Kernel estimation of the intensity of Cox processes

Gaspar Massiot

IRMAR, ENS Rennes, Crest-Ensai
gaspar.massiot@ens-cachan.fr

Résumé. Les processus de comptage souvent notés $N = (N_t)_{t\in\mathbb{R}^+}$ sont utilisés dans de nombreuses applications en biostatistique, notamment dans l'étude des maladies chroniques. Dans le cadre de maladies respiratoires il est naturel de supposer que le nombre de consultations d'un patient suit un tel processus dont l'intensité dépend de covariables environnementales. Les processus de Cox (ou processus de Poisson doublement stochastiques) permettent de modéliser de telles situations. L'intensité aléatoire s'écrit alors sous la forme $\lambda(t) = \theta(t, Z_t)$ où θ est une fonction déterministe, $t \in \mathbb{R}^+$ est la variable de temps et $(Z_t)_{t\in\mathbb{R}^+}$ est le processus des covariables de dimension d. Lors d'une étude longitudinale sur n patients, on observe $(N_t^k, Z_t^k)_{t\in\mathbb{R}^+}$ pour $k = 1, \ldots, n$. On se propose d'estimer l'intensité du processus sur la base de ces observations et d'étudier les propriétés de l'estimateur construit.

Mots-clés. Processus de Cox, Estimateur de Nadaraya-Watson

Abstract. Counting processes often written $N = (N_t)_{t\in\mathbb{R}^+}$ are used in several applications of biostatistics, notably for the study of chronic diseases. In the case of respiratory illness it is natural to suppose that the count of the visits of a patient can be described by such a process which intensity depends on environmental covariates. Cox processes (also called doubly stochastic Poisson processes) allows to model such situations. The random intensity then writes $\lambda(t) = \theta(t, Z_t)$ where θ is a non-random function, $t \in \mathbb{R}^+$ is the time variable and $(Z_t)_{t\in\mathbb{R}^+}$ is the d-dimensional covariates process. For a longitudinal study over n patients, we observe $(N_t^k, Z_t^k)_{t\in\mathbb{R}^+}$ for $k = 1, \ldots, n$. The intention is to estimate the intensity of the process using these observations and to study the properties of this estimator.

Keywords. Cox Process, Nadaraya-Watson estimator

1 Introduction

A random process $N = (N_t)_{t\in[0,T]}$ defined until a fixed time (horizon) $T > 0$ is a counting process if its trajectories are almost surely (*a.s.*) right-continuous and piecewise constant, starting at 0 and if the jump size of N at time t is either 0 or 1 *a.s.*. It can model for instance the visits in a hospital. In the case of respiratory illness it is natural to suppose that the intensity of such a process depends on environmental covariates such as the pollen concentration of the air or the weather. Cox processes come in handy to take these covariates into consideration. They are formally defined as follows.

Definition 1 (Cox process)**.** *Consider a probability space $(\Omega, \mathcal{F}, P)$, carrying a counting process N as well as a non negative process $\lambda = (\lambda_t)_{t\in[0,T]}$. We say that N is a* **Cox process** *with intensity process λ if the relation*

$$\mathbb{E}\left[e^{iu(N_t-N_s)}\,\middle|\,\mathcal{F}_s^N \vee \mathcal{F}_\infty^\lambda\right] = e^{\Lambda_{s,t}(e^{iu}-1)}$$

holds for all $s < t$, where

$$\Lambda_{s,t} = \int_s^t \lambda_u \mathrm{d}u,$$

and $\mathcal{F}^N = (\mathcal{F}_u^N)_{u\in\mathbb{R}^+}, \mathcal{F}^\lambda = (\mathcal{F}_u^\lambda)_{u\in\mathbb{R}^+}$ are the natural filtrations of N and λ.

Roughly speaking, conditionally to the entire λ-trajectory, a Cox process is a Poisson process with intensity λ. Thus, a Cox process is a generalization of a Poisson process where the time-dependent intensity λ is itself a stochastic process. The process is named after the statistician David Cox, who first published the model in 1955 (Cox 1955a,b).

A kernel inference of Cox process data with large arrival rates is proposed in Zhang and Kou (2010). In some cases we feel the necessity to consider covariates in the model. To model the arrival process of claims an insurance company must for example take into consideration personal data of its customers or climatic, geographic data. The introduction of covariates such as the age, the sex or other physiological characteristics of the patients can also improve the inference for clinical trials problems. We consider a non-parametric estimation of λ based on i.i.d. observations when it can be written as $\lambda(t) = \theta(t, Z_t)$ where θ is a non-random function, $t \in \mathbb{R}^+$ is the time variable and $Z = (Z_t)_{t\in\mathbb{R}^+}$ is a d-dimensional covariates process.

2 Model and estimate

2.1 Model and data

We suppose that we observe n independent copies of $(N, Z) = (N_t, Z_t)_{t\in\mathbb{R}^+}$ on $[0, \tau]$. These observations are denoted by $(N^1, Z^1), \ldots, (N^n, Z^n)$. The jumping times of N^k are denoted by $T_1^k, T_2^k, \ldots$ Given these informations we propose to construct a kernel based estimator of $\theta : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}$.

2.2 Assumptions on the model

We suppose that for all t, Z_t has a density f_{Z_t} and denote $\phi(t, z) = f_{Z_t}(z)\theta(t, z)$.

To state our results we make several assumptions on the function θ and the processes N and Z which are summed up here.

(H1) $\theta(t, z)$ and $f_{Z_t}(z)$ are continuous functions of t and z;

(H2) For all $t \in \mathbb{R}^+$,

$$\int_0^t \mathbb{E}[\theta(u, Z_u)\theta(v, Z_v)|Z_t = y]\mathrm{d}v + \mathbb{E}[\theta(u, Z_u)|Z_t = y]$$

is a continuous function of (u, y);

(H3) $\mathbb{E}[\theta(u, Z_u)|Z_t = y]$ and $f_{Z_t}(y)$ are twice differentiable in u and y and have bounded continuous partial derivatives.

Remark 1. *A sufficient condition to get assumption* **(H2)** *is that θ is a Lipschitz continuous function and Z_t has a Lipschitz continuous density. In a same way, a sufficient condition to get assumption* **(H3)** *is that θ and f_{Z_t} are twice differentiable and have Lipschitz continuous partial derivatives and $f_{Z_t}(y) \neq 0$.*

2.3 Estimation strategy

Suppose that we observe $\theta(t, Z_t^k)$ for all $k = 0, \ldots, n$. Then our problem of estimation can simply be viewed as a regression estimation problem. A usual estimator would then

be the Nadaraya-Watson estimator

$$\hat{\theta}_{NW}(t,z) = \frac{\sum_{k=1}^{n} \theta(t, Z_t^k) H_\eta(z - Z_t^k)}{\sum_{l=1}^{n} H_\eta(z - Z_t^l)},$$

where $H : \mathbb{R}^d \to \mathbb{R}$ is a kernel, η is a bandwidth and $H_\eta(\cdot) = \frac{1}{\eta^d} H\left(\frac{\cdot}{\eta}\right)$.

As we do not observe $\theta(t, Z_t^k)$, we suggest to estimate the function $\theta(\cdot, Z_t^k)$ using another kernel $K : \mathbb{R} \to \mathbb{R}$ and another bandwidth h. This gives the estimator we study in this presentation

$$\hat{\theta}_{h,\eta}(t,z) = \frac{\sum_{k=1}^{n} \sum_{i=1}^{N_t^k} K_h(t - T_i^k) H_\eta(z - Z_t^k)}{\sum_{l=1}^{n} H_\eta(z - Z_t^l)}.$$

3 Main results

The following usual assumptions are made on the two kernels.

(H4) $\mathrm{supp} K = [0,1]$, $K \in \mathbb{L}^3(\mathbb{R})$, $\int K = 1$ and $\int uK(u)\mathrm{d}u = 0$.

(H5) $\mathrm{supp} H = [-1,1]^d$, $H \in \mathbb{L}^3([-1,1]^d)$, $\int H = 1$ and $\int uH(u)\mathrm{d}u = 0$ where the multiplication is coordinate-wise.

A study of the properties of our estimator is related below.

3.1 Mean square error

For fixed $t \leq \tau$ denote $\psi(u,y) = f_{Z_t}(y)\mathbb{E}[\theta(u, Z_u)|Z_t = y]$.

Under the hypothesis introduced before and if $f_{Z_t}(z) > 0, h \to 0, \eta \to 0$ and $nh\eta^d \to +\infty$ then the mean square error defined by

$$\mathrm{MSE}(t,z) = \left(\mathbb{E}\hat{\theta}_{h,\eta}(t,z) - \theta(t,z)\right)^2 + \mathrm{Var}\,\hat{\theta}_{h,\eta}(t,z),$$

varies as

$$\eta^4 + h^2\eta^2 + h\eta^3 + \frac{1}{nh\eta^d},$$

when $n \to +\infty$.

Its minimization over h and η leads to h and η of the same order $n^{-\frac{1}{5+d}}$. We can remark that we get the usual convergence rate for the kernel estimation of the regression function

with a covariate vector of dimension $d+1$. The fact that we get $h^* = \eta^*$ comes from the assumptions we make on the kernels which are quite similar. We should be able to get a different result by changing these assumptions and adapt the proof consequently.

3.2 Consistency and convergence in distribution

It is straightforward that our estimator is biased. We can nevertheless show that it is consistent and asymptotically normal.

Theorem 1. *Assume that* **(H1)**–**(H6)** *are satisfied.*
If $f_{Z_t}(z) > 0$, $h \to 0, \eta \to 0, nh\eta^d \to +\infty$ *then*

$$\hat{\theta}_{h,\eta}(t,z) \xrightarrow{\mathbb{P}} \theta(t,z)$$

Theorem 2. *Assume that* **(H1)**–**(H6)** *are satisfied.*
If $nh^3 \to 0$, $nh\eta^d(h\eta+\eta^2)^2 \to 0$ *and* $nh\eta^d \to +\infty$ *then*

$$(nh\eta^d)^{1/2} \frac{\hat{\theta}_{h,\eta}(t,z) - \theta(t,z)}{\left[\hat{v}_{h,\eta}(t,z) \int K^2 H^2 / \hat{f}_\eta^2(z)\right]^{1/2}} \xrightarrow{D} \mathcal{N}(0,1),$$

where $\hat{v}_{h,\eta}(t,z) = \hat{\phi}_{h,\eta}(t,z) \left(\int_0^t \mathbb{E}[\theta(v, Z_v) | Z_t = z] \mathrm{d}v + 1 \right)$.

Remark 2. *Remark that the assumption* $nh\eta^d(h\eta+\eta^2)^2 \to 0$ *writes* nh^{d+1+4} *for* $h = \eta$ *which is the exact assumption of the usual result for the kernel estimator of the regression.*

References

Cox, D.R. (1955a). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 129–164.

— (1955b). The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. 51, pp. 433–441.

Zhang, T. and S.C. Kou (2010). Nonparametric inference of doubly stochastic Poisson process data via the kernel method. *The annals of applied statistics* 4.4, p. 1913.