Petr Volf

Hazard rate model and statistical analysis of a compound point process

Kybernetika, Vol. 41 (2005), No. 6, [773]--786

Persistent URL: http://dml.cz/dmlcz/135692

KYBERNETIKA — VOLUME 41 (2005), NUMBER 6, PAGES 773 - 786

HAZARD RATE MODEL AND STATISTICAL ANALYSIS OF A COMPOUND POINT PROCESS

PETR VOLF

A stochastic process cumulating random increments at random moments is studied. We model it as a two-dimensional random point process and study advantages of such an approach. First, a rather general model allowing for the dependence of both components mutually as well as on covariates is formulated, then the case where the increments depend on time is analyzed with the aid of the multiplicative hazard regression model. Special attention is devoted to the problem of prediction of process behaviour. To this end, certain results on risk processes and crossing probabilities are recalled and utilized. The application deals with the process of financial transactions and the problem of detection of outlied trajectories.

Keywords: counting process, compound process, Cox regression model, financial series, intensity, prediction

AMS Subject Classification: 62G05, 62M09

1. INTRODUCTION, COMPOUND POINT PROCESS

We study the compound point process composed from random increments Y_j occurring at random time points $0 < T_1 < T_2 < \dots$. It can be written as a random sum $C(t) = \sum Y_j \cdot 1[T_j \leq t]$, $(C(0) = 0)$, or formally also as

$$C(t) = \int_0^t Y(s)\,\mathrm{d}N(s), \tag{1}$$

where $N(t)$ is the counting process corresponding to the random point process of time moments. The model is suitable for the description of many real-world engineering, environmental, economic, biological, and also financial processes (particularly from the field of insurance, cf. Asmussen, [3], Rolski et al. [12].

The model of a compound point process is standardly given by the intensity of a random point process and by the distribution of increments. A general model should consider also the mutual dependence of both process components as well as their dependence on covariates. Such a dependence uses the notion of the filtration $(\mathcal{F}(t)$, say), a non-decreasing sequence of σ-algebras defined on the sample space of

$\{N(s), I(s), Z(s), Y(s),\ 0 \le s \le t\}$. Here $I(t)$ and $Z(t)$ are $\mathcal{F}(t^-)$ measurable predictable processes, the indicator of observability of $C(t)$ and a covariate process, by $\mathcal{F}(t^-)$ we denote the left-continuous version of filtration, a 'history'. The behaviour of the counting process $N(t)$ is governed by its hazard rate $h(t, z)$. As regards the distribution of random variables $Y(t)$, it is as a rule assumed that the conditional distribution of $Y(t)$, given $\mathcal{F}(t^-)$, can be described via a density function $f(y, t, Z(t))$, i. e. it can depend on a set of covariates, too, and it possesses the first and second conditional moments $E(Y(t)|\mathcal{F}(t^-)) = m(t, Z(t))$, $var(Y(t)|\mathcal{F}(t^-)) = \sigma^2(t, Z(t))$. In Volf [15] it has been proven (for a slightly less general setting) that the rate of cumulation of $C(t)$, given by $k(t, z) = h(t, z) \cdot m(t, z)$, is estimable consistently, further, that the process $\int_0^t k(s, Z(s))I(s)\,\mathrm{d}s$ is the compensator of $C(t)$ and the residual martingale has the variance process $\langle \mathcal{M} \rangle(t) = \int_0^t (\sigma^2(s, Z(s)) + m^2(s, Z(s)))\, h(s, Z(s))I(s)\,\mathrm{d}s$.

The process of such a type can also be treated as a marked point process (e. g. Brémaud [4]), however, in our case such a characterization of increments as marks does not help us much to solve main problems, namely estimation, testing and outlied trajectories detection (though it can be useful for classification and selection of sub-processes).

Though the model outlined above is rather general, we still feel certain inconsistency of such a description combining two styles of characterization of probability distribution. That is why, the objective of the present paper is to propose a model characterizing both process parts, i. e. also the distribution of increments, with the help of hazard rates. We shall also consider the mutual dependence of both components via a regression model. Hence, the process can be regarded also as a 2D random point process, though its two components are not 'balanced', in the sense that the event 'increment' is not possible without the occurrence of the 'time point'. Processes of such a kind are studied for instance in the monograph of Jacod and Shiryaev [7].

The paper has the following parts: Part 2 collects certain well known results concerning the simple case of compound Poisson process, which will be used in further sections, namely the results on the ruin probability problem and the construction of prediction lines. Then, in Part 3, a quite general model with covariates will be formulated. Further, Part 4 presents a main results of the paper, it studies a particular case with time-dependent increments described with the aid of a nonparametric version of the multiplicative hazard regression (Cox) model, namely with the response function estimated as a histogram function and then secondary smoothed. We shall recall the methods of estimation of Cox model components and we shall show the consistency of estimates for our case. The practical application then will deal with the process of financial transactions and with the problem of detection of outlied (atypical) trajectories. The method uses the fact that the cumulated intensity actually represents the transformation of the process to the scale of Poisson process with intensity one, which, in our case, holds for both components of the considered two-dimensional process.

In certain cases it will be necessary to distinguish between the hazard rate (or hazard function) and the intensity. By the hazard rate of a continuous random variable we mean $h(x) = f(x)/(1 - F(x))$, where $f(x)$, $F(x)$ are the corresponding

density and distribution function. More generally, the hazard rate is a nonnegative function used in the model of random point process. The intensity then denotes the actual rate (local probability) of random event occurrence, it can depend on the (past) development of the process and on the covariates. Then, each trajectory of the process can have its own intensity (though they have the same model, the same hazard rate).

The methodology for Cox model analysis is collected for instance in Andersen et al. [1]), the treatment of the nonparametric version can be based for instance on the results on consistent spline models derived by C. J. Stone (e. g. Stone [13], Kooperberg et al. [8]). We shall use the histogram approximation which is actually a special case of spline model. There also exists a number of papers and monographs dealing with crossing probabilities of special types of compound processes and with connected problems, for instance Embrechts et al. [5], Asmussen [3], Rolski et al. [12]. In the next part we shall recall certain useful results from this field.

2. CROSSING PROBABILITIES FOR COMPOUND POISSON PROCESS

Let $C(t)$ represent a process of insurance claims, they should be covered from a fund $u + vt$ at time t. Then $R(t) = u + vt - C(t)$ is the risk process and the event $R(t) < 0$ means the ruin. Hence, the problem is how parameters u, v (initial capital and income rate) should be selected in order to keep the ruin probability $P(\inf_t R(t) < 0)$ less than a given α (either on $[0, T]$ or $[0, \infty)$). Notice that the problem of selection of u, v is actually equivalent to the construction of a linear prediction band, i. e. such a line that the process $C(t)$ lies below it with probability at least $1 - \alpha$. Though the basic results concerning the Pollaczek–Khinchin convolution formula for ruin probability, its Cramér–Lundberg approximation, etc., date back to 20-ties and 30-ties of the last century, the problem is solved explicitly only for the simplest cases. Namely, let us consider the compound Poisson process consisting of a homogeneous Poisson process of random time points with constant hazard rate h and of increments distributed identically and exponentially (with hazard rate g, i. e. $EY = 1/g$). Moreover, let all components be mutually independent. Notice also that the projection of $C(t)$ to the vertical axis is again a homogeneous Poisson process. For such a compound Poisson process and infinite time interval the exact formula for the crossing probability has been derived (see for instance Rolski et al. [12]):

$$P\left(\inf_t R(t) < 0\right) = \frac{h}{gv} \exp\left(-\frac{(v - h/g) \cdot u}{v/g}\right) = (1 - p)\, e^{-pug}, \tag{2}$$

where $p = (v - h/g)/v$ compares the growth rate of the reserve v with the mean of claims (per time unit) $h \cdot EY$, $p > 0$ is the basic condition for the existence of solution. Hence, for given α and selected $v > h/g$ the corresponding $u = -\ln(\alpha/(1-p))/(pg)$.

As regards another simple case, namely the upper prediction limit for the trajectories of the standard Poisson process (i. e. with hazard rate $h = 1$ and non-random increments equal to one), it can be obtained (computed or randomly generated) from the corresponding Pollaczek–Khinchin formula. Namely, the probability of crossing the line $u + v \cdot t$, when $u \geq 0, v > 1$, is given by a finite sum (e. g. Asmussen [3],

Part III. 3d):

$$P(u) = 1 - (1-\rho)\sum_{k=0}^{[u]} e^{-\rho(k-u)}\frac{(\rho(k-u))^k}{k!},$$

where $\rho = 1 - p$ from above. Other possibility is to use the Cramér–Lundberg approximation $P(u) \sim (v-1)/(rv+1-v)\cdot\exp(-ru)$, where r is the positive solution of $\exp(r) - 1 = r \cdot v$.

At present the research in the area of compound processes is focused mainly to the cases with sub-exponentially distributed increments (i. e. distributions with heavy tails) corresponding to many real situations. On the other hand, the models allowing for dependence (mutual or on common covariates and history) of increments and times are not so frequent, due many both theoretical and methodological difficulties. In Parts 4 and 5 we shall study one such a case and propose at least an approximate method of solution. Nevertheless, let us first consider just a slight generalization of the preceding case, the nonhomogeneous Poisson process of times with general, but still i.i.d. increments, in order to recall the way of estimation of hazard rates.

2.1. Nonhomogeneous compound Poisson process

Let the process of times $N(t)$ be nonhomogeneous Poisson process with hazard rate $h(t)$ and the distribution of increments $Y(t)$ be given by hazard rate $g(y)$. We still assume that both components are independent mutually; this is the most serious restriction of such a model. Increments are therefore i.i.d. random variables, however, we describe them as a set of point processes $S_j(y)$, say, with only (maximally, in the case of censoring) one point – the value of increment. Denote $H(t) = \int_0^t h(s)\,\mathrm{d}s$ and $G(y) = \int_0^y g(x)\,\mathrm{d}x$ cumulated hazard rates.

Let n realizations be observed in $[0,T]$, without censoring, so that the data are $N_i(t)$, $C_i(t)$, $i = 1,\ldots,n$, each with time points $0 < T_{i1} < \cdots < T_{im} < T$, $(m = m_i = N_i(T))$, and increments Y_{ij}, $j = 1,\ldots,m_i$. Let us first recall the likelihood process of $h(t)$ based on observed counting processes $N_i(t)$:

$$V_t = \prod_{i=1}^{n}\left\{\prod_{t>0} h(t)^{\mathrm{d}N_i(t)}\cdot\exp\left(-\int_0^T h(t)\,\mathrm{d}t\right)\right\},$$

and the corresponding Nelson–Aalen estimate of cumulated rate $H(t)$:

$$\widehat{H}(t) = \sum_{i=1}^{n}\int_0^t \frac{\mathrm{d}N_i(s)}{n} = \frac{1}{n}\sum_{i=1}^{n} N_i(t^-) = \frac{1}{n}\sum_{i=1}^{n}\sum_j 1[T_{ij} < t]. \tag{3}$$

Similarly, the likelihood of rate $g(y)$ of the increments distribution can be written as

$$V_y = \prod_{i=1}^{n}\prod_{j=1}^{m_i}\left\{\prod_{y>0} g(y)^{\mathrm{d}S_{ij}(y)}\cdot\exp\left(-\int_0^\infty g(y)J_{ij}(y)\,\mathrm{d}y\right)\right\},$$

where $J_{ij}(y)$ are random indicators, $J_{ij}(y) = 1$ for $0 \le y \le Y_{ij}$, $J_{ij}(y) = 0$ otherwise. Hence, at fixed y, $\overline{J}(y) = \sum_{i=1}^{n}\sum_j J_{ij}(y)$ is the number of increments larger than y

($\overline{J}(y)$ is called the risk set in the field of survival analysis). Again, the estimator of Nelson–Aalen type for $G(y)$ yields

$$\widehat{G}(y) = \sum_{i=1}^{n} \sum_{j} \int_0^y \frac{\mathrm{d}S_{ij}(x)}{\sum_{k=1}^{n} \sum_{l} J_{kl}(x)} = \sum_{i=1}^{n} \sum_{j} \frac{1[Y_{ij} < y]}{\sum_{k=1}^{n} \sum_{l} J_{kl}(Y_{ij})}.$$

These estimators are consistent (uniformly on each bounded interval $[0,T] \times [0,Y]$) and asymptotically normal (in the sense of convergence of properly normalized residual process to the Wiener process). Estimates of hazard rates $h(t)$ or $g(y)$ are as a rule computed with the aid of kernel smoothing of increments $\Delta\widehat{H}(T_{ij})$, $\Delta\widehat{G}(Y_{ij})$.

3. A GENERAL MODEL OF CUMULATIVE PROCESS

Let us now formulate a rather general description of process $C(t)$. We shall model its time component, $N(t)$, as a counting process, with (possibly random) intensity process $\lambda(t) = h(t, Z(t)) \cdot I(t)$, where $h(t,z)$ is a hazard rate, $Z(t)$ is a covariate, which may depend on time and be random, too, $I(t)$ is the indicator of observability ($I(t) = 0$ if $N(t)$ is censored or terminated). The corresponding filtration $\mathcal{F}(t)$ is a nondecreasing sequence of σ-algebras generated by $\{N(s), I(s), Z(s), s \leq t\}$. Then $N(t) = M(t) + L(t)$, where $M(t)$ is a martingale adapted to $\mathcal{F}(t)$ and $L(t) = \int_0^t \lambda(s)\,\mathrm{d}s$ is the cumulated intensity, $\lambda(t)$, $Z(t)$, $I(t)$ are adapted (measurable with respect) to $\mathcal{F}(t^-)$. If $i = 1, \ldots, n$ processes are considered, we shall take $\mathcal{F}(t)$ as constructed jointly over all of them. Naturally, covariate processes $Z_i(t)$, indicators $I_i(t)$, and therefore also intensities $\lambda_i(t)$, $L_i(t)$ of ith process differ from each other. The actual state of the process, $C_i(t^-)$, can play the role of one of covariates influencing the future intensity (at $s \geq t$). Martingale innovations $\mathrm{d}M_i(t)$ are mutually conditionally independent, given $\mathcal{F}(t^-)$. This is a common schema of counting process models considered for instance in Andersen et al. [1]).

Simultaneously, let us again describe the increment at time t as a result of counting process $S(y,t)$, with an intensity $\mu(y,t)$. The 'history' for $S(y,t)$ can be constructed as $\boldsymbol{\sigma}_t(y^-) = \mathcal{F}(t) \otimes \mathcal{B}[0,y)$, i.e. as generated by relevant functions measurable with respect to it and left continuous at y, with $\mathcal{B}$ denoting the Borel σ-algebra. Finally, the filtration is $\boldsymbol{\sigma}_t(y) = \sigma\{\boldsymbol{\sigma}_t(y^-) \cup \mathrm{d}\mathcal{M}_t(y)\}$, where $\mathrm{d}\mathcal{M}_t(y) = \mathrm{d}S(y,t) - \mu(y,t)\,\mathrm{d}y$ is a martingale increment, the innovation of $S(y,t)$ at y.

If n processes are observed, each with 0, 1 or more increments, we then describe the increment of ith process at time t, $Y_i(t)$, via the counting process $S_i(y,t)$, with an intensity $\mu_i(y,t)$ ($i = 1, \ldots, n$). The model of intensity is based on a hazard rate $g(y,z,t)$, the same for all increments, the actual intensity at y for given covariate $Z_i(t)$, indicator $J_i(y,t)$, at time t, is then $\mu_i(y,t) = g(y, Z_i(t), t) \cdot J_i(y,t)$. Notice that the time t is actually regarded as one of covariates influencing the magnitude of an increment.

The formulation of such a general form of the model can be helpful for clearing up the structure of possible mutual dependencies of cumulative process components and the sequential dynamics of their development, through the notions like 'history'

and innovation depending on it. We can also imagine the compound process as one point process possessing a mechanism switching its direction, at points $(T_j, C(T_j^-)$ from horizontal to vertical, and at $(T_j, C(T_j) = C(T_j^-) + Y(T_j))$ back to horizontal.

In the sequel, we shall return to a simpler situation without regression on covariates, however, with the dependence of increments on the time.

4. MODEL WITH TIME–DEPENDENT INCREMENTS

In the rest of the paper we shall consider the case without covariates but such that the increments depend on the time of their occurrence. Hence, the process is described by hazard rates $h(t)$ and $g(y;t)$. We again assume that n processes $N_i(t), C_i(t)$ are observed, in an interval $[0,T] \times [0,\infty]$, fully, without any censoring, so that the corresponding indicators are $I_i(t) = 1$ on $[0,T]$, $J_i(y,t) = 1$ for $y \in [0, Y_i(t)]$. On the other hand, we shall study the properties of estimates on a chosen bounded interval $[0,T] \times [0,Y]$ only. We also assume that functions h and g are bounded on that interval. The consequence of the boundedness of g is that $\overline{J}(y) = \int_0^T \sum_{i=1}^n J_i(y;t)\,\mathrm{d}N_i(t)$ is $O_P(n)$ for each $y \in [0,Y]$. In other words, $\overline{J}(y)/n$ has a P-limit, which is bounded and bounded away from zero when n increases to infinity (while $\overline{I}(t) = \sum_{i=1}^n I_i(t) = n \sim O_P(n)$ directly).

As regards the statistical analysis, the process of times is actually a nonhomogeneous Poisson one, the estimate of cumulated hazard rate $H(t) = \int_0^t h(s)\,\mathrm{d}s$ is then obtained from the Nelson–Aalen estimator (3). On the other end, the hazard rate of increments is a function of two variables, of y as a leading one and of t as a covariate. There exist quite general methods of estimation in such a case, for instance the method of nonparametric estimation of doubly-cumulative hazard rate proposed in McKeague and Utikal [10], with consistent results. Their estimate is actually of the Nelson–Aalen type w. r. to y and of the kernel type w. r. to t. In the sequel, we shall consider the Cox model specification of $g(y;t)$.

4.1. Proportional hazard model for increments

Let us assume that the hazard rate of the distribution of an increment at time t can be written as

$$g(y;t) = g_0(y) \cdot e^{b(t)}, \tag{4}$$

where $g_0(y)$ is a baseline hazard rate and $b(t)$ is a (nonparametric, in general) response function. It is seen that functions in (4) are not given uniquely, some normalization is necessary. For instance, we can keep $b(t_0) = 0$ at a chosen point t_0. The standard case deals with a parametrized function $b(t)$, the nonparametric maximum likelihood problem can be solved via the local scoring method (e. g. Hastie and Tibshirani [6], also Volf [15]), or, alternatively, $b(t)$ can be constructed from an appropriate functional basis (e. g. polynomials , splines etc.). The simplest approach uses a histogram-like estimator of $b(t)$, i. e. taking $\widehat{b}(t) = \widehat{b}_r$ constant in selected equidistant intervals $\mathcal{T}_r$, $r = 1, \ldots, m$, dividing $[0,T]$ (while it is assumed that the actual unknown $b(t)$ is a continuous function). Let us denote for $i = 1, \ldots, n$

$\overline{S}_i(y,r) = \int_{\mathcal{T}_r} S_i(y,t)\,\mathrm{d}N_i(t)$ the (marked) counting process (in argument y) registering all increments in $\mathcal{T}_r$, then denote $\overline{S}(y,r) = \sum_{i=1}^{n} \overline{S}_i(y,r)$, and express also the corresponding risk sets as $\overline{J}(y,r) = \sum_{i=1}^{n} \int_{\mathcal{T}_r} J_i(y,t)\,\mathrm{d}N_i(t)$. The constants $\widehat{b}_r$ are obtained from the maximization of the logarithm of Cox partial likelihood

$$L_p = \sum_{r=1}^{m} \int_0^Y \ln\left(\frac{e^{b_r}}{\sum_{s=1}^{m} e^{b_s}\,\overline{J}(y,s)}\right) \mathrm{d}\overline{S}(y,r). \tag{5}$$

Finally, the cumulated baseline hazard rate $G_0(y) = \int_0^y g_0(x)\,\mathrm{d}x$ is estimated with the aid of the Breslow–Crowley estimator as

$$\widehat{G}_0(y) = \int_0^y \sum_{r=1}^{y} \frac{\mathrm{d}\overline{S}(x,r)}{\sum_{s=1}^{m} e^{\widehat{b}_r}\,\overline{J}(x,s)}. \tag{6}$$

Let us note that the piecewise constant function $\widehat{b}(t)$ has here also the character of a heterogeneity variable describing the departures of the distribution of increments in certain time intervals from the baseline distribution given by $g_0(y)$.

Naturally, the assumption of proportional hazard should be verified, there exists a number of tests of proportionality of two subsamples as well as the goodness-of-fit tests. Commonly used is the graphical test assessing the fit of Cox model (see Arjas [2]), a more complicated numerical specification of this graphical method for the parametrized Cox model has been proposed in Marzec and Marzec [9].

The graphical test uses the generalized residuals, the differences between actual intensities computed from the tested model (here $\mu_i(y,t) = g_0(y)\exp(b(t))\,J_i(y,t)$) and observed number of counts (points of process). These residuals are summarized through a selected subsample of data and cumulated from 0 to kth from ordered counts. Then they are plotted in such a way that numbers $k = 1, 2, \ldots$, are on the abscissa and cumulated intensities on the ordinate. If the model holds, the difference is a martingale, therefore the curve of cumulated intensities should be close to the diagonal line of the graph. On the contrary, large differences then indicate the lack of fit. The selection of different subsamples allows to reveal where the departure from the model is significant.

4.2. Consistency of nonparametric estimate

As regards the asymptotic properties, i. e. the situation when $n \to \infty$, it is as a rule assumed that the number of histogram intervals $m = m_n \to \infty$ and $n/m_n \to \infty$. Under certain more-less technical conditions the consistency of estimation can be proven. The theory for the parametrized case is already well developed, the asymptotics for the more complex nonparametric case can use several sources. One of them is based on already mentioned results of McKeague and Utikal [10] which has been developed further in McKeague and Utikal [11], Section 6 for the multiplicative specification of the general model, namely $g(y,t) = g_0(y) \cdot c(t)$. Naturally, the standard case has been studied, so that the first variable was the time and the second a covariate, the notation therefore differed from that used here. McKeague and Utikal constructed their histogram-like estimate c_r of function $c(t)$ in the following manner:

They set $\sum_r c_r = 1$, consequently $\widehat{G}_0(y) = \widehat{\mathcal{G}}(y,T)$. Further, they selected suitably one value y_1 in the domain of variable y and they proposed $c_r = \widehat{G}(y_1,r)/\widehat{\mathcal{G}}(y_1,T)$, where $\widehat{G}(y,r)$ is the Nelson–Aalen estimate of C.H.R. in $\mathcal{T}_r$, $\widehat{\mathcal{G}}(y,t)$ is the estimate of doubly-cumulative hazard rate. It is seen that such a method utilizes just a part of the available information, nevertheless, the authors proved its consistency and asymptotic normality following from the asymptotics of the doubly-cumulative hazard rate estimator. The results were then used for the goodness-of-fit test of the multiplicative model.

A variant of such an estimator has been studied and its limit properties shown in Volf [14]. The estimator of each c_r was computed from the least squares method, minimizing the following expressions

$$\sum_{r=1}^{m} \int_0^Y (\widehat{G}(y,r) - \widehat{G}_0(y) \cdot c_r)^2 \mathrm{d}\bar{S}(y,r).$$

In the present case we are interested in the estimator of values b_r based on (5). The analysis of its properties can utilize the results derived in Kooperberg et al. [8], which are actually the application of C. J. Stone's idea of "dimensionality reduction principle" (D.R.P., Stone [13]) to the log-additive spline model of hazard rate. The substance of the D.R.P. is rather general and claims that under suitable conditions there exists a unique additive function closest to the actual response function, and that this additive function is consistently estimable. Stone prefers the maximum likelihood estimator constructed from polynomial splines of order M and shows the optimal rate of convergence depending both on M and on the smoothness of a target function.

The situation considered in the present paper is just a particular case. Thus, our solution maximizing (5) is equivalent to the estimate of function b obtained by the approach of Kooperberg et al. [8]. Notice also that for finite m the maximization of (5) has a unique solution – it is easy to show that the second derivatives of L_p yield a negative definite matrix. In accordance with Kooperberg et al., let us assume the following minimal requirements:

Assumptions A1.

1. Let the actual function $b(t)$ be Lipschitz continuous on the interval $[0,T]$ (it then follows that the first of "smoothness parameters" considered in Stone and in Kooperberg et al. can be taken as $\beta = 1$).
2. The following conditions of stability hold: There exist limits in probability

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} J_i(y,T_i), \qquad \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} e^{b(T_i)} J_i(y,T_i),$$

uniformly in $[0,Y]$, such that they are bounded and bounded away from zero. Actually, this condition, together with boundedness of $b(t)$, yields a variant formulation of Conditions 1 and 2 of Kooperberg et al..

3. The function $b(t)$ will be estimated from (5), i. e. with the aid of the ML principle and in a piece-wise constant form (the second smoothness parameter considered both in Stone and Kooperberg et al. is then $M = 0$). Then, the function $G_0(y)$ will be estimated along (6).
4. The number of (equidistant) intervals dividing $[0, T]$ will be chosen as $m_n = n^\gamma$, with $\gamma = 1/3$ (it corresponds to the optimal $\gamma = 1/(2p+1)$, with $p = \beta + M$, required in Stone and then also in Kooperberg et al., so that we obtain $p = 1$).

Proposition 1. Let assumptions A1 hold. Then the estimator $\widehat{b}(t)$) computed from (5) is consistent, with the rate of convergence in probability with respect to the L_2 norm

$$||\widehat{b}(t) - b(t)|| = O_p(n^{-1/3}).$$

Proposition 2. The estimator $\widehat{b}(t)$ is P-consistent also w. r. to the supremum norm, namely for $t \in [0, T]$

$$\sup_t |\widehat{b}(t) - b(t)| = o_p(1).$$

Corollary 1. The estimator $\widehat{G}_0(y)$ computed from (6) is P-consistent as well, uniformly on $[0, Y]$.

The statement of Proposition 1 follows from Theorem 3 of Kooperberg et al. [8], the Proposition 2 then from Lemmas 2 and 7 there. Corollary 1 is the consequence of the boundedness of $g_0(y)$, of uniform consistency of $\widehat{b}(t)$ and of uniform consistency of the Breslow–Crowley estimator (6) in the case of the known function $b(t)$. Let us also mention here Conditions 3 (existence of maximizers of the log-likelihood) and 4 ($\gamma < 0.5$) of Kooperberg et al., however, in the context studied here they hold automatically.

5. EXAMPLE

As an example, we analyzed the processes of credit cards payments at a gas station. One process corresponded to payments on one day, data are from $n = 90$ days, $t \in [0, 24]$ hours (data can be downloaded from *http://siprint.utia.cas.cz/public/income/volf/trans5541.txt*). Figure 1 shows a part of observed processes $N_i(t)$ and $C_i(t)$, together with estimated cumulated rates $\widehat{H}(t)$ and $\widehat{K}(t)$, actually the means from observed trajectories. Certain interesting trajectories are denoted by their numbers. Figure 2 then contains kernel smoothed rates $\widehat{h}(t)$ and $\widehat{k}(t)$, the third subplot then shows, just for comparison, estimated mean increment $\widehat{\mu}(t) = \widehat{k}(t)/\widehat{h}(t)$ (it could be obtained also directly from the analysis of increments, by their kernel smoothing in the time domain). The estimates indicate that the frequency of increments as well as their magnitude depend on time.

No other covariates were considered, we assumed that the model (4) described the case sufficiently, which was also tested in the final phase of analysis. Now, the next step consists in the estimation of model components.

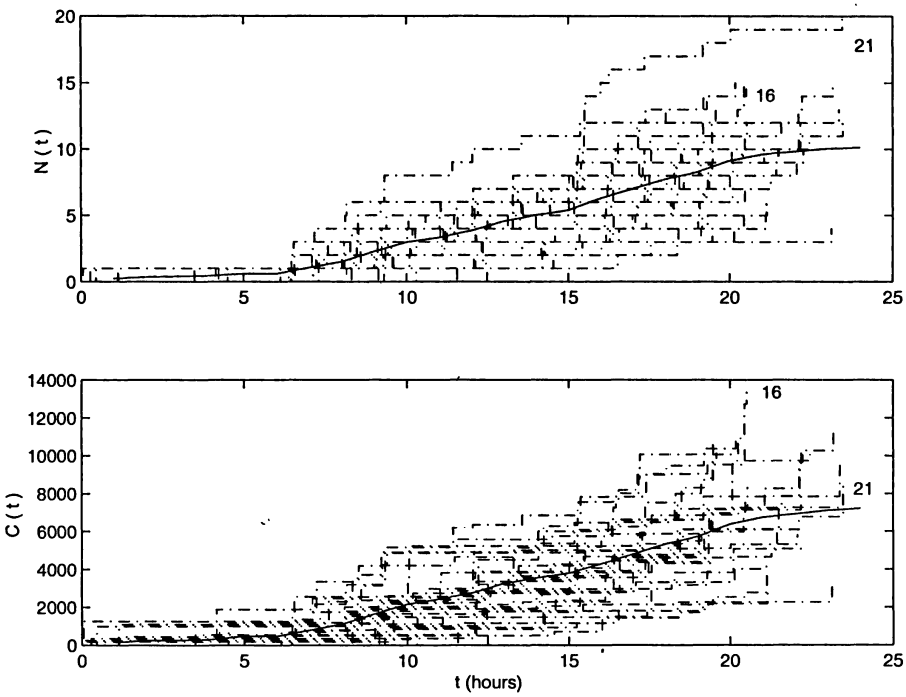

Fig. 1. Observed processes.

The time domain $[0, 24]$ hours was divided to $m = 24$ one-hour intervals. The piecewise constant estimator $\widehat{b}_r$ has been obtained from (5), then it was smoothed by a moving window. The result is on the lower subplot of Figure 3. We kept $b_1 = 0$ in order to guarantee uniqueness of solution. Further, the cumulated baseline rate estimate $\widehat{G}_0(y)$ has been computed along (6), its plot is in the upper subplot, the estimate of baseline distribution function $\widehat{F}_0(y) = 1 - \exp(-\widehat{G}_0(y))$ is displayed below it. Notice that greater $b(t)$ corresponds to smaller mean increments and vice versa.

For all trajectories their actual times and increments have been observed, and actual intensities have been just estimated. Hence, the behaviour of observed trajectories can be at once compared with their expected behaviour derived from the model. More precisely, if a process has times T_j and increments $Y_j = Y(T_j)$, and model with $H(t)$ and $G(y,t)$ is the right one, then times $\tau_j = H(T_j)$ should be the times of Poisson(1) process and values $Z_j = G(Y_j, T_j)$ should be the values corresponding to Exp(1) distribution. Therefore, the scale of each trajectory can be transformed to the scale of compound Poisson (1,1) process. We used the estimate $\widehat{H}(t)$ and, simultaneously, took $G(y,t) = \widehat{G}_0(y) \cdot \exp(\widehat{b}(t))$. The results of such transformations, for four interesting trajectories, are displayed in Figure 4, first subplot. The dashed line is the 95 % prediction line $u + v\tau$, $u = 9.308$, $v = 1.4$ (for selected v and $\alpha = 0.05$, u has been computed with the aid of (2) of Section 2).

Simultaneously, we transformed also the times of corresponding counting processes and compared them with one of 95% prediction lines. In subplot 2 of Figure 4, such a line is given by $u + v\tau$, $u = 8.17, v = 1.2$ (again, for selected v and $P(u) = 0.05$, u has been derived from the Cramér–Lundberg approximation given in Section 2). Finally, the third subplot shows the increments of selected four trajectories, transformed again to $Z(\tau) = \widehat{G}_0(y) \cdot \exp(\widehat{b}(t))$ at points $\tau = \widehat{H}(t)$, and compares them with the mean (dashed line) and 99% quantile (full line) of the standard exponential distribution.

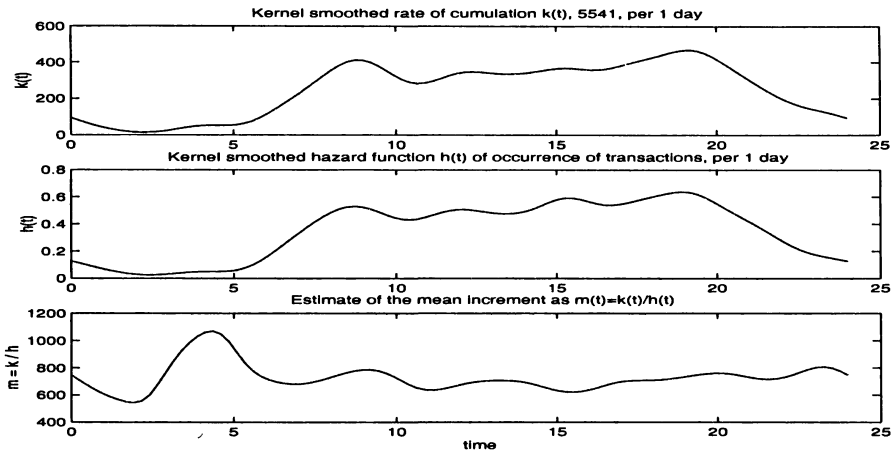

Fig. 2. Kernel estimates of rates and of mean increment as a function of time.

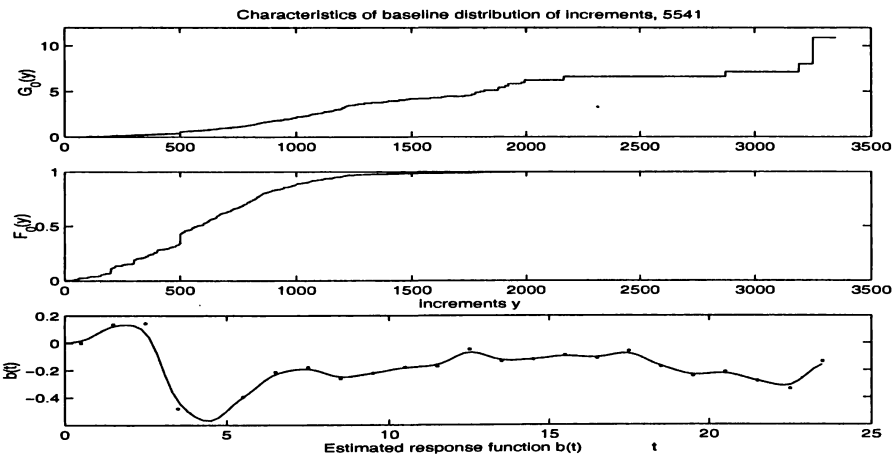

Fig. 3. Estimates of cumulated baseline rate $G_0(y)$, of baseline distribution function $F_0(y)$, and of response function $b(t)$.

Such a diagnostics revealed that, as expected, process No 21 had atypical number

of events – transactions, while their sum was close to average. Process No 16, on the contrary, had maximal $C(t)$, which was caused by rather high (but not atypical) both number and amounts of transactions (compare it also with Figure 1). The third interesting case was the process No 29, with large increments and a mild $N(t)$ resulting in relatively large (but not extremal) $C(t)$. Finally, process No 17 behaved quite standardly except one extremal increment. Naturally, the more convenient variant of such a diagnostics should use the cross-validation, i. e. the model would be evaluated only from trajectories not selected for the test.

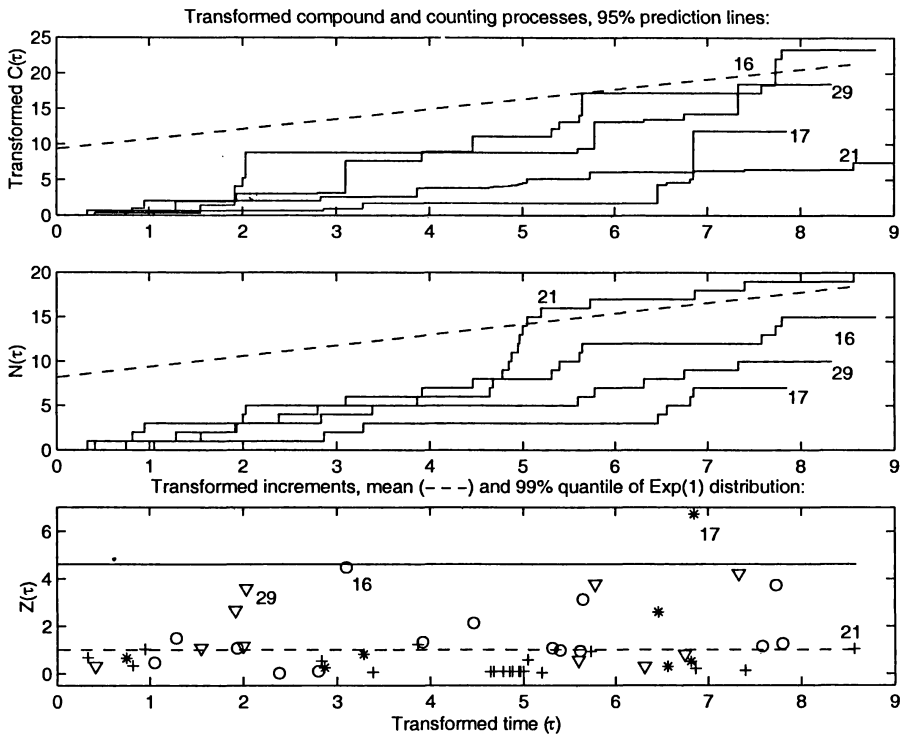

Fig. 4. Selected trajectories, comparison with prediction lines.

Finally, the graphical goodness-of-fit test described in Section 4.1 was utilized. Several subsamples of transactions in different time periods were analyzed, the test confirmed a good fit of the model to our data. Figure 5 shows a test graph corresponding to two subsamples – payments before and after 3 p.m. The graph indicates that the actual intensity of the distribution of increments after 3 p.m. was slightly larger than the intensity given by our model (which means that increments were actually smaller than the model assumed), and similarly the payments before 3 p.m. were (again slightly) larger than suggested by the model.

6. CONCLUSION

The main purpose of the paper was to offer a simple model for the cumulative processes consisting in the combination of the counting process with random increments dependent on it and to show an application to the analysis of the sequence of financial transactions. A successful use of such models requires the development of the methods for estimation of the model characteristics and also the methods for the prediction of the process behaviour under different conditions. Then we are also able to classify the processes and, eventually, to detect atypical ones. The practical application for instance to the fraud detection problem is quite straightforward.

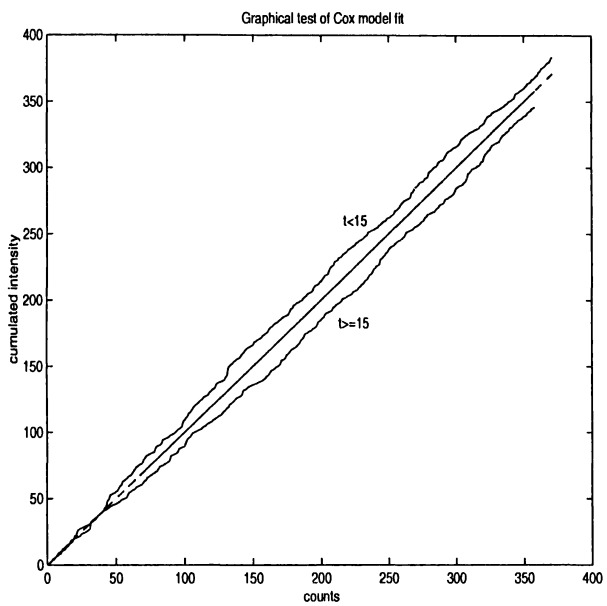

Fig. 5. Graphical assessing the goodness-of-fit of Cox model.

ACKNOWLEDGEMENT

The work was partially supported by the Grant Agency of the Czech Republic under Grant 402/04/1294.

(Received December 13, 2004.)

REFERENCES

[1] P. K. Andersen, O. Borgan, R. D. Gill, and M. Keiding: Statistical Models Based on Counting Processes. Springer, New York 1993.

[2] E. Arjas: A graphical method for assessing goodness of fit in Cox's proportional hazards model. J. Amer. Statist. Assoc. *83* (1988), 204–212.

[3] S. Asmussen: Ruin Probabilities. World Scientific, Singapore 2000.
[4] P. Brémaud: Point Processes and Queues: Martingale Dynamics. Springer, Berlin 1981.
[5] P. Embrechts, K. Klüppelberg, and T. Mikosch: Modeling Extremal Events. Springer, Berlin 1997.
[6] T. J. Hastie and R. J. Tibshirani: Generalized Additive Models. Wiley, New York 1990.
[7] J. Jacod and A. N. Shirjajev: Limit Theorems for Stochastic Processes. Springer, Berlin 2003.
[8] C. Kooperberg, C. J. Stone, and Y. K. Truong: The L_2 rate of convergence for hazard regression. Scand. J. Statist. *22* (1995), 143–157.
[9] L. Marzec and P. Marzec: Generalized martingale-residual processes for goodness-of-fit inference in Cox's type regression model. Ann. Statist. *25* (1997), 683–714.
[10] I. W. McKeague and K. J. Utikal: Inference for a nonlinear counting regression model. Ann. Statist. *18* (1990), 1172–1187.
[11] I. W. McKeague and K. J. Utikal: Goodness-of-fit tests for additive hazard and proportional hazard models. Scand. J. Statist. *18* (1991), 177–195.
[12] T. Rolski, H. Schmidli, V. Schmidt, and J. Teugels: Stochastic Processes for Insurance and Finance. Wiley, New York 1999.
[13] C. J. Stone: The use of polynomial splines and their tensor products in multivariate function estimation. With discussion. Ann. Statist. *22* (1994), 118–184.
[14] P. Volf: A nonparametric analysis of proportional hazard regression model. Problems Control Inform. Theory *18* (1989), 311–322.
[15] P. Volf: Analysis of generalized residuals in hazard regression models. Kybernetika *32* (1993), 501–510.
[16] P. Volf: On cumulative process model and its statistical analysis. Kybernetika *36* (2000), 165–176.

Petr Volf, Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, 182 08 Praha 8. Czech Republic.
e-mail: volf@utia.cas.cz