Petr Volf

On cumulative process model and its statistical analysis

Kybernetika, Vol. 36 (2000), No. 2, [165]--176

Persistent URL: http://dml.cz/dmlcz/135342

KYBERNETIKA — VOLUME 36 (2000), NUMBER 2, PAGES 165 - 176

ON CUMULATIVE PROCESS MODEL AND ITS STATISTICAL ANALYSIS

PETR VOLF

The notion of the counting process is recalled and the idea of the 'cumulative' process is presented. While the counting process describes the sequence of events, by the cumulative process we understand a stochastic process which cumulates random increments at random moments. It is described by an intensity of the random (counting) process of these moments and by a distribution of increments. We derive the martingale – compensator decomposition of the process and then we study the estimator of the cumulative rate of the process. We prove the uniform consistency of the estimator and the asymptotic normality of the process of residuals. On this basis, the goodness-of-fit test and the test of homogeneity are proposed. We also give an example of application to analysis of financial transactions.

1. INTRODUCTION

A counting process is a stochastic point process registering random events and counting their number. The trajectory of such a process starts at zero and has jumps $+1$ at random moments. The main characteristic is the intensity of the stream of events. A review of theory and applications of counting process models is given, for instance, in Andersen et al [2], or in Fleming and Harrington [5].

In the present paper, we consider a random process

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

where $N(t)$ is a counting process and $Y(t)$ is a set of random variables. We assume that the time runs through $[0, T]$ and starting value is again $C(0) = 0$. From this point of view, we deal with a process having the random increments at disjoint random moments. The process $C(t)$ will be called the cumulative process. In the special case when the process of times, $N(t)$, is the Poisson one, the process (1) is known as the compound Poisson process (see for example Embrechts et al [4]).

The objective of the present paper is to describe the process (1) with the aid of characteristics of both its components, i. e. the hazard function of $N(t)$ and the distribution of $Y(t)$. The paper is organized as follows: In part 2 the process (1) is defined more accurately. Its martingale-compensator decomposition is presented

and the variance process of the martingale is computed. Then the estimator of the mean trajectory of the process (actually representing the cumulative rate) is constructed and its uniform consistency is proved. In this we generalize the results of Volf [7] achieved for the case of underlying Poisson process of events. The main result consists in the derivation of the weak convergence of the residual process to a Wiener one. Finally, based on this convergence, a test procedure is proposed both for assessing the goodness-of-fit of the model and for testing the homogeneity of two processes. In this we follow the method of analysis of generalized residuals for counting processes proposed in Arjas [3] and also in Volf [6].

2. THE MODEL OF CUMULATIVE PROCESS

In order to define the process (1), we consider a measurable, nonnegative and bounded function $h(t), t \geq 0$, the hazard function, and the indicator process $I(t)$ which equals 1 if $N(t)$ is in the risk of count, $I(t) = 0$ otherwise. Actually, $I(t)$ is an indicator of observability of counting process $N(t)$. Then, the behaviour of the counting process $N(t)$ in (1) is governed by a random (in general) intensity process $\lambda(t) = h(t)I(t)$.

Further, let us consider a right–continuous nondecreasing sequence of σ-algebras, $\mathcal{S}(t)$, where each $\mathcal{S}(t)$ is defined on the sample space of $\{N(s), I(s), Y(s),\ 0 \leq s \leq t\}$. We assume that process $N(t)$ is $\mathcal{S}(t)$-measurable. Following Andersen and Borgan [1], we denote by $\mathrm{d}N(t)$ the increments of $N(t)$ over the small time interval $[t, t+\mathrm{d}t)$. Then we can write that $\lambda(t)\,\mathrm{d}t = P(\mathrm{d}N(t) = 1 \mid \mathcal{S}(t^-))$. The trajectories of $N(t)$ are right–continuous, the trajectories of $I(t)$ and also the "histories" collected in $\mathcal{S}(t^-)$ are left–continuous.

As regards the variables $Y(t)$, we assume that

1. $Y(t)$ are distributed with (unknown) densities $f(y;t)$.
2. Their means $\mu(t)$, variances $\sigma^2(t)$, and also $E(|Y(t)|^3)$ exist and are measurable and bounded functions on $[0, T]$.
3. Each $Y(t)$ is independent of $\mathcal{S}(t^-)$, i.e. of the history of the process $C(s)$ up to t (on the other hand, $\mathrm{d}N(t)$ can depend on history of $Y(s), s < t$).

Remark. Point 3 is a rather strong condition which in some cases is not fulfilled. On the other hand, we can imagine a number of processes (especially in the area of natural sciences) for which such an independence of increments on the history is a quite realistic property.

The assumption on existence and boundedness of the 3-rd absolute moments is actually a condition of the Lyapunov version of the central limit theorem. Here, it will be utilized for the proof of the condition of Lindeberg required in the central limit theorem for martingales (Proposition 3.).

Let us now recall the martingale-compensator decomposition of the counting process $N(t)$. Define first its cumulative intensity $L(t) = \int_0^t \lambda(s)\,\mathrm{d}s$. Then $N(t) = L(t) + M(t)$, where $M(t)$ is a martingale adapted to the σ-algebras $\mathcal{S}(t)$ (i.e. it

is $\mathcal{S}(t)$-measurable). It holds that $E\,M(t) = 0$ and variance processes $\langle N\rangle(t) = \langle M\rangle(t) = L(t)$, where the notation $\langle N\rangle(t)$ means $var(N(t)|\mathcal{S}(t^-))$. Similarly as the paths of $N(t)$, the paths of $M(t)$, and also of $C(t)$, are right–continuous (while the paths of $L(t)$ are continuous).

The first task is to derive a compensator of the process $C(t)$. We utilize the decomposition $N(t) = L(t) + M(t)$ of the counting process. Denote $Y^*(t) = Y(t) - \mu(t)$. Then we can decompose

$$C(t) = \int_0^t (Y^*(s) + \mu(s))\,\mathrm{d}N(s) = \int_0^t \mu(s)\,\mathrm{d}L(s) + \mathcal{M}(t),$$

where

$$\mathcal{M}(t) = \mathcal{M}_1(t) + \mathcal{M}_2(t) = \int_0^t Y^*(s)\,\mathrm{d}N(s) + \int_0^t \mu(s)\,\mathrm{d}M(s).$$

Proposition 1. The processes $\mathcal{M}(t)$, $\mathcal{M}_1(t)$, $\mathcal{M}_2(t)$ are martingales adapted to σ-algebras $\mathcal{S}(t)$ on $[0,T]$, the variance process of $\mathcal{M}(t)$ is

$$\langle\mathcal{M}\rangle(t) = \int_0^t (\sigma^2(s) + \mu^2(s))\,\mathrm{d}L(s).$$

Proof. Evidently, $E\,\mathcal{M}(t) = 0$. As regards the property defining the martingale, we have for $0 < s < t$ that

$$E(\mathcal{M}(t)\,|\,\mathcal{S}(s)) = \mathcal{M}(s) + E\left(\int_s^t \mathrm{d}\mathcal{M}(\tau)|\mathcal{S}(s)\right) = \mathcal{M}(s),$$

because $E\left(\int_s^t \mathrm{d}\mathcal{M}(\tau)|\mathcal{S}(s)\right) = 0$ holds for both parts of $\mathcal{M}(t)$: For $\mathcal{M}_1(t)$ it follows from the centering of $Y^*(t)$ and from the independence of $Y^*(t)$ on $\mathrm{d}N(t)$. Properties of $\mathcal{M}_2(t)$ follow directly from properties of $M(t)$. From the independence of $Y^*(t)$ on the past up to t it also follows that $Y^*(t)$ is orthogonal both to $\mathrm{d}M(t)$ and to $\mathrm{d}N(t)$, distribution of $\mathrm{d}N$ and $\mathrm{d}M$ being given by predictable process $\mathrm{d}L$. Therefore

$$\langle\mathcal{M}_1\rangle(t) = \int_0^t \sigma^2(s)\,\mathrm{d}L(s), \quad \langle\mathcal{M}_1,\mathcal{M}_2\rangle(t) = 0.$$

Further, from martingale properties of $M(t)$ we have that $\langle\mathcal{M}_2\rangle(t) = \int_0^t \mu^2(s)\,\mathrm{d}L(s)$. Then

$$\begin{aligned}\langle\mathcal{M}\rangle(t) &= E\,\{\mathcal{M}_1(t)^2 + \mathcal{M}_2(t)^2 + 2\mathcal{M}_1(t)\,\mathcal{M}_2(t)\,|\,\mathcal{S}(t^-)\}\\ &= \int_0^t \sigma^2(s)\,\mathrm{d}L(s) + \int_0^t \mu^2(s)\,\mathrm{d}L(s).\end{aligned}$$

□

Corollary. Process $\int_0^t \mu(s)\,\mathrm{d}L(s)$ is the compensator of process $C(t)$ on $[0,T]$.

Evidently, process $\int_0^t \mu(s)\,\mathrm{d}L(s)$ fulfils the conditions of compensator. Its substraction from $C(t)$ yields a martingale, process is predictable and its paths are uniformly continuous on $[0,T]$ (which is more than is needed).

3. LARGE SAMPLE PROPERTIES

In the follow-up, let us imagine that n realizations $C_i(t)$ of a cumulative process $C(t)$ are observed in interval $[0,T]$. More precisely, we observe moments of events T_{ij} of counting processes $N_i(t)$, corresponding indicators $I_i(t)$ and "jumps" $Y_i(T_{ij})$, $(i = 1,\ldots,n,\ j = 1,\ldots,n_i = N_i(T))$. Formally, observed trajectories are

$$C_i(t) = \int_0^t Y_i(s)\,\mathrm{d}N_i(s) = \sum_{j=1}^{n_i} Y_i(T_{ij}).$$

It is assumed that random variables $Y_i(t), i = 1,2,\ldots,n$ have the same distributions, with densities $f(y;t)$ and moments $\mu(t)$ and $\sigma^2(t)$. Further, we assume that $Y_i(t)$ are independent of the common history of the processes $N_i(s), Y_i(s), I_i(s), s < t, i = 1,2,\ldots,n$ stored now in σ-algebras $\mathcal{S}(t^-)$. Finally, we assume that $N_i(t)$ are characterized by the same hazard function $h(t)$. Corresponding intensities of $N_i(t)$ are then $\lambda_i(t) = h(t)I_i(t)$. As it is assumed that the hazard function is finite, the compensator is a continuous process. The consequence is also that there are not two events at one moment and, further, that for $i \neq j$ $\mathrm{d}\langle M_i, M_j\rangle(t) = 0$, $\mathrm{d}\langle \mathcal{M}_i, \mathcal{M}_j\rangle(t) = 0$ and even $\mathrm{cov}\{Y_i(t)\,\mathrm{d}\mathcal{M}_i(t), Y_j(t)\,\mathrm{d}\mathcal{M}_j(t)|\mathcal{S}(t^-)\} = 0$.

The likelihood process (which is actually the generalization of the likelihood of Poisson process) is

$$\mathcal{L} = \prod_{i=1}^{n}\prod_{j=1}^{n_i}\{\lambda_i(T_{ij})f(Y_i(T_{ij});T_{ij})\}\cdot\exp\left\{-\int_0^T \lambda_i(t)\,\mathrm{d}t\right\}.$$

It is seen that the part containing the intensities and the part containing the distribution of Y's can be separated (and therefore both characteristics can be estimated independently). Denote by $\mathcal{L}(f) = \prod_{i=1}^{n}\prod_{j=1}^{n_i} f(Y_i(T_{ij});T_{ij})$. Then we obtain the following log-likelihood:

$$\ln\mathcal{L} = \sum_{i=1}^{n}\left\{\int_0^T \ln\lambda_i(t)\,\mathrm{d}N_i(t) - \int_0^T \lambda_i(t)\,\mathrm{d}t\right\} + \ln(\mathcal{L}(f)). \tag{2}$$

In the case of parametrized function f, its parameters can be estimated from the maximum likelihood estimation procedure based on $\mathcal{L}(f)$ only. In a nonparametrized case, estimates of functions $\mu(t)$, $\sigma^2(t)$ can be obtained with the help of a smoothing (kernel) technique. Even the density $f(y;t)$ can be then estimated via the kernel method. In the follow-up, we shall try to characterize the process $C(t)$ jointly and to derive some asymptotic properties which depend on $h(t)$, $\mu(t)$ and $\sigma^2(t)$.

3.1. Estimates and their convergence

Let us recall here the well-known Nelson–Aalen estimator of the cumulative hazard function $H(t) = \int_0^t h(s)\,\mathrm{d}s$:

$$\widehat{H}_n(t) = \sum_{i=1}^{n}\int_0^t \frac{1[I(s) > 0]}{I(s)}\,\mathrm{d}N_i(s),$$

where $I(s) = \sum_{i=1}^{n} I_i(s)$ characterizes the risk set at moment s. Let us make the following assumption:

A1. There exists the limit $r(s) = \lim_{n\to\infty} \frac{I(s)}{n}$ in probability such that

a) the limit is uniform on $[0, T]$,

b) $1 \geq r(s) \geq \varepsilon$ on $[0, T]$, for some $\varepsilon > 0$.

Then it is proved elsewhere (for instance in Andersen and Borgan [1], Andersen et al [2]) that $\widehat{H}_n(t)$ is a consistent estimate of $H(t)$. Further, such an estimate is asymptotically normal on $[0, T]$, namely $\sqrt{n}(\widehat{H}_n(t) - H(t)) \approx \sqrt{n} \sum_{i=1}^{n} \int_0^t \frac{\mathrm{d}M_i(s)}{I(s)}$ converges weakly to a Wiener process with variance function $\int_0^t \frac{\mathrm{d}H(s)}{r(s)}$, when $n \to \infty$. It follows from the central limit theorems for martingales (e. g. Andersen et al [2], chapter II). Moreover, it is due A1 and due the boundedness of all involved functions that $\widehat{H}_n(t)$ is a uniformly consistent estimator of $H(t)$ on $[0, T]$ (see also Winter, Földes and Rejtö [8], and their variant of Glivenko–Cantelli theorem).

Inspired by these results, we consider the average of observed processes

$$\overline{C}_n(t) = \sum_{i=1}^{n} \int_0^t \frac{1[I(s) > 0]}{I(s)} Y_i(s)\, \mathrm{d}N_i(s)$$

as an estimator of the function $K(t) = \int_0^t \mu(s)\, \mathrm{d}H(s)$. Actually, $K(t)$ represents the cumulative rate describing the risk and the mean size of jumps of $C(t)$.

From A1 and boundedness of $H(s)$ and $\mu(s)$ we easily see that $P\left\{ \int_0^T 1[I(s) = 0]\, \mathrm{d}s = 0 \right\} \to 1$ for $n \to \infty$, so that even $P\left\{ \sqrt{n} \int_0^t 1[I(s) = 0]\, \mathrm{d}K(s) = 0 \right\} \to 1$, uniformly w.r.t. $t \in [0, T]$.

Proposition 2. Under A1, $\overline{C}_n(t)$ is a uniformly consistent estimate of $K(t)$ on $[0, T]$, that is

$$\lim_{n\to\infty} \sup_{t\in[0,T]} |\overline{C}_n(t) - K(t)| = 0 \text{ in probability.}$$

Proof.

$$\overline{C}_n(t) = \sum_{i=1}^{n} \int_0^t \frac{1}{I(s)} \{\mu(s)\, \mathrm{d}L_i(s) + \mathrm{d}\mathcal{M}_i(s)\}$$

$$= K(t) + \int_0^t \sum_{i=1}^{n} \frac{\mathrm{d}\mathcal{M}_i(s)}{I(s)} - \int_0^t 1[I(s) = 0]\, \mathrm{d}K(s). \tag{3}$$

Here $\mathcal{M}_i(t)$ are square integrable martingales, with the same distribution as $\mathcal{M}(t)$ defined in the preceding part. They are mutually orthogonal, $\langle \mathcal{M}_i, \mathcal{M}_j \rangle(t) = 0$ for $i \neq j$, their variances are uniformly bounded on $[0, T]$. From Lenglart's inequality (cf. Andersen and Borgan [1], or Andersen et al [2], part II.5.2) it follows that for each $\delta, \varepsilon > 0$ and for sufficiently large $n > n(\delta, \varepsilon)$

$$P\left(\sup_{t\in[0,T]} |\frac{1}{n} \sum_{i=1}^{n} \mathcal{M}_i(t)| \geq \varepsilon \right) \leq \frac{\delta}{\varepsilon^2}.$$

This, together with the uniform convergence of $\frac{I(s)}{n}$ assumed in A1, leads to the convergence of $\sup_{t\in[0,T]} |\int_0^t \sum_{i=1}^n \frac{\mathrm{d}\mathcal{M}_i(s)}{I(s)}|$ in probability to zero. □

Proposition 3. Under A1 the process $\sqrt{n}(\overline{C}_n(t) - K(t))$ converges weakly on $[0,T]$ to a Wiener process (i.e. the continuous Gaussian process with zero mean and independent increments) which has the variance function $W(t) = \int_0^t \frac{1}{r(s)}(\mu^2(s) + \sigma^2(s))\,\mathrm{d}H(s)$.

Proof. From (3) we have that

$$\sqrt{n}(\overline{C}_n(t) - K(t)) = \sqrt{n}\int_0^t \sum_{i=1}^n \frac{\mathrm{d}\mathcal{M}_i(s)}{I(s)} - \sqrt{n}\int_0^t 1[I(s)=0]\,\mathrm{d}K(s).$$

The convergence follows from the central limit theorem for martingales. We use the version stated in Andersen et al [2], namely Theorem II.5.1 (Rebolledo's). The proof requires two convergences in probability to hold, namely that for all $t \in [0,T]$

(i) $\langle M^{(n)}\rangle(t) \to W(t)$,

(ii) $\langle M_\varepsilon^{(n)}\rangle(t) \to 0$ (Lindeberg condition).

(i) Here $M^{(n)}(t) = \sqrt{n}\int_0^t \sum_{i=1}^n \frac{\mathrm{d}\mathcal{M}_i(s)}{I(s)}$. From Proposition 1 and from A1 it follows that

$$\langle M^{(n)}\rangle(t) = \int_0^t \sum_{i=1}^n \frac{(\sigma^2(s)+\mu^2(s))\,\mathrm{d}L_i(s)}{I^2(s)/n} =$$

$$= \int_0^t \frac{(\sigma^2(s)+\mu^2(s))\,1[I(s)>0]}{I(s)/n}\,\mathrm{d}H(s) \to W(t)$$

in probability, (i) is proved.

(ii) By $M_\varepsilon^{(n)}(t) = \int_0^t \sum_{i=1}^n \frac{\sqrt{n}}{I(s)} Q_{i,\varepsilon}(s)\mathrm{d}\mathcal{M}_i(s)$, with $Q_{i,\varepsilon}(s) = 1[|\sqrt{n}\mathrm{d}\mathcal{M}_i(s)/I(s)| > \varepsilon]$, we mean the process (martingale) containing all jumps of $M^{(n)}(t)$ larger than chosen $\varepsilon > 0$. For "zero-one" random variables $Q_{i,\varepsilon}(s)$ we have from the Chebyshev inequality that, for each $k = 1, 2, \ldots$

$$E(Q_{i,\varepsilon}^k(s)|\mathcal{S}(s^-)) = P(Q_{i,\varepsilon}(s) = 1|\mathcal{S}(s^-)) =$$

$$= P\left\{|\frac{\sqrt{n}\,\mathrm{d}\mathcal{M}_i(s)}{I(s)}| > \varepsilon|\mathcal{S}(s^-)\right\} \le \left\{\frac{n}{\varepsilon^2}\frac{(\mu^2(s)+\sigma^2(s))}{I^2(s)}\right\}\mathrm{d}L_i(s), \qquad (4)$$

where the expression in brackets is of size $\mathcal{O}_P\left(\frac{1}{n}\right)$ uniformly in $s \in [0,T]$, as a consequence of A1 (by the notation $B_n \sim \mathcal{O}_P(a_n)$, for sequences B_n of random variables and a_n of numbers, we mean that the sequence B_n/a_n is asymptotically bounded in probability).

From the conditional orthogonality of increments $\mathrm{d}\langle \mathcal{M}_i, \mathcal{M}_j\rangle(s)$ for $i \neq j$, we obtain that

$$\langle M_\varepsilon^{(n)}\rangle(t) \leq \int_0^t \frac{n}{I^2(s)} \sum_{i=1}^n E\{Q_{i,\varepsilon}^2(s)(\mathrm{d}\mathcal{M}_i(s))^2 | \mathcal{S}(s^-)\}.$$

Further, the Hölder inequality yields that

$$E\{Q_{i,\varepsilon}^2(s)(\mathrm{d}\mathcal{M}_i(s))^2 | \mathcal{S}(s^-)\} \leq [E\{|\mathrm{d}\mathcal{M}_i(s)|^3 | S(s^-)\}]^{\frac{2}{3}} \cdot [E\{Q_{i,\varepsilon}^6(s) | S(s^-)\}]^{\frac{1}{3}}.$$

From our assumptions on the boundedness of (absolute) moments of variables $Y(s)$ up to the 3-rd moment it follows that $E\{|\mathrm{d}\mathcal{M}_i(s)|^3 | S(s^-)\} = E\{|Y_i(s)|^3\}\mathrm{d}L_i(s) + \mathcal{O}((\mathrm{d}s)^2)$, taking into account that $\mathrm{d}L_i(s) = h(s)I_i(s)\mathrm{d}s$. Finally, we obtain

$$\langle \mathcal{M}_\varepsilon^{(n)}\rangle(t) \leq \int_0^t \frac{n}{I^2(s)} \sum_{i=1}^n \{E|Y_i(s)|^3 \mathrm{d}L_i(s) + \mathcal{O}((\mathrm{d}s)^2)\}^{\frac{2}{3}} \{\mathcal{O}_P(\frac{1}{n})\mathrm{d}L_i(s)\}^{\frac{1}{3}}$$

$$\sim \int_0^t \frac{n}{I(s)} \{E|Y_1(s)|^3 h(s)\,\mathrm{d}s\}^{\frac{2}{3}} \mathcal{O}_P(n^{-\frac{1}{3}})(h(s)\,\mathrm{d}s)^{\frac{1}{3}} \sim \int_0^t \frac{n}{I(s)} \mathcal{O}_P(n^{-\frac{1}{3}}) h(s)\,\mathrm{d}s,$$

which is of size $\mathcal{O}_P(n^{-\frac{1}{3}})$ uniformly in $t \in [0, T]$. This proves the condition (ii). □

4. STATISTICAL TESTS

In the following part, the asymptotic normality of the residual process will be utilized for the construction of statistical tests, namely the goodness-of-fit test and the test of homogeneity of two samples of cumulative processes.

From two parts of martingale $\mathcal{M}(t)$ the first reflects the variability of Y's and the second equals $\mu(t)$-times the 'residual' martingale known from the counting processes scheme. Hence, the variance function of residuals $\sqrt{n}(\overline{C}(t) - K(t))$ contains also two parts, expressed by σ^2 and μ^2. They can significantly influence the power of test procedures. Therefore, for the purpose of tests, we recommend to normalize the residuals, i. e. to divide them by $\sqrt{\sigma^2 + \mu^2}$.

4.1. The goodness-of-fit test

Arjas [3] and later Volf [6] derived goodness-of-fit tests for the counting processes model, and generalized them for the case of hazard regression models (namely Arjas considered the Cox model, Volf a general case of hazard regression model and the Aalen model). From this point of view, the case considered here is much simpler, because the regression is not involved.

Let the model be given by functions $H(t)$, $\mu(t)$, $\sigma^2(t)$, we want to decide whether the data correspond to it. The data are represented by the observed trajectories $C_i(t)$ and indicators $I_i(t)$, $i = 1, \ldots, n$. The tests are quite naturally based on the comparison of $\overline{C}_n(t)$ with expected $K(t)$. The process of differences $\overline{C}_n(t) - K(t)$ is called the residual process.

Graphical test: Let us order all moments of events into one nondecreasing sequence T_k, $k = 1, \ldots, K$. For a graphical comparison, we plot $K(T_k) = \int_0^{T_k} \mu(s)\,\mathrm{d}H(s)$ and $\overline{C}(T_k)$ into one figure, against k on the abscissa. If the model fits the residual process is a martingale asymptotically tending to zero. Then it is expected that both plots will be close to each other. An opposite case (i.e. an increasing distance of both curves) indicates that the model $K(t)$ does not correspond to the data. Of course, a more precise test will need a specification of critical limits for the distance of compared curves. Such critical bounds can be derived from the large sample properties, for instance in the following way.

Numerical test: Numerical test is based on asymptotic distribution. Define the normalized residual process by

$$R_n(t) = \int_0^t \frac{\mathrm{d}(\overline{C}_n(s) - K(s))}{\sqrt{\mu^2(s) + \sigma^2(s)}}.$$

From Proposition 3 it follows that $\sqrt{n}R_n(t)$ is asymptotically distributed as a Wiener process with the variance function $V(t) = \int_0^t \mathrm{d}H(s)/r(s)$. Then the process

$$D_n(t) = \sqrt{n}R_n(t)\,/(1 + V(t))$$

is (if the model holds) asymptotically distributed as a Brownian bridge process $\mathcal{B}((V(t)/(1 + V(t))$, in $t \in [0, T]$. Hence, a test of Kolmogorov–Smirnov type can be used. From the theory of Brownian bridge it follows, for instance, that for any $d \geq 0$,

$$P\left(\max_t D_n(t) \geq d\right) = P\left(\min_t D_n(t) \leq -d\right) \approx \exp(-2d^2)$$

approximately. So that the value $\exp(-2d^2)$, where d is the observed $\max_k |D_n(T_k)|$, is an approximate P-value for the test of hypothesis of the goodness-of-fit against a proper one-sided alternative. Unknown limit function $r(s)$ needed for computation of $V(t)$ is consistently estimated by $I(s)/n$ from A1.

4.2. Test of homogeneity

Besides the goodness-of-fit tests, we can also consider the tests of homogeneity. They compare two (sets of) realizations of the process. Both graphical and numerical comparison can be based on slight modifications of the methods described above. On the other hand, the performance of a test of homogeneity is influenced by the fact that, as a rule, certain characteristics of the joint model have to be estimated. The properties of the test procedure then depend strongly on the properties of the estimator.

Let us consider two independent sets of cumulative processes, $C_i^{(k)}(t)$, $k = 1, 2$, $i = 1, 2, \ldots, m_k$, each representing a certain model characterized by $H^{(k)}(t), \mu^{(k)}(t)$, $\sigma^{(k)}(t)$. The test of homogeneity assesses the hypothesis $\boldsymbol{H}_0$ that $H^{(k)}(t)$, $\mu^{(k)}(t)$, $\sigma^{(k)}(t)$ are the same for $k = 1, 2$, on a given interval $[0, T]$. To confirm it, we analyse

the averaged processes

$$\overline{C}^{(k)}(t) = \int_0^t \sum_{i=1}^{m_k} \frac{Y_i^{(k)}(s)\,\mathrm{d}N_i^{(k)}(s)}{I^{(k)}(s)}\; 1[I^{(k)}(s) > 0],$$

and their difference

$$\sqrt{\frac{m_1 m_2}{m}}\left\{\overline{C}^{(1)}(t) - \overline{C}^{(2)}(t)\right\}$$

$$\approx \sqrt{\frac{m_1 m_2}{m}}\left\{K^{(1)}(t) - K^{(2)}(t) + \int_0^t \sum_{i=1}^{m_1} \frac{\mathrm{d}\mathcal{M}_i^{(1)}(s)}{I^{(1)}(s)} - \int_0^t \sum_{i=1}^{m_2} \frac{\mathrm{d}\mathcal{M}_i^{(2)}(s)}{I^{(2)}(s)}\right\},$$

where $m = m_1 + m_2$. Let us now assume that:

1. $\boldsymbol{H}_0$ holds, so that $K^{(1)}(t) = K^{(2)}(t)$, μ, σ and H are common for both processes.
2. m_1, m_2 tend to infinity in such a way that $\frac{m_1}{m} \to \alpha \in (0,1)$.
3. Assumption A1 is fulfilled for both sets of processes (possibly with different $r^{(k)}(s)$).

Then, from Proposition 3 it follows that

$$\sqrt{\frac{m_1 m_2}{m}} \int_0^t \frac{\mathrm{d}(\overline{C}^{(1)}(s) - \overline{C}^{(2)}(s))}{\sqrt{\mu^2(s) + \sigma^2(s)}} \tag{5}$$

tends weakly to the Wiener process with zero mean and variance function $V^*(t) = (1-\alpha)V^{(1)}(t) + \alpha V^{(2)}(t)$, where $V^{(k)}(t) = \int_0^t \mathrm{d}H(s)/r^{(k)}(s)$, $k = 1, 2$. In order to estimate (5), we need the estimates of joint characteristics of the processes. As regards $H(t)$, the Nelson–Aalen estimator is available, cf. part 3.1. The moments $\mu(t)$ and $\sigma^2(t)$ can be estimated e. g. with the help of the moving window (or kernel) approach. We can then compute (approximatelly) the test process

$$D_m(t) = \sqrt{\frac{m_1 m_2}{m}} \int_0^t \frac{\mathrm{d}(\overline{C}^{(1)}(s) - \overline{C}^{(2)}(s))}{\sqrt{\mu^2(s) + \sigma^2(s)}} \Big/ (1 + V^*(t)), \tag{6}$$

which again behaves asymptotically as the process of Brownian bridge. Therefore, the test of $\boldsymbol{H}_0$ is then performed in a quite similar way as the goodness-of-fit test, i. e. by evaluation of $d = \max |D_m(t)|$ on $[0, T]$ and taking $\exp(-2d^2)$ as an approximate P-value of the test against a one-sided alternative.

5. EXAMPLE OF THE TEST OF HOMOGENEITY

As an example, let us consider one-day processes of financial transactions performed via credit cards at two different gas stations, both for $m_1 = m_2 = 100$ days. We follow both the number of transactions (forming the counting process) and also the

cumulation of transferred amounts (from this we obtain the cumulative process), $t \in [0, 24]$ hours. Figure 1 contains the averaged processes $\overline{N}^{(k)}(t) = \frac{1}{m_k} \sum_{i=1}^{m_k} N_i^{(k)}(t)$ and estimated corresponding hazard functions $\widehat{h}^{(k)}(t)$, $k = 1, 2$. Estimates of hazard function were obtained from the estimated cumulative hazard functions by a smoothing (kernel) technique.

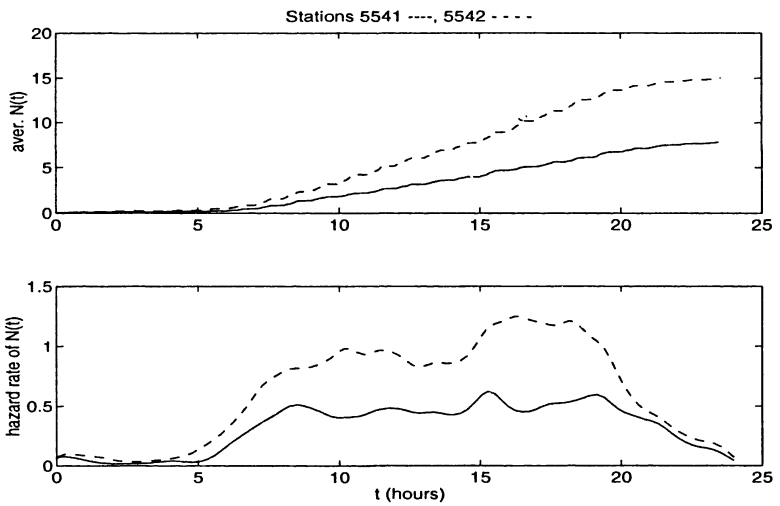

Fig. 1. Averaged counting processes and estimates of their hazard rates.

Figure 2 shows the averaged cumulative processes $\overline{C}^{(k)}(t) = \frac{1}{m_k} \sum_{i=1}^{m_k} C_i^{(k)}(t)$, and estimated and smoothed derivatives of functions $K^{(k)}(t)$.

From the graphical comparison we already see the difference between both sets of processes. By the numerical test of homogeneity computed in accordance with (6) we obtained that the minimum of $D_n(t)$ was -3.841, which was highly significant (P-value was $\sim 10^{-13}$). Functions $\mu(t)$ and $\sigma^2(t)$ were estimated with the aid of the moving window procedure, $H(t)$ by the Nelson–Aalen estimator.

6. CONCLUSION

The main advantage of the counting processes is their dynamics resulting from the conditioning the actual intensities by the history of the system. This area of statistical methods has a well developed theoretical background as well as the techniques of computational analysis.

The main purpose of the paper was to describe and analyze the random process (called here the cumulative process) consisting in the combination of the counting process with the process of random increments. Such models are suitable for description of many real-world technological, environmental, biological (and also financial) processes. We derived tools for modelling and statistical analysis of such situations,

namely we proposed the estimator of the rate of the cumulative process and proved its large sample properties. These properties were utilized in the proposal of procedures for the test of agreement of the data with the cumulative process model and for the tests of homogeneity of two cumulative processes.

As regards the generalization of the case studied in the present paper, the first one should consider a functional model for the hazard function describing also the influence of the history of $C_i(t)$ on the actual intensity. For instance, regression models (variants of Cox model, say) are available for such a case. Another generalization should omit the assumption of the independence of variables $Y_i(t)$ on the history and should deal with increments generated by a specific random process model.

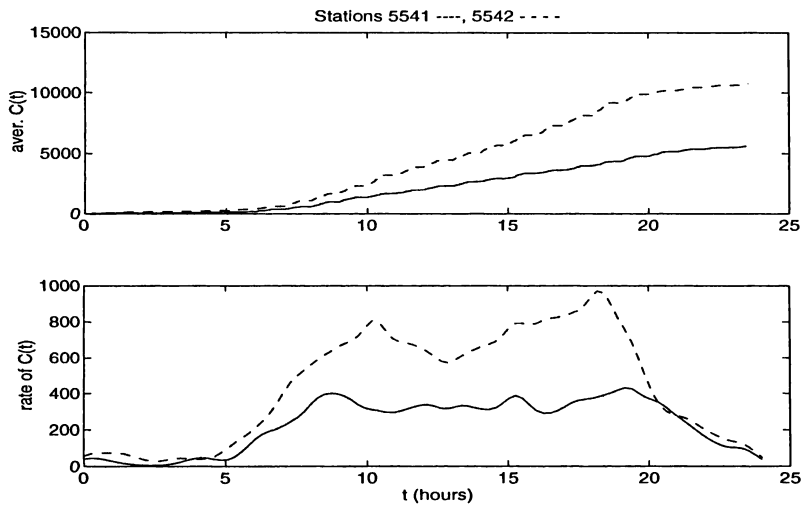

Fig. 2. Averaged cumulative processes and estimates of their rates.

ACKNOWLEDGEMENT

The present study was supported by the Grant Agency of the Czech Republic under grants 201/97/0354 and 402/98/0742.

(Received October 20, 1998.)

REFERENCES

[1] P. K. Andersen and O. Borgan: Counting process models for life history data: A review. Scand. J. Statist. *12* (1985), 97–158.

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

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

[4] P. Embrechts, C. Klüppelberg and T. Mikosch: Modelling Extremal Events. Springer, Heidelberg 1997.
[5] T. R. Fleming and D. P. Harrington: Counting Processes and Survival Analysis. Wiley, New York 1991.
[6] P. Volf: Analysis of generalized residuals in hazard regression models. Kybernetika *32* (1996), 501–510.
[7] P. Volf: On counting process with random increments. In: Proceedings of Prague Stochastics'98, Union of Czech Math. Phys., Prague 1998, pp. 587–590.
[8] B. B. Winter, A. Földes and L. Rejtö: Glivenko–Cantelli theorems for the product limit estimate. Problems Control Inform. Theory *7* (1978), 213–225.

Petr Volf, CSc., 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