ConvSCCS: a convolutional self-controlled case series model for lagged adverse event detection in large databases

Emmanuel Bacry[1], Stéphane Gaiffas[1], Agathe Guilloux[1], Maryan Morel*[1], Fanny Leroy[2]

1. Centre de Mathématiques Appliquées, École Polytechnique, Palaiseau, France
2. Caisse Nationale de l'Assurance Maladie des Travailleurs Salariés, Paris, France

* maryan.morel@polytechnique.edu

I. Overview

Motivation

Long-term adverse drug reactions (ADRs) are hard to anticipate, a postmarketing research effort is necessary.

- Current detection system based on spontaneous reports, under-reporting
- Large Observational healthcare Databases (LODs) contain large, rich data

Goal

Automatically detect unknown lagged ADRs in large healthcare observational databases. This is a non-trivial task, as

- ADRs are rare events occuring at random times
- Temporal pattern can take many forms
- Many confoundings, hard to control

Contributions

An interpretable multivariate scalable model with nice properties:

- Scalable
- Time-invariant
- Robust to nonlongitudinal confoundings

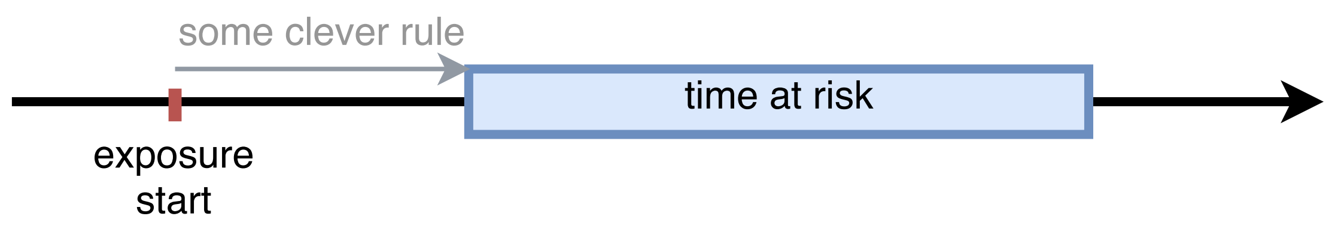
II. Self Control Case Series

Conditional Poisson model

$$P_{(0,T_i]}(t_{i,1},\dots,t_{i,n_i}|n_i) = n_i!\prod_{j=1}^{n_i}\frac{\lambda(t_{i,j}|X_i)}{\Lambda((0,T_i]|X_i)}$$

$$\lambda_i(t_i|X_i) = \exp(\rho_{t_i} + \beta x_i(t_i) + \psi_i + \gamma^T z_i)$$

- **Interpretability**
  - Estimate the relative incidence of longitudinal features
  - Patients act as their own control
- **Convenient for large observational databases study**
  - Ignore nonlongitudinal effects: robust to nonlongitudinal confoundings
  - Scalable: use only cases
  - Numerical stability: no overflow risk



III. Our approach

Discrete-time SCCS

Intensity assumed constant over Time intervals $I_b = [t_{b-1}, t_b),\ b = 1,\dots,B$

$$\lambda_b := \frac{\Lambda(I_b)}{|I_b|} \qquad \Lambda_b := \Lambda(I_b) = \int_{I_b}\lambda(t)dt \qquad \boldsymbol{\Lambda}_b := \Lambda\left(\bigcup_{b'=1}^{b} I_{b'}\right) = \sum_{b'=1}^{b}\Lambda_{b'}$$

Right censoring

We assume there is a right censoring time C

$$\widetilde{N}([0,t)) = N([0,t\wedge C)) \qquad \widetilde{Y}_b = \widetilde{N}(I_b)$$

$$\widetilde{\Lambda}_b = \Lambda_b\mathbf{1}_{t_b<t^*} + \Lambda^*\mathbf{1}_{t_b=t^*} \qquad \widetilde{\boldsymbol{\Lambda}}_b = \sum_{b'=1}^{b}\widetilde{\Lambda}_{b'}$$

$$ll\left(\psi,\theta,\alpha|X,Y,C\right) = \sum_{i=1}^{m}\sum_{b=1}^{B} y_b^i \log\left(\frac{\widetilde{\Lambda}_b^i}{\widetilde{\boldsymbol{\Lambda}}_B}\right)$$

Time dependence through a convolution

Weighted cumulative exposure (WCE) of feature j:

$$WCE^j(t) = \int_0^t X^j(u)\theta^j(t-u)\mathrm{d}u$$

We assume point exposures to avoid kernel overlapping issues, thus

$$WCE_b^j = \theta_{b-b_k}^j\mathbf{1}_{b\geq b_k}$$

$$\lambda_b(X) = e^{\phi_b + \sum_{j=0}^{J} WCE_b^j + \alpha Z_b}$$

Penalization

Group fused lasso over each feature's coefficient groups, ordered by time are penalized with Fused Lasso (FL) and Group Lasso (GL).

$$g(\theta) = \sum_{j=1}^{J}\sum_{b=0}^{B-1}|\theta_{b+1}^j - \theta_b^j| + ||\theta^j||_2$$

Estimation

We use proximal SVRG [5] to estimate the parameters. The proximal operator associated with our penalization is easy to compute:

- The proximal operator can be decomposed [6]: apply Prox$_{FL}$ then Prox$_{GL}$
- Efficient algorithm to compute Prox$_{FL}$ [2] and Prox$_{GL}$ has a closed form

IV. Simulation study

State of the art

SmoothSCCS [3] is the only SCCS model designed to detect lagged effects to our knowledge. It is a spline-based intensity to model the WCE, allowing only one longitudinal feature and a time drift at a time.

Protocol

- Correlated longitudinal features simulated from a Hawkes process using [1] on 730 time intervals
- ADRs events are simulated using a Poisson distribution
- Censoring time simulated with an exponential distribution
- 360 time intervals, max lag of the effect of 50 time intervals

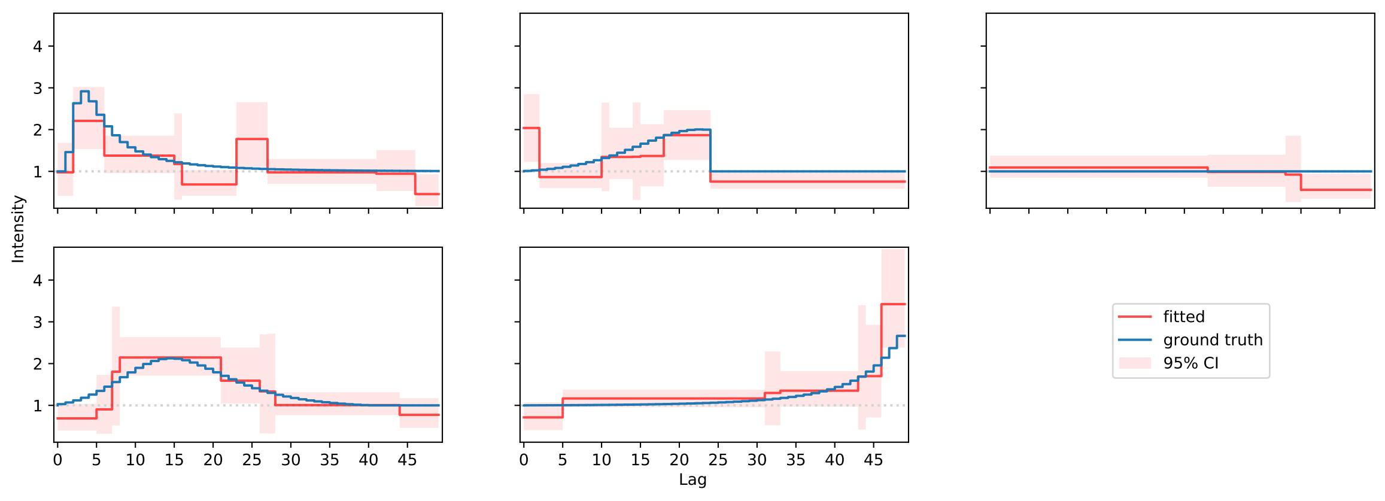


Figure 1: Estimation on simulated data (2000 samples)

Table 1: Performance comparison (MSE)

Model	effect	n = 500	n = 1000
SmoothSCCS	rapid	7.71 *(2.68)*	7.79 *(1.67)*
	early	7.43 *(3.07)*	7.73 *(1.97)*
	intermediate	7.45 *(2.33)*	7.40 *(1.41)*
	late	7.10 *(2.31)*	7.40 *(1.36)*
	null	6.74 *(1.97)*	7.51 *(1.79)*
	overall	**7.29 *(2.51)***	**7.57 *(1.66)***
ConvSCCS	rapid	2.36 *(1.20)*	2.04 *(1.05)*
	early	2.05 *(1.12)*	1.76 *(0.64)*
	intermediate	1.90 *(1.17)*	1.61 *(0.94)*
	late	1.57 *(1.07)*	1.42 *(1.04)*
	null	1.91 *(0.91)*	1.54 *(0.71)*
	overall	**1.96 *(1.12)***	**1.68 *(0.91)***

V. Application

- We apply our model the the SNIIRAM database (French healthcare system database)
- 3.5 millions of diabetic patients
- Study the effect of glucose-lowering drugs on bladder cancer risk
- Results consistent with [4]

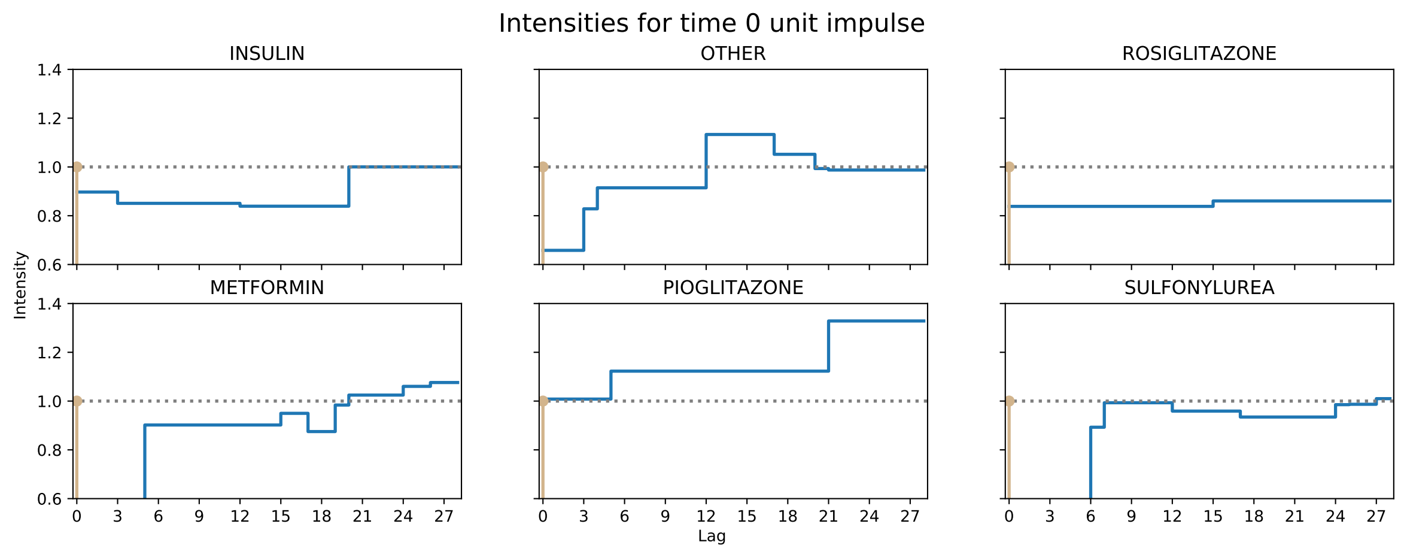


Figure 2: Estimated relative incidences of glucose-lowering drugs on bladder cancer risk

Limitations

- Drug exposures constrained by kernel size to avoid kernel overlap
- Need to use parametric bootstrap to compute approximate confidence intervals: very slow
- Feature design is easier than common medical studies practice, but is still an heavy task

References

[1] Emmanuel Bacry, Martin Bompaire, Stéphane Gaïffas, and Soren Poulsen. tick: a Python library for statistical learning, with a particular emphasis on time-dependent modeling. jul 2017.

[2] Laurent Condat. A Direct Algorithm for 1-D Total Variation Denoising. *IEEE Signal Processing Letters*, 20(11):1054–1057, nov 2013.

[3] Yonas Ghebremichael-Weldeselassie, Heather J. Whitaker, and C. Paddy Farrington. Flexible modelling of vaccine effect in self-controlled case series models. *Biometrical Journal*, 58(3):607–622, 2016.

[4] A. Neumann, A. Weill, P. Ricordeau, J. P. Fagot, F. Alla, and H. Allemand. Pioglitazone and risk of bladder cancer among diabetic patients in France: a population-based cohort study. *Diabetologia*, 55(7):1953–1962, 2012.

[5] Lin Xiao and Tong Zhang. A Proximal Stochastic Gradient Method with Progressive Variance Reduction. *SIAM Journal on Optimization*, 24:2057—-2075, 2014.

[6] Jiayu Zhou, Jun Liu, Vaibhav A. Narayan, and Jieping Ye. Modeling disease progression via fused sparse group lasso. *Kdd*, pages 1095–1103, 2012.

Aknowledgement

This research benefited from the CNAMTS-Polytechnique research contract. We thank Aurélie Bannay, Hélène Caillol, Joel Coste, Claude Gissot, Anke Neumann, Jérémie Rudant, and Alain Weill for interesting discussions and their help with the SNIIRAM data understanding. We also thank Firas Ben Sassi, Prosper Burq, Xristos Giastidis, Sathiya Kumar, and Daniel de Paula, the research engineers working on this project.