta \in \mathbb{R}^{p*B}} \left\{ \ell_n^P(T\theta) + \gamma \sum_{j=1}^{p} \sum_{b=1}^{B} |\theta^j(b)| \right\}$$

Problem : the Hessian term is multiplied by up to B.

Consider that

$$\beta^j(t) = \sum_{k=1}^{D} \beta^j_k \mathbf{1}_{\left(\frac{k-1}{D}, \frac{k}{D}\right]}(t)$$

and

$$\hat{\beta}_{\mathsf{TV}} = \arg \min_{\theta \in \mathbb{R}^{p*B}} \left\{ \ell_n^P(\beta) + \gamma \sum_{j=1}^{p} \sum_{k=1}^{D} |\beta^j(b) - \beta^j(b-1)| \right\}.$$

Wa have

$$|\hat{\beta} - \beta^*|_2^2 \sim \|\beta^*\|_{TV0} \frac{\log(Dpn^k)}{n}.$$

Algorithm

- The gradient of the Cox partial likelihood is Lipschitz, proximal gradient descent algorithms can be considered (e.g. ISTA or FISTA).
- The prox operator of our TV is (almost) linearly computable (cf. Condat 2013, AGG2014b)

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

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

- Gradients

$$\nabla\Big(\left(\ell_n^P(\beta)\right)_{(i)}\Big) = -X_i(T_i) + \sum_{j:X_j\geq T_i} X_j(T_i)\pi_\beta^i(j)$$

with

$$\pi_\beta^i(j) = \frac{\exp(X_j(T_i)\beta)}{\sum_{j':T_{j'}\geq T_i}\exp(X_{j'}(T_i)\beta)}, \quad \forall j : X_j \geq T_i.$$

Doubly stochastic

- Pick $i \sim \mathcal{U}[n]$
- $\widehat{\nabla} f_i(\beta^t) \leftarrow$ approximation of $\nabla f_i(\beta^t)$ using N_k Monte-Carlo Markov-Chain iterations

2SGPD - Achab, G., Gaïffas, Bacry (2015)

Our doubly stochastic algorithm has the convergence rate

$$\mathbb{E}[F(\tilde{\beta}^K)] - F(\beta^*) \leq D' \rho^K.$$

Atchade, Y. F., Fort, G., and Moulines, E. (2014), SAGA (Defazio el al., 2014), Prox-SVRG (Xiao and Zhang , 2014)

1-step procedures

Estimation with high dimensional covariates - Lemler (2014)

Procedure lasso to estimate $\lambda^*(t) = \alpha^*(t)\exp(X(t)\beta^*)$, the error term is of order

$$s_\beta \log p/n + s_\gamma \log M/n.$$

Likelihood

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

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

$$= \dots$$

1-step procedures

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

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 : no iRLS.

Multitask leanrning (K)

"Signal + error" representation

$$
\begin{pmatrix} dN_1^1(t) & \ldots & dN_1^K(t) \\ \ldots & & \\ dN_n^1(t) & \ldots & dN_n^K(t) \end{pmatrix}
$$

$$
= \underbrace{Y(t) \begin{pmatrix} \lambda^{*,1}(t, X_1(t)) & \ldots & \lambda^{*,K}(t, X_1(t)) \\ \ldots & & \\ \lambda^{*,1}(t, X_n(t)) & \ldots & \lambda^{*,K}(t, X_n(t)) \end{pmatrix} dt}_{\text{"signal"}}
$$

$$
+ \underbrace{\begin{pmatrix} dM_1^1(t) & \ldots & dM_1^K(t) \\ \ldots & & \\ dM_n^1(t) & \ldots & dM_n^K(t) \end{pmatrix}}_{\text{"error"}}.
$$

Featuring

We assumed

$$\lambda^*(t, X(t)) = \mathbb{P}(dN^*(t) = 1 | N^*(s), s \leq t, X(t)).$$

- When $X^j(t)$ is the dose (taken during a week) for a drug, this is not raisonable.
- Madigan et al. proposed models where $\int_0^t dX^j(s)\beta^j(t)$ replace $X^j(t)\beta^j(t)$: accumulated exposure.
- We propose $\int_0^t \beta^j(t-s)dX^j(s)$
- Or $(X(s), s \leq t)$?