

# Sequential testing of Cox proportional hazard models

## Victor D. Zurkowski

### University of Toronto

**Abstract.** We study survival processes that follow a parametric model in which parameters are estimated. We introduce the *sequential (maximum likelihood) estimator*, discuss "geometric" large sample limits of resulting residual processes, and illustrate computations by testing data for compliance with a constant rate Poisson process. We state, without proofs, large sample limit properties of the sequential estimator and residual martingales in the case of the semi-parametric Cox proportional hazard model. We also single out a (finite) number of residual processes and show that the simultaneous vanishing of their means is equivalent to compliance with Cox proportional hazards model.

**1. Introduction.** [Durbin] considers the limit distribution of the process

$$y(t) = \sqrt{n}(\hat{F}(t) - t), \quad 0 \leq t \leq 1,$$

where  $\hat{F}(t)$  is the empirical distribution of a size  $n$  sample of a random variable with probability distribution  $F(\cdot, \mathbf{q})$ ,  $\mathbf{q} \in \mathcal{Q}$  (after a suitable data-dependent transformation). The value of  $\mathbf{q}$  is unknown, and it is estimated from the sample. Under reasonable assumptions,  $y$  converges weakly towards a gaussian process. However, unlike the case in which  $\mathcal{Q}$  reduces to a single point, in general, the limit depends on the unknown parameter  $\mathbf{q}$ .

About a decade earlier, electrical engineering had begun to actively apply linear (Kalman) filters. The problem of linear filtering is to estimate the state  $u(t)$  of a system at time  $t$ , given the value of some linear function  $Y$  of  $u$  before  $t$ . This leads naturally to the consideration of  $E(du(t) | \mathcal{F}_t^-)$  where  $\mathcal{F}_t^- = \sigma$ -algebra generated by  $\{Y(s) | 0 \leq s < t\}$ . The method of regressing a process onto past observation has since then being apply to many models.

Starting around 1975, Aalen applies continuous time martingales, stochastic integration, and counting processes to survival analysis. From an abstract point of view, Aalen's approach to event history analysis is a form of regression, where the dependent variable is a random measure  $dN(t)$  on a line interval  $[0, T]$ , induced by a counting process  $N$ , whereas the independent variable is an increasing family  $\{\mathcal{F}_t\}_{t \in I}$  of  $\sigma$ -algebras. For many models, the expected value  $E(dN(t) | \mathcal{F}_t^-)$  can be computed explicitly, which makes possible the testing of hypothesis by means of statistics of the form

$$(1.1) \quad Z = \int_I K(s) [dN(s) - E(dN(s) | \mathcal{F}_s^-)],$$

see [Andersen et al], section VI.3.3.

The archetypal example of the situation covered by this method is as follows: let  $\{T_i\}_{1 \leq i \leq n}$  be a random sample drawn from a non-negative random variable  $T$ , with density  $f$  and distribution  $F$ .  $T$  represents the waiting time until an event of interest occurs. Let  $N(t) = \sum_{i=1}^n I(T_i \leq t)$  be the number of events that occurred in the interval  $[0, t]$ . Let  $\mathcal{F}_t^- = \sigma$ -algebra generated by  $\{N(s) | 0 \leq s < t\}$ . Then – abusing notation –

$$E(dN(t) | \mathcal{F}_t^-) = Y(t)h(t)dt,$$

where  $Y(t) = n - N(t)$  is the number of events yet to occur after or at time  $t$ , and  $h(t) = f(t)/(1 - F(t))$  is the hazard function. Note that  $dN(t) - Y(t)h(t)dt$  is an instance of an "observed – expected" statistic.  $Z$  can be seen as a weighted form of a goodness-of-fit statistics.

The weak converge results applied by Durbin are also available for the large sample limit of (1.1). [Hjort] considers parametric models with hazard  $h(t) = \mathbf{a}(t, \mathbf{q}_0)$ . Here is a brief summary of Hjort results. Suppose that observations are made during a finite interval  $[0, T]$ . Let

$$(1.2) \quad H_n(t) = \sqrt{n} \int_0^t K_n(s) J(s) \left\{ \frac{dN(s)}{Y(s)} - \mathbf{a}(s, \hat{\mathbf{q}}) ds \right\},$$

where  $J(s) = I(Y(s) > 0)$ ,  $K_n(s) \xrightarrow{p} k(s, \mathbf{q})$ ,  $\frac{Y(s)}{n} \xrightarrow{p} y(s)$ ,

and  $\hat{\mathbf{q}}$  is an estimation of  $\mathbf{q}$ . Hjort takes  $\hat{\mathbf{q}}$  to be the maximum likelihood estimator of  $\mathbf{q}$ , i.e.:  $\hat{\mathbf{q}}$  is the solution of the score equation:

$$(1.3) \quad \int_0^T \frac{\partial \ln(\mathbf{a}(s, \hat{\mathbf{q}}))}{\partial \mathbf{q}} \{dN(s) - Y(s)\mathbf{a}(s, \hat{\mathbf{q}}) ds\} = 0.$$

Set  $\mathbf{y}(s, \mathbf{q}) = \frac{\partial \ln(\mathbf{a}(s, \mathbf{q}))}{\partial \mathbf{q}}$  (a row vector). Under suitable

(smoothness, integrability, and non-vanishing of  $y$ )

conditions,  $\hat{\mathbf{q}}$  is a consistent estimator of  $\mathbf{q}_0$ , and

$$(1.4) \quad \sqrt{n}(\hat{\mathbf{q}} - \mathbf{q}_0) \rightarrow_d N(0, \Sigma^{-1})$$

where  $\Sigma$  is the matrix

$$\Sigma = \int_0^T \mathbf{y}(s, \mathbf{q}_0)^* \mathbf{y}(s, \mathbf{q}_0) y(s) \mathbf{a}(s, \mathbf{q}_0) ds.$$

Finally, let  $V$  be a gaussian martingale with  $\text{Var}(dV(s)) = y(s)\mathbf{a}(s, \mathbf{q}_0)ds$ . Hjort finds that under the hypothesis that  $h(t) = \mathbf{a}(t, \mathbf{q}_0)$ ,

$$H_n \xrightarrow{d} H \text{ in } D[0, T],$$

where:

$$(1.5) \quad H(t) = \int_0^t \frac{k(s, \mathbf{q}_0)}{y(s)} dV(s) - \left( \int_0^t k(s, \mathbf{q}_0) y(s, \mathbf{q}_0) \mathbf{a}(s, \mathbf{q}_0) ds \right) \Sigma^{-1} \left( \int_0^T \mathbf{y}(s, \mathbf{q}_0)^* dV(s) \right).$$

The first term in  $H$  is the limit of  $H_n$  if in the definition of  $H_n$  one uses the true value of  $\mathbf{q}$  instead of the estimate  $\hat{\mathbf{q}}$ . The distribution of  $H$  coincides with the

distribution of  $H_1(t) = \int_0^t \frac{k(s, \mathbf{q})}{y(s)} dV(s)$  subject to the

constraint:

$$(1.6) \quad \int_0^T \mathbf{y}(s, \mathbf{q})^* \frac{y(s)}{k(s, \mathbf{q})} dH_1(s) = 0.$$

It is worth noticing that the distribution of  $H$  depends on the sampling method, since it involves the large sample proportion  $y(t)$  of events yet to be observed after  $t$ , in particular the effect of "censoring" enters  $H$  through  $y$ . More than one way has been suggested to deal with the complicated distribution of  $H$ . We will review some of the suggestions, and add our contribution to the list.

## 2. Tests involving constrained gaussian martingales.

This section is not a comprehensive review of the subject. References are only an indication for further readings. Familiarity with martingale calculus is assumed, as presented in [Andersen et al.], and only "geometric" arguments will be considered, disregarding issues of smoothness, boundness, existence of solutions, etc. For the most part, arguments can be made rigorous by applying the canonical representation of gaussian process as in Hida et al (although the form of  $H$  in 1.5 suffices), stochastic integration with respect to a semimartingale, and Ito's formula.

We consider a process  $H$ . The question is to assess whether  $H$  is as in 1.5. Processes like the one 1.5 arise as large sample limits in the context of goodness-of-fit test. In practice, whether the observed (finite sample) process has the fractal roughness exhibited by almost all paths of the right side of 1.5 is usually not an issue. Attention is placed in determining whether the observed process is consistent with a process with mean and covariance as in 1.5.

Here is a list (far from complete) of suggested methods that lead to formal statistical tests.

$V, H$  are as in (1.5).

**2.1 Simulation.** Since  $V$  is a gaussian martingale, an approximation  $V^\#$  to a realization of  $V$  can be obtained as a linear combination of independent standard normals. Since the coefficients  $\mathbf{y}, k, y, \mathbf{a}$  are known, it is easy to derive an approximation  $H^\#$  of a realization of  $H$  from  $V^\#$ . By generating many  $H^\#$ 's, one approximates the distribution of  $H$ . This crude simulation approach should work given enough computational power.

A more interesting simulation could follow the idea in [Lin, Wei, Yin], [Lin, Spikerman], [Li, Sun], etc.: consider the counting process  $N_i$  that describes the progress of subject  $i$ . The martingale  $M_i$  associated to  $N_i$  satisfies  $E(M_i(t)) = 0$  and  $\text{var}(M_i(t)) = E(N_i(t))$ . Replace  $M_i$  with  $G_i N_i$ , where  $\{G_i\}_{1 \leq i \leq n}$  are independent standard normal variables (independent of the data). The resulting process  $\hat{W}$  has the asymptotic limit (1.5).  $\hat{W}$  given the data is a gaussian process whose distribution is obtained by simulation.

**2.2  $c^2$  projection.** Let  $0 = t_0 < t_1 < \dots < t_m = T$ . Then

$$U = (H(t_1), H(t_2) - H(t_1), \dots, H(t_m) - H(t_{m-1}))^*$$

is a gaussian vector, whose covariance  $R$  can be obtained from (1.5):  $R_{ij} = \text{cov}(U_i, U_j) =$

$$= \int_{I_i \cap I_j} \frac{k^2}{y} \mathbf{a} ds - \left( \int_{I_i} k y \mathbf{a} ds \right) \Sigma^{-1} \left( \int_{I_j} k y^* \mathbf{a} ds \right)$$

where  $I_i = (t_{i-1}, t_i)$ ,  $i = 1, \dots, m$ . It follows that if  $R$  is full rank,  $X^2 = U R^{-1} U$  has a  $\chi^2$  distribution with  $df = m$ . See [Hjort] for details.

**2.3 Innovation process.** As mentioned in 2.2, the process  $H$  has correlated increments. With notation as in 2.2, assuming  $R$  has full rank, it is possible to find a lower triangular matrix  $L$  with 1's along the diagonal, such that

$$D = L^{-1} R (L^{-1})^*$$

is diagonal (this is either the LU decomposition applied to symmetric, positive definite matrices; or the Gram-Schmidt orthogonalization process). Then  $U = L D^{1/2} Z$ , where  $Z$  is a vector of independent standard normals. Since  $L$  is lower triangular, for  $j = 1, \dots, m$ ,  $\sigma\{U_i | i \leq j\} = \sigma\{Z_i | i \leq j\}$ . The same idea applies to the process  $H$  in 1.5, (assuming the measures induced by  $H$  and by  $V$  on function spaces are equivalent), there is a gaussian martingale  $W$  such that:

$$(2.3.1) \quad dW(t) = dH(t) - \left( \int_{(0,t]} L(t, \mathbf{x}) dH(\mathbf{x}) \right) dt$$

and for all  $t \in [0, T]$ :

$$\sigma\{W(s) | s \leq t\} = \sigma\{H(s) | s \leq t\}.$$

See [Hida et al]. Because of constraints 1.6 satisfied by  $H$ , the measures induced by  $H$  and  $V$  are only equivalent on  $[0, t]$  for  $t < T$ . [Khmaladze, 1981] establishes that for 1.5,

$$(2.3.2) \quad L(t, \mathbf{x}) =$$

$$= k(t) \mathbf{y}(t) \mathbf{a}(t) \left( \int_0^t \mathbf{y}(x) \mathbf{y}(x)^* \mathbf{y}(x) \mathbf{a}(x) dx \right)^{-1} \mathbf{y}(x)^* \frac{\mathbf{y}(x)}{k(x)}$$

(the parameter  $\mathbf{q}$  was omitted to lighten the notation,). It is remarkable that for 1.5,  $L$  can be given explicitly.

Informally,  $dW(t)$  is the novel part of "information" added by  $dH(t)$  to  $\sigma\{H(s) | 0 \leq s < t\}$ . The innovation process  $W$  has the same distribution as a time-changed brownian motion.

**2.4 Imputation.** Randomized tests introduce "extra" randomness in the data to facilitate the analysis. In [Li, Sun] the introduction of external randomness is used to simplify working with a tied brownian motion. We find that judicious addition of "noise" to  $H$  leads to a formal test. Explicitly, let  $u$  be a random vector with distribution  $N(0, \Sigma^{-1})$ , independent of  $\sigma\{H(s) | 0 \leq s \leq T\}$ , and consider the process:

$$W(t) = H(t) + \left( \int_0^t k(\mathbf{x}) \mathbf{y}(\mathbf{x}) \mathbf{a}(\mathbf{x}) d\mathbf{x} \right) u$$

It is easy to verify that  $W$  is a gaussian process with uncorrelated increments, and  $\text{var}(W(t)) = \int_0^t k(\mathbf{x}) \mathbf{a}(\mathbf{x}) d\mathbf{x}$ .

The two terms that comprise  $W(t)$  can be recovered

from  $W$ . Indeed,  $\int_0^t \frac{\mathbf{y}}{k} \mathbf{y}^* dW = \Sigma u$ , therefore

$$H(t) = W(t) - \left( \int_0^t k \mathbf{y} \mathbf{a} d\mathbf{x} \right) \Sigma^{-1} \left( \int_0^t \frac{\mathbf{y}}{k} \mathbf{y}^* dW \right).$$

Geometrically,  $dH$  is the projection of  $dW$  onto the space orthogonal to  $\mathbf{y}\mathbf{y}^*/k$ . The component of  $dW$  parallel to  $\text{span}\{\mathbf{y}\mathbf{y}^*/k\}$  is imputed randomly.

A formal statistical test can be implemented as follows: let  $\mathbf{A}$  be a set of functions defined on  $[0, T]$  such that  $P(W \in \mathbf{A}) = \alpha$ . Let  $H$  be a process defined on  $[0, T]$ .

Simulate an imputation  $a(t) = \left( \int_0^t k(\mathbf{x}) \mathbf{y}(\mathbf{x}) \mathbf{a}(\mathbf{x}) d\mathbf{x} \right) u$ . If  $H$

belongs to the (random) set  $\mathbf{A} - a$ , then reject that  $H$  is of the form prescribed in 1.5, otherwise conclude  $H$  is consistent with 1.5. Of course, the choice of  $\mathbf{A}$  should be dictated by power considerations. The size of this test is

$$P(\text{reject } H_0 | H_0) = E_0\{P(H \in \mathbf{A} - a | H_0, a)\} = P(W \in \mathbf{A} | H_0) = \alpha.$$

**3. Comments on the previous testing methods.** Because of dependence of 1.5 upon  $\mathbf{q}_0$  and  $\mathbf{y}$ , simulations have to be carried anew each time a process is considered. Moreover, questions of choice of  $k$  to maximized power against contiguous alternatives cannot be addressed easily by simulations. Then, simulation studies do not exhaust the analysis. Furthermore, if for no reason other than to verify the result of simulations, it is worth having alternative means of testing (the conclusion in [Ferrenberg, Landau, Wong]. comes to mind). See [Lin, Kosorok] for a recent simulation study pertaining to survival analysis.

[Li, Sun] give references to  $\chi^2$  and to innovation process tests. It also raises the key objections against those tests, to wit:  $\chi^2$  tests involve a subjective partition of the data, and could lead to loss of information (power), whereas the innovation process approach "lacks a clear statistical interpretation". The last statement requires an explanation. In practice, one works with a finite sample, and a statistic  $H_n$  like (1.2). By considering the large sample limit, one ends up looking at  $H$  as in (1.5). The innovation process approach applies to (1.5), and leads to a process  $W$  as implied by (2.3.1), (2.3.2).  $W$  involves unknown parameters, which have to be replaced by empirical estimators. The result is a process  $W_n$  that has  $W$  as large sample limit. The transformation from  $H$  to  $W$  has a clear interpretation in terms of the flow of information (i.e. filtration of events). However, the transformation from  $H_n$  to  $W_n$  lacks a simple direct interpretation. Andersen et al. give a short heuristic explanation of Khmaladze's goodness of fit idea (innovation process approach).

The idea of imputing to  $H$  an orthogonal component will be explored further elsewhere.

A host of graphical methods is in use. See [Fleming, Harrington], Chapter 4. We will not review them here, since we are more interest in analytical formal statistical tests.

#### 4. Sequential estimator in parametric survival models.

We propose to follow the strategy underlying the innovation process approach to hypothesis testing, namely: use information as it accrues in time. As a first example, we consider statistics of the form (1.1), where the counting process  $N$  follows the parametric model

$$(4.1) \quad E(dN(s) | \mathcal{F}_{s-}) = Y(s) \mathbf{a}(s, \mathbf{q}_0) ds$$

for some  $\mathbf{q}_0 \in \Theta$ .

Suppose observations are made during a finite interval of time  $[0, T]$ . For each  $t \in (0, T]$ , let  $\hat{\mathbf{q}}(t)$  be the

maximum likelihood estimator of  $\mathbf{q}$  based on observations during  $[0, t]$ , i.e.:  $\hat{\mathbf{q}}(t)$  is the solution of the score equation:

$$(4.2) \quad \int_{(0,t]} \frac{\partial \ln(\mathbf{a}(s, \hat{\mathbf{q}}))}{\partial \mathbf{q}} \{dN(s) - Y(s)\mathbf{a}(s, \hat{\mathbf{q}})\} ds = 0$$

Compare the domain of integration with the one in 1.3.

We will refer to  $\hat{\mathbf{q}}(\cdot)$  as the sequential estimator of  $\mathbf{q}_0$ . Note that the sequential estimator is adapted to the filtration associated to the counting process  $N$ . For now we are only interested in the "geometry" behind sequential estimators, thus we disregard all considerations about existence of  $\hat{\mathbf{q}}(t)$ , etc.

The same Taylor expansion that yields (1.4) implies:

$$(4.3) \quad \sqrt{n}(\hat{\mathbf{q}}(t) - \mathbf{q}_0) \rightarrow_d \Sigma(t)^{-1} \int_0^t \mathbf{y}(s, \mathbf{q}_0)^* dV(s)$$

where  $\Sigma(t) = \int_0^t \mathbf{y}(s, \mathbf{q}_0)^* \mathbf{y}(s, \mathbf{q}_0) y(s) \mathbf{a}(s, \mathbf{q}_0) ds$ , and  $V$  is a gaussian martingale with  $\text{Var}(dV(s)) = y(s) \mathbf{a}(s, \mathbf{q}_0) ds$ .

Following standard ideas, in order to assess (4.1), we introduce a weight function  $K_n$  (see [Andersen et al.], VI.3, and references therein) and let

$$(4.4) \quad H_n(t) = \sqrt{n} \int_0^t K_n(s) J(s) \left\{ \frac{dN(s)}{Y(s)} - \mathbf{a}(s, \hat{\mathbf{q}}(t)) \right\} ds,$$

where  $J(s) = I(Y(s) > 0)$ ,  $K_n(s) \rightarrow_p k(s, \mathbf{q})$ ,  $\frac{Y(s)}{n} \rightarrow_p y(s)$  as in (1.2). Then,

$$(4.5) \quad H_n \rightarrow_d H \text{ in } D[0, T],$$

where:

$$(4.6) \quad H(t) = \int_0^t \frac{k}{y} dV(s) - \left( \int_0^t k \mathbf{y} \mathbf{a} ds \right) \Sigma(t)^{-1} \left( \int_0^t \mathbf{y}^* dV(s) \right).$$

The difference between (1.5) and the last expression is the interval of integration in the last integral. The difference is due to the fact that the sequential estimator  $\hat{\mathbf{q}}(t)$  does not introduce in  $H_n$  uncertainty beyond  $t$ .

The following remarkable fact holds:

**4.7 Lemma.** The process  $H$  given in (4.6) has uncorrelated increments.

**Proof.** We recognize

$$\mathbf{p}_t(f)(\cdot) = \left( \int_0^t f \mathbf{y} \mathbf{y} \mathbf{a} ds \right) \Sigma(t)^{-1} \mathbf{y}(\cdot)^*$$

as the orthogonal projection of the function  $f$  onto the subspace spanned in  $L^2\{[0, t], \mathbf{y} \mathbf{a} ds\}$  by the components of  $\mathbf{y}(\cdot, \mathbf{q}_0)$ . So if we set  $\mathbf{r} = k/y$ , we can write

$$H(t) = \int_0^t (\mathbf{r}(s) - \mathbf{p}_t(\mathbf{r})(s)) dV(s).$$

Let  $\langle \cdot, \cdot \rangle_t$  denote the inner product in  $L^2\{[0, t], \mathbf{y} \mathbf{a} ds\}$ . Let  $0 < u < v < T$ . We have:  $E(H(u)H(v)) =$

$$= E \left( \left( \int_0^u (\mathbf{r} - \mathbf{p}_u(\mathbf{r})) dV \right) \left( \int_0^v (\mathbf{r} - \mathbf{p}_v(\mathbf{r})) dV \right) \right) =$$

$$\langle \mathbf{r} - \mathbf{p}_u(\mathbf{r}), \mathbf{r} - \mathbf{p}_v(\mathbf{r}) \rangle_u =$$

$$\langle \mathbf{r} - \mathbf{p}_u(\mathbf{r}), \mathbf{r} - \mathbf{p}_u(\mathbf{r}) \rangle_u + \langle \mathbf{r} - \mathbf{p}_u(\mathbf{r}), \mathbf{p}_u(\mathbf{r}) - \mathbf{p}_v(\mathbf{r}) \rangle_u.$$

But over the interval  $[0, u]$ , both  $\mathbf{p}_v(\mathbf{r})$  and  $\mathbf{p}_u(\mathbf{r})$  are linear combinations of the components of  $\mathbf{y}(\cdot, \mathbf{q}_0)$ . Hence  $\mathbf{p}_u(\mathbf{r}) - \mathbf{p}_v(\mathbf{r})$  is orthogonal to  $\mathbf{r} - \mathbf{p}_u(\mathbf{r})$  with respect to  $\langle \cdot, \cdot \rangle_u$ , and so:

$$E(H(u)H(v)) = \langle \mathbf{r} - \mathbf{p}_u(\mathbf{r}), \mathbf{r} - \mathbf{p}_u(\mathbf{r}) \rangle_u = E(H(u)^2)$$

does not depend on  $v$ . This proves the lemma.

**4.8 Corollary.** The process  $H$  given in (4.6) is a gaussian martingale.

**Proof.** Clearly  $H$  is a gaussian process. For gaussian processes, having uncorrelated increments imply having independent increments.

**4.9 Comments.** The interpretation of the sequential estimator  $\hat{\mathbf{q}}(t)$  is clear:  $\hat{\mathbf{q}}(t)$  is the MLE estimator of  $\mathbf{q}_0$  based on observations up to time  $t$ , and it reflects the way information is accrued.

The martingale  $H$  has variance

$$\text{var}(H(t)) = \int_0^t (\mathbf{r} - \mathbf{p}_t(\mathbf{r}))^2 \mathbf{y} \mathbf{a} ds,$$

while the innovation process mentioned in 2.3 has variance

$$\int_0^t \mathbf{r}^2 \mathbf{y} \mathbf{a} ds.$$

Thus, the martingale obtained by using the sequential estimator is not quite equivalent to the innovation process.

What we have called the "geometry" of the asymptotic properties of  $H_n$  is but a set of properties of projections in  $L^2([0, T], \mathbf{y} \mathbf{a} ds)$ . [Khmaladze, 1993] uses a family of projections in  $L^2$  to construct a *scanning innovation* for (1.5) viewed as function of  $\mathbf{r} = I_{(0,t]} K/y$ . The scanning innovation transform generalizes the innovation approach in more than one way: first, it applies to processes in more than one ("time") dimension, and second, it

considers an entire class of filtrations. From a conceptual point of view, the use of sequential estimators and the scanning innovation transform are similar. However, the two approaches lead to different residual martingales.

## 5. Examples.

**5.1 Testing a distribution for constancy of hazard.** Let  $\{T_i / i = 1, \dots, n\}$  be iid observations drawn from a non-negative random variable with hazard  $h$ . We want to test the hypothesis  $H_0: h(t) = \text{constant} = \mathbf{q}$ . Let

$$N(t) = \sum_{i=1}^n I(T_i \leq t),$$

$$F_t = \sigma\{N(s) / 0 \leq s \leq t\}.$$

Then (abusing notation)  $E(dN(s) / F_{s^-}) = Y(s)h(s)ds$ , where

$$Y(s) = \sum_{i=1}^n I(T_i \geq s) = n - N(s^-).$$

Likelihood considerations [Zurkowski] indicates that goodness of fit tests could be based on:

$$(5.1.1) \quad H_n(t) =$$

$$= \sqrt{n} \int_{(0,t]} \ln\left(\frac{Y}{n}\right) \frac{Y}{n} I(Y > 0) \left( \frac{dN}{Y} - \hat{\mathbf{q}}(t) ds \right) =$$

$$= \int_{(0,t]} \ln\left(\frac{Y}{n}\right) \left( \frac{dN - Y\hat{\mathbf{q}}(t) ds}{\sqrt{n}} \right),$$

which is of the form in (4.4). We have:

$$H_n \xrightarrow{d} H \text{ in } D[0,T],$$

where:

$$H(t) = \int_0^t \ln y dV(s) - \frac{\int_0^t y \ln y ds}{\int_0^t y ds} \left( \int_0^t dV(s) \right).$$

It is easy to verify directly that if  $0 < s < t$ , then  $E(H(s)H(t)) = E(H(s)^2)$ , thus  $H$  has uncorrelated increments.

The variance of  $H$  can be estimated by  $\hat{\mathbf{a}}(t)$  defined as:

$$(5.1.2) \quad \hat{\mathbf{a}}(t) = \left( \frac{N(t)}{\int_0^t Y ds} \right) \int_0^t (\ln(Y(s)) - m(t))^2 \frac{Y(s)}{n} ds$$

with

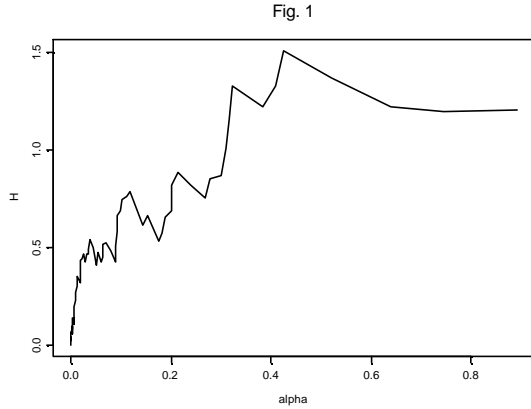
$$m(t) = \frac{\int_0^t \ln(Y)Y ds}{\int_0^t Y ds}$$

**5.2 Numerical example.** We will explore whether the rate of usage of the word "the" in [Kmaladze, 1993] is constant. We use the PostScript version available at the JSTOR database (<http://www.jstor.org/>) to measure the distance between occurrences of the word "the" along the written line. The unit of distance is *point*. We ignore the Abstract and start measuring from "1. Introduction. This....". We disregard white space around formulas and between sections. The measurements are done by hand, which introduce measurement errors. However, the order of magnitude of the errors is negligible in comparison with the distances being measured. Also, some further small rounding errors are introduced when measuring the length of formulas. Both measurement errors will be ignored. We consider the word "the" (with the space to the next word included) as well as "The". For the purpose of this example we look only at pages 798 to 802, including the latter. We count respectively 8, 25, 22, 22, and 17 occurrences of "the".

The distances between successive occurrences are (in order of occurrence): 2576, 253, 336, 483, 388, 560, 872, 197, 311, 399, 65, 1115, 407, 83, 207, 520, 68, 176, 348, 136, 266, 337, 380, 314, 145, 309, 448, 657, 325, 605, 239, 166, 1232, 912, 434, 200, 261, 115, 2050, 64, 1275, 75, 549, 769, 382, 895, 863, 303, 276, 429, 307, 456, 85, 638, 520, 1048, 241, 578, 1299, 82, 101, 561, 971, 75, 154, 248, 641, 702, 201, 335, 784, 1092, 895, 200, 665, 1109, 458, 1099, 257, 338, 641, 110, 204, 1771, 383, 845, 638, 132, 1030, 364, 64, 314, 685, 1501.

The counting process methodology assumes ties in the data occur with probability 0. The presence of ties requires modifications of the cumulative hazard estimators. The idea is to take a suitable limit of a process without ties. Alternatively, we can introduce a small perturbation of the data. We opt for the latter solution, and add to each datum  $T_i$  a random noise  $u_i \sim \text{uniform}[-0.05, 0.05]$ . Then we proceed as indicated in section 5.1. Figure 1 shows the graph of  $H$  vs.  $a = \text{var}(H)$ .

Fig. 1 depicts a graph that is not atypical for an approximation to a brownian motion obtained from a sample of about 100 observations that are consistent with the hypothesis of constant hazard. In an Appendix we include the definition of a function written in Splus, which, given a vector  $X$  of event times, computes  $H$  and  $\mathbf{a}$  as in (5.1.1), (5.1.2).



We can quantify how extreme the observed process  $H$  is by means of the statistic  $\sup_{0 < t < T} |H(t)|$ . We have:

$$P(\sup_{0 < t < T} |H(t)| \geq I) \approx P(\sup_{0 < t < T} |B(\mathbf{a}(t))| \geq I) =$$

$$P(\sup_{0 < u \leq \mathbf{a}(T)} |B(u)| \geq I) = P(\sup_{0 < x < 1} |B(x)| \geq \frac{I}{\sqrt{\mathbf{a}(T)}}) =$$

$$= 1 - \frac{4}{p} \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \exp\left(-\frac{p^2(2k+1)^2 \mathbf{a}(T)}{8I^2}\right).$$

Thus,  $p$  defined as:

$$(5.2.1) \quad p = 1 - \frac{4}{p} \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \exp\left(-\frac{p^2(2k+1)^2}{8I^2}\right) \Bigg|_{I = \frac{\sup_{0 < t < T} |H(t)|}{\sqrt{\mathbf{a}(T)}}}$$

has a distribution that is asymptotically uniform  $[0,1]$ , with small values of  $p$  corresponding to deviations from the null hypothesis (see Appendix 2). For the data set under consideration,  $p = 0.2209458$ .

## 6. Cox proportional hazards model.

In section 5 we applied the method outlined in section 4 to test for constancy of the hazard. The choice of a Poisson model for the example is not ingenuous, for the sequential estimator of the unknown rate  $\mathbf{q}$  can be given explicitly, therefore avoiding the need for extensive computations. By contrast, testing for compliance with Cox proportional hazard model requires more computations. From the point of view of complexity, the sequential estimator approach and the innovation martingale approach are of the same order: both methods require some computation at each failure time. The innovation martingale requires at each event time the inversion of a matrix (whose dimension is determined by the dimension of the unknown parameter  $\mathbf{q}$ , and remains the same for all sample sizes). The

sequential estimator requires at each time the estimation of  $\mathbf{q}$ . Thus, for a data set comprised of  $n$  event times, the number of computations required by both methods are roughly proportional (although the proportionality constant might be sizable).

The "geometric" argument outlined in section 4 applies to the semi-parametric Cox model after suitable changes. Much of what we will add specifically for the Cox model pertains to the issue of showing that sequential estimator exists. We state our results. Full proofs will be provided elsewhere.

**6.1 Notation, assumptions.** We consider  $n$  "event" times (failure times):  $T_1, \dots, T_n$  produced by  $n$  observational units. We posit the existence of a probability space  $(\Omega, \mathcal{F}, P)$ , and that  $\mathcal{F}, P$  are such that  $\{T_i\}_{1 \leq i \leq n}$  are  $\mathcal{F}$ -measurable, and independent. In addition we are given the following data:

- i) a "risk status" process (covariate process)  $z_i$  for each observational unit  $i$ ;
- ii) an event mode ("failure mode", or "cause of failure")  $\mathbf{d}_i$  for each  $i$ ;
- ii) a filtration of  $\sigma$ -algebras  $\mathcal{F}_* = \{\mathcal{F}_t\}_{0 \leq t \leq T}$  satisfying the usual conditions.

We also assume:

- iii) the processes  $t \mapsto N_{iq}(t) = I(T_i \leq t, \mathbf{d}_i = q)$  is  $\mathcal{F}_*$ -adapted, the processes  $z_i$  is  $\mathcal{F}_*$ -predictable.

The process  $N_{iq}$  is non-increasing, hence it admits a Doob-Meier decomposition. The risk  $dR_{iq}$  of subject  $i$  experiencing the event under consideration in mode  $q$  (the hazard process of  $i$  failing due to cause  $q$ ) is associated to the predictable part of  $N_{iq}$ :

$$(6.1.1) \quad d\Lambda_{iq}(t) = E(d^- N_{iq}(t) | \mathcal{F}_{t-}) = Y_i(t) dR_{iq}(t)$$

where the process  $Y_i$  is predictable. When  $\mathcal{F}_t$  is generated by  $\{N_{iq}\}_q$ , we have  $Y_i(t) = I(T_i \leq t)$ . In any case, the process  $Y_i$  has the heuristic interpretation of describing the survivorship of subject  $i$  at time  $t$ . We assume the risk  $R_{iq}$  is determined by the risk status process  $z_i$ , specifically, we consider only the case in which  $R_{iq}$  has the functional form:

$$dR_{iq}(t) = dR_q(t, z_i(t))$$

where  $R_q(\cdot, \mathbf{x})$  is a non-decreasing, non-random function defined on the interval  $[0, T]$ .

Finally, we assume  $z_i$  takes values in a finite set  $Z \subset \mathbb{R}^d$ , and that  $\mathbf{d}_i$  takes values in the finite set  $\{1, \dots, Q\}$ .

**6.2 Cox model.** We are interested in assessing whether the data are consistent with Cox proportional hazard model. Explicitly, we have a formal statistical problem, with null hypothesis:

$\mathbf{H}_0$ : for each event mode  $q$ , there is  $\mathbf{b}_q \in (\mathbb{R}^d)^*$ , and a Borel measure  $d\Lambda_q$  on  $[0, T]$  such that for all risk status  $\mathbf{x}$

$$dR_q(\cdot, \mathbf{x}) = e^{\mathbf{b}_q \mathbf{x}} d\Lambda_q(\cdot)$$

i.e. for a given  $q$ , and for  $\mathbf{x}_1, \mathbf{x}_2 \in \mathbf{Z}$ , the two risks  $dR_q(\cdot, \mathbf{x}_1)$  and  $dR_q(\cdot, \mathbf{x}_2)$  are not only proportional, but also the proportionality constants depend (one may say) *log-linearly* upon the risks.

The alternative hypothesis is that the risks  $\{R_q(\cdot, \mathbf{x})\}_{q, \mathbf{x}}$  do not have the functional form proposed in the null hypothesis.

**6.3 Test statistic.** As in the example presented in section 5, likelihood considerations (see [Zurkowski]) indicates a statistic upon which goodness of fit tests could be based. We need some assumptions and a few definitions before proceeding further:

Let:

$$\Delta N_{q\mathbf{x}}(s) = \sum_{i=1}^n \Delta N_{iq}(s) I(z_i(s) = \mathbf{x})$$

$$Y_{\mathbf{x}}(s) = \sum_{i=1}^n Y_i(s) I(z_i(s) = \mathbf{x}).$$

Assume:

(6.3.1) there exist functions  $y_{\mathbf{x}}$  (with  $\xi \in \mathbf{Z} =$  set of risk states) such that for all  $t > 0$ ,

$$\lim_{n \rightarrow \infty} \sup_{s \in [0, t]} \left| \frac{1}{n} \sum_{i=1}^n Y_i(s) I(z_i(s) = \mathbf{x}) - y_{\mathbf{x}}(s) \right| = 0 \text{ a.s. [P]}$$

(6.3.2) there is  $\varepsilon > 0$  such that  $y_{\mathbf{x}}(s) > 0$  for all  $s \in [0, T]$ .

(6.3.3) (identifiability of  $\{\beta_q\}_q$ ) the set  $\mathbf{Z}$  of risk states is not contained in a hyperplane.

(6.3.4) Defined  $Z$  as follows:

i) wait until at time  $t_0$  by which at least one *failure* has been observed at each failure mode  $q$ , and each failure state  $\xi$  (the intuitive meaning of  $t_0$  is clear:  $t_0$  is the first time by which estimates of the hazards  $dR_q(\cdot, \mathbf{x})$  do not vanish). We will assume that the sample size is large enough that  $Y_i(s) > 0$  for all  $s \in [0, T]$ .

ii) for  $t \geq t_0$ , let  $\hat{\mathbf{b}}(t) = (\hat{\mathbf{b}}_q(t))_q$ , be the value of  $\mathbf{b} = (\mathbf{b}_q)_q$  that maximizes the *partial log-likelihood*:

$$\hat{\ell}_0(t, \mathbf{b}) = \sum_{q, \mathbf{x}} \int_{(0, t]} \left( \mathbf{b}_q \mathbf{x} - \ln \left( \sum_{\mathbf{n}} Y_{\mathbf{n}}(s) e^{\mathbf{b}_q \mathbf{n}} \right) - 1 \right) dN_{q\mathbf{x}}(s).$$

It can be shown [Zurkowski], that  $\mathbf{b} \mapsto \hat{\ell}_0(t, \mathbf{b})$  attains a maximum value when  $t \geq t_0$ .

iii) let  $d\hat{R}_q(\cdot, \mathbf{x})$  be the Breslow estimator of  $dR_q(\cdot, \mathbf{x})$  under

$\mathbf{H}_0$  ( $dR_q(\cdot, \mathbf{x})$ ), i.e. the estimator given by:

$$d\hat{R}_q(s, \mathbf{x}) = e^{\hat{\mathbf{b}}_q \mathbf{x}} \frac{\sum_{\mathbf{n}} dN_{q\mathbf{n}}(s)}{\sum_{\mathbf{n}} Y_{\mathbf{n}}(s) e^{\hat{\mathbf{b}}_q \mathbf{x}}}$$

iv) define  $Z$  as:

if  $t \geq t_0$ , set  $Z(t) =$

$$= \sum_{q, \mathbf{x}} \int_{(t_0, t]} \ln \left( \frac{Y_{\mathbf{x}}(s)}{n} \right) \left( dN_{q\mathbf{x}}(s) - Y_{\mathbf{x}}(s) d\hat{R}_q(s, \mathbf{x} | t) \right),$$

and if  $t < t_0$  then set  $Z(t) = 0$ .

**6.4 Large sample distribution of  $Z$ .** Suppose that for each  $n$  (sufficiently large) we are given a sample of size  $n$ . Let  $X_n, Y_{n, \mathbf{x}}, t_{n0}$ , etc. be the objects defined above corresponding to the sample of size  $n$ .

In order to describe the asymptotic form of the distribution of  $Z$  we will assume:

(6.4.1) There are processes  $\mathbf{w}_{\mathbf{x}}$  for  $\xi \in \mathbf{Z}$ , such that:

$$\sqrt{n} \left\{ \frac{Y_{\mathbf{x}}(\cdot)}{n} - y_{\mathbf{x}}(\cdot) \right\}_{\mathbf{x}} \xrightarrow{weak} \left\{ \mathbf{w}_{\mathbf{x}}(\cdot) \right\}_{\mathbf{x}} \text{ in } D([0, \tau])^{\#(\mathbf{Z})}$$

We introduce the following notation:

Let  $\{\mathbf{b}_q^0(t)\}_{1 \leq q \leq Q}$  be the vector that maximizes:

$$(6.4.2) \sum_{q, \mathbf{x}} \int_{(0, t]} \left\{ \mathbf{b}_q \mathbf{x} - \ln \left( \sum_{\mathbf{n}} y_{\mathbf{n}}(s) e^{\mathbf{b}_q \mathbf{n}} \right) \right\} y_{\mathbf{x}}(s) dR_q(s, \mathbf{x}).$$

Note that under the null hypothesis,

$$\mathbf{b}_q^0(t) = \text{constant in } t \text{ (= "true" } \mathbf{b}_q).$$

Let

$$(6.4.3) J_q(t) = \int_{(0, t]} M(s | t) \left( \sum_{\mathbf{n}} y_{\mathbf{n}} dR_q(s, \mathbf{n}) \right)$$

where  $M(s | t)$  is the positive semi-definite matrix:

$$M(s | t) = \frac{\sum_{\mathbf{n}} y_{\mathbf{n}}(s) e^{\mathbf{b}_q \mathbf{n}} \mathbf{n}^{\otimes 2}}{\sum_{\mathbf{n}} y_{\mathbf{n}}(s) e^{\mathbf{b}_q \mathbf{n}}} - \left( \frac{\sum_{\mathbf{n}} y_{\mathbf{n}}(s) e^{\mathbf{b}_q \mathbf{n}} \mathbf{n}}{\sum_{\mathbf{n}} y_{\mathbf{n}}(s) e^{\mathbf{b}_q \mathbf{n}}} \right)^{\otimes 2}$$

**Note:**  $\mathbf{x}^{\otimes 2}$  is short for  $\mathbf{x} \mathbf{x}^*$

**Facts.** Under assumptions 6.3.1-6.3.3, 6.5.1, notation as in 6.5.2-4, the following holds.

i)  $t_{n0} \rightarrow 0$  a.s.[P]

ii) **asymptotic form of b.** Let

$$B_q(t) = \left\{ \sum_{\mathbf{x}} \int_{(0,t]} \left[ \mathbf{x} - \frac{\sum_n y_n \mathbf{n} e^{b_q \mathbf{n}}}{\sum_n y_n e^{b_q \mathbf{n}}} \right] (w_{\mathbf{x}}(s) dR_q(s, \mathbf{x}) + dW_{q\mathbf{x}}(s)) + \sum_{\mathbf{x}} \int_{(0,t]} \left[ \frac{(\sum_n y_n e^{b_q \mathbf{n}})(\sum_n w_n e^{b_q \mathbf{n}})}{(\sum_n y_n e^{b_q \mathbf{n}})^2} - \frac{(\sum_n w_n e^{b_q \mathbf{n}})}{(\sum_n y_n e^{b_q \mathbf{n}})} \right] y_{\mathbf{x}} dR_q(s, \mathbf{x}) \right\}$$

where  $\{W_{q\mathbf{x}}\}$  are the weak limits of

$$\frac{1}{\sqrt{n}} \int_{(0,t]} dN_{q\mathbf{x}} - Y_{\mathbf{x}}(s) dR_q(s, \mathbf{x}), \text{ i.e.: } \{W_{q\mathbf{x}}\} \text{ are gaussian, zero-}$$

mean martingales, with  $\text{Var}(W_{q\mathbf{x}}(t)) = \int_{(0,t]} y_{\mathbf{x}}(s) dR_q(s, \mathbf{x}),$

$\text{cov}(W_{q\mathbf{x}}(t), W_{qn}(s)) = 0$  if  $(q, \mathbf{x}) \neq (p, \mathbf{n})$ ; and

$\mathbf{b}_q = \mathbf{b}_q^0(t)$  (evaluated at  $t$ ). Then:

$$(6.5.4) \quad \sqrt{n}(\hat{\mathbf{b}}_q(t) - \mathbf{b}_q^0(t)) \xrightarrow{\text{weak}} J_q(t)^{-1} B_q(t) \text{ in } D[\varepsilon, T] \text{ for any } \varepsilon \in (0, T).$$

iii) **asymptotic form of Z and similar processes.** Consider Z, and, more generally, a process G of the form:

$$(6.5.5) \quad G(t) = \sum_{q, \mathbf{x}} \int_{(t_0, t]} K_{q\mathbf{x}}(s) (dN_{q\mathbf{x}}(s) - Y_q(s) d\hat{R}_q(s, \mathbf{x} | t))$$

with  $t_0$  as in (6.3.4). Assume there are functions  $k_{q\mathbf{x}}$  such that

$$(6.5.6) \quad \sqrt{n} \{K_{q\mathbf{x}} - k_{q\mathbf{x}}\} \xrightarrow{\text{weak}} \{d_{q\mathbf{x}}\} \text{ in } (D[0, T])^a$$

where  $a = \#\{(q, \mathbf{x}) | q \text{ a failure mode, } \mathbf{x} \text{ a risk state}\}$ . Then

$$(6.5.7) \quad G(t) = nU(t) + \sqrt{n}H(t) + o_{\text{weak}}(1)$$

where:

$$U(t) = - \sum_{q, \mathbf{x}} \int_{(0,t]} \left[ k_{q\mathbf{x}} - \frac{\sum_n k_{qn} y_n e^{b_q \mathbf{n}}}{\sum_n y_n e^{b_q \mathbf{n}}} \right] y_{\mathbf{x}} dR_q(s, \mathbf{x})$$

with  $\mathbf{b}_q = \mathbf{b}_q^0(t)$  (evaluated at  $t$ ); and *under the null hypothesis*

$$H(t) = \sum_{q, \mathbf{x}} \int_{(0,t]} \left[ k_{q\mathbf{x}} - \frac{\sum_n k_{qn} y_n e^{b_q \mathbf{n}}}{\sum_n y_n e^{b_q \mathbf{n}}} \right] dW_{q\mathbf{x}}(s) +$$

$$+ \sum_q \left\{ \left( \int_{(0,t]} \sum_{\mathbf{x}} k_{q\mathbf{x}} \left( \mathbf{x} - \frac{\sum_n y_n \mathbf{n} e^{b_q \mathbf{n}}}{\sum_n y_n e^{b_q \mathbf{n}}} \right) y_{\mathbf{x}} e^{b_q \mathbf{x}} d\Lambda_q \right)^* \right. \\ \left. J_q(t)^{-1} \left( \int_{(0,t]} \sum_{\mathbf{x}} \left( \mathbf{x} - \frac{\sum_n y_n \mathbf{n} e^{b_q \mathbf{n}}}{\sum_n y_n e^{b_q \mathbf{n}}} \right) dW_{q\mathbf{x}} \right) \right\}$$

**Proof** will be given elsewhere.

**6.6 Comments.** Consider a process like G in (6.5.5-6). i) under the null hypothesis, the leading term U in the asymptotic form of G vanishes identically, and we find:

$$\frac{G}{\sqrt{n}} \xrightarrow{\text{weak}} H \text{ in } D[0, T].$$

ii) under the null hypothesis, H is a gaussian process. The expression of H in (6.5.7) is somewhat overwhelming. Our earlier "geometric" analysis in section 4 suggests that H should have a simple expression in terms of projections, and indeed, that is the case. Consider the Hilbert space H consisting of families  $\mathbf{f} = \{f_{q\mathbf{x}}\}_{q, \mathbf{x}}$  of functions with  $f_{q\mathbf{x}} \in L^2([0, T], y_{\mathbf{x}} \exp(\mathbf{b}_q \mathbf{x}) d\Lambda_q)$ , with inner product:

$$\langle \mathbf{f}, \mathbf{g} \rangle = \sum_{q, \mathbf{x}} \int_{(0, T]} f_{q\mathbf{x}} g_{q\mathbf{x}} y_{\mathbf{x}} \exp(\mathbf{b}_q \mathbf{x}) d\Lambda_q.$$

Let  $\mathbf{k}, \mathbf{g}^i$  be the element of H whose  $q\mathbf{x}$  component is given by:

$$\begin{aligned} \mathbf{k}_{q\mathbf{x}} &= k_{q\mathbf{x}}; \\ \mathbf{g}^i_{q\mathbf{x}} &= \mathbf{x}_i; \\ \mathbf{u}^p_{q\mathbf{x}} &= 1 \text{ if } p = q, 0 \text{ otherwise} \end{aligned}$$

and let  $\mathbf{p}_i$  be the orthogonal projection in H onto the subspace spanned by

$\{\mathbf{g}^i I_{[0,t]} | i = 1, \dots, d\} \cup \{\mathbf{u}^p I_{[0,t]} | p = 1, \dots, Q\}$ . Then,

$$(6.6.1) \quad H(t) = \sum_{q, \mathbf{x}} \int_{(0,t]} (\mathbf{k}_{q\mathbf{x}} - \mathbf{p}_i(\mathbf{k})_{q\mathbf{x}}) dW_{q\mathbf{x}}$$

and H is a gaussian martingale, with

$$(6.6.2) \quad \alpha(t) = \text{Var}(H(t)) = \sum_{q, \mathbf{x}} \int_{(0,t]} (\mathbf{k}_{q\mathbf{x}} - \mathbf{p}_i(\mathbf{k})_{q\mathbf{x}})^2 y_{\mathbf{x}} \exp(\mathbf{b}_q \mathbf{x}) d\Lambda_q.$$

We estimate  $\alpha(t)$  by making the following replacements in the last formula:

Quantity	Replacement (empirical estimator)
$k_{q\mathbf{x}}$	$K_{q\mathbf{x}}$

$y_{\xi}(s)$	$\frac{Y(s)}{n}$
$\beta_q$	$\hat{\mathbf{b}}_q(t)$
$d\Lambda_q(s)$	$\frac{dN_q(s)}{\sum_n Y_n(s) / n \exp(\hat{\mathbf{b}}_q(t)\mathbf{n})}$

The resulting estimator  $\hat{\mathbf{a}}(t)$  of  $\text{Var}(H(t))$  is the residual sum of (weighted) squares in a standard linear regression problem that we describe next.

iii) linear regression problem to compute  $\hat{\mathbf{a}}(t)$ .

a) Fix an event time  $t \geq t_0$  ( $t_0$  as in 6.6). Find  $\hat{\mathbf{b}}_q(t)$  for  $q = 1, \dots, Q$ .

b) Consider the linear regression problem:  $\underline{\mathbf{a}} = \sum_j c_j \underline{\boldsymbol{\varphi}}^j + \underline{\boldsymbol{\varepsilon}}$ , where the row-vectors  $\underline{\mathbf{a}}, \underline{\boldsymbol{\varphi}}^j, \underline{\boldsymbol{\varepsilon}}$  have components indexed by  $(q, \mathbf{x}, s)$  with  $q$  a failure mode,  $\mathbf{x}$  a risk state, and  $s \in (0, t]$  a failure time. The components of  $\underline{\mathbf{a}}$  are:

$$a_{q\mathbf{x}s} = \ln(Y_{\mathbf{x}}(s) \exp(\hat{\mathbf{b}}_q(t)\mathbf{x})).$$

The explanatory variables  $\{\underline{\boldsymbol{\varphi}}^j\}$  are:  $\underline{\mathbf{x}}^i$  ( $i = 1, \dots, d$ ) and  $\underline{\mathbf{b}}^{\text{ru}}$  ( $r = 1, \dots, Q; u \in (0, t]$  a failure time) and have components:

$$x_{q\mathbf{x}s}^i = \mathbf{x}_i$$

$$b_{q\mathbf{x}s}^{\text{ru}} = \mathbf{d}_{(ru),(qs)} = \begin{cases} 1 & \text{if } (r,u) = (q,s) \\ 0 & \text{otherwise} \end{cases}$$

The coefficients  $c_j$  are found by minimizing

$$\sum_{q,\mathbf{x},s} e^{2q\mathbf{x}s} \left( \frac{Y_{\mathbf{x}}(s) e^{\hat{\mathbf{b}}_q(t)\mathbf{x}}}{\sum_n Y_n(s) e^{\hat{\mathbf{b}}_q(t)\mathbf{n}}} \right)^2,$$

and  $\hat{\mathbf{a}}(t)$  is the minimum value of this sum of squares.

iv)  $Z$  is obtained when  $K_{q\mathbf{x}}(s) = \ln(Y_{\mathbf{x}}(s) \exp(\hat{\mathbf{b}}_q(s)\mathbf{x}))$ . If  $n$  is large, under the null hypothesis, the graph  $(t_0, T] ? t ? (\hat{\mathbf{a}}(t), Z(t) / \sqrt{n})$  approaches the graph of a standard brownian motion on  $[0, \alpha(T)]$ . Tests are based on this fact.

**7 The issue of power.** We look at the power of the test against a fixed alternative, i.e. we fix  $\{R_q(., \mathbf{x})\}_{q,\xi}$  and look at the large sample form of  $Z$ . If the leading part  $Z$ , say  $nU$ , is not identically 0, then  $Z / \sqrt{n}$  will tend to  $\infty$  at some point and such a large deviation from 0 can be detected with any reasonable test. The question is whether  $U$  can be identically 0 when  $dR_q(., \mathbf{x})$  does not have the form prescribe by the null

hypothesis.  $U$  depends on the large sample proportion  $y_{\mathbf{x}}$  of survivors, which in turn depends not only on the failure rate, but also on the censoring rate. Failure and censoring are confounded in  $U$ , thus it may happen that both rates combine in such way that  $U$  vanishes identically without the failure rates in each given mode  $q$  being proportional. In this section we propose a way of supplementing the information about the failure rates contained in  $U$  to be able to distinguish the Cox model from any alternative. As earlier, we are not seeking the most general set of assumptions, we are after the "geometric" picture, and so we are justified in making the following simplifying assumption:

(7.0.1) The failure rates  $R_q(., \mathbf{x})$  are absolutely continuous with respect to Lebesgue measure:

$$dR_q(s, \mathbf{x}) = \mathbf{j}_{q\mathbf{x}}(s) ds.$$

Moreover,  $\sum_{\mathbf{x}} y_{\mathbf{x}}(s) \varphi_{q\mathbf{x}}(s) > 0$  for all  $q$ .

Fix a constant  $\gamma$ , and consider the following weighted residual processes  $e_{q\mathbf{x}}$ :

For  $q = 1, \dots, Q; \mathbf{x} \in \mathbf{Z}; t \in (t_0, T]$ ,

(7.0.2) if  $t > t_0$  then

$$e_{q\mathbf{x}}(t) = \frac{1}{n} \int_{(t_0, t]} e^{\mathbf{g}s} \left( dN_{q\mathbf{x}}(s) - \frac{Y_{\mathbf{x}}(s) e^{\hat{\mathbf{b}}_q(t)\mathbf{x}}}{\sum_n Y_n(s) e^{\hat{\mathbf{b}}_q(t)\mathbf{n}}} dN_q(s) \right)$$

and if  $t \leq t_0$  then  $e_{q\mathbf{x}}(t) = 0$ .

We recognize  $n e_{q\mathbf{x}}$  as a special case of (6.7.5). In particular,

$$e_{q\mathbf{x}}(t) \rightarrow U_{q\mathbf{x}}(t) =$$

$$\int_{(0, t]} \left[ e^{\mathbf{g}s} y_{\mathbf{x}}(s) \varphi_{q\mathbf{x}}(s) - \frac{e^{\mathbf{g}s} y_{\mathbf{x}}(s) e^{b_q^0(t)\mathbf{x}}}{\sum_n y_n(s) e^{b_q^0(t)\mathbf{n}}} \left( \sum_n y_n(s) \varphi_{q\mathbf{x}}(s) \right) \right] ds$$

uniformly in  $[0, T]$  a.s..

**Note:** Instead of  $\exp(\gamma s)$ , any strictly monotone function can be used as weight. For the proof of the following theorem to work, for a given  $q$ , all  $e_{p\mathbf{x}}$  should use the same weight function.

**7.1 Theorem.** Assume (7.0.1), (6.3.1-3).

i) if the null hypothesis holds, for all  $q\mathbf{x}$   $U_{q\mathbf{x}}$  vanishes identically

ii) if for some  $\mathbf{g} \neq 0$ , for all  $q\mathbf{x}$   $U_{q\mathbf{x}}$  vanishes identically, then the null hypothesis holds.

**Proof:** will be given elsewhere.

## 8. Acknowledgements.

Part of this research was carried while the author received support from the government of Canada through a MITACS grant. The author is grateful for that support.

The author expresses his gratitude to John Hsieh for many helpful suggestions. The author is accountable for any errors, oversights, etc. in this work.

## Appendix 1. A function written in Splus to test for constant hazard using the sequential estimation of the rate.

```
sequential.test.Poisson_function(failT)
{
# failT is the vector of inter-failure times,
# it is assume that length(failT) > 1,
# there is no provision for censoring.
#
  n <- length(failT)
  failT <- sort(failT)
  Y <- n:1
#-----
# Brake ties, find Dt = inter-failure times
# we use Dt as an auxiliary variable until the definitive
assignment
#
  Dt <- failT - c(0, failT[1:(n - 1)])
  eps <- min(Dt[Dt > 0])
  Dt <- failT
  count <- 1
  while(count <= 5 & length(unique(Dt)) != n) {
    count <- count + 1
    Dt <- failT + eps * 0.01 * (runif(n) - 0.5)
  }
  if(count == 6) {
    stop(message = " Brake ties manually")
  }
  Dt <- Dt - c(0, Dt[1:(n - 1)])
#-----
#Total time on test:
# if N(t) = j, t a failure time,
# Integral Y ds over [0,t] = sum Dt[i]*Y[i] for i = 1,...,j
#
  ToTi <- cumsum(Dt * Y)
#-----
# Non-parametric estimator of the log-likelihood process:
# Ll = - Integral (ln(Y) + 1) dN over (0,t)
#
  el <- - cumsum(log(Y) + 1)
#-----
# Parametric estimator of the log-likelihood process:
# Ll0 = N*{-1 + ln(N/(Total time on test))}
# sequential estimator of l: N(t) / TTT(t)
```

```
#
  el0 <- 1:n * (log(1:n/ToTi) - 1)
#-----
# The test statistic:
#
  H <- - cumsum(log(Y/n)) + 1:n/ToTi *
cumsum(log(Y/n) * Y * Dt)
  H <- H/sqrt(n)
#-----
# Estimator of alpha = Var(H):
#
  alpha <- (1:n/ToTi * 1)/n *
(cumsum((log(Y))^2 * Y * Dt) -
(cumsum(log(Y) * Y * Dt))^2/ToTi)
#
# assume Y/n converges in probability to y
# Fact: H converges weakly in Skorohod space D[0,T] to
B(alpha)
# where B is a standard brownian motion on the line, and
# alpha = lambda*Integral {ln(y) - pj_t}^2y ds over (0,t]
# with pj_t = orthogonal projection of ln(y) in L^2([0,t],
lambda y ds) in the direction of 1
#-----
#
  list(H = H, alpha = alpha)
}
```

## Appendix 2. Confidence regions for H.

To assess how extreme an observation is, we can use statistics of the form  $I = \sup_{0 < t < T} \frac{|H(t)|}{w(\mathbf{a}(t))}$  [Fleming, Harrington], section 6.3). The asymptotic distribution of  $\lambda$  coincides with the distribution of  $\sup_{0 < u < \mathbf{a}(T)} \frac{|B(u)|}{w(u)}$  (where  $B$  is a standard brownian motion). For arbitrary  $w$ , this distribution can be estimated through simulations. For special values of  $w$  the distribution can be given in closed form. For example when  $w(t) = 1$  for all  $t$ , we have:

$$P\left(\sup_{0 < u \leq \mathbf{a}(T)} |B(u)| \geq I\right) = 1 - \frac{4}{p} \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \exp\left(-\frac{p^2(2k+1)^2}{8I^2}\right).$$

In particular,

$$(A.2.1) \quad p = 1 - \frac{4}{p} \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \exp\left(-\frac{p^2(2k+1)^2}{8I^2}\right) \Bigg|_{I = \sup_{0 \leq t} |H(t)|}$$

has a distribution that is asymptotically uniform [0,1], with small values of  $p$  corresponding to deviations from the null hypothesis.

An alternative is to define confidence regions as follows. Consider the family of curves

$$h_a(u) = u \frac{\ln(e^{\frac{a^2}{2u}} + \sqrt{e^{\frac{a^2}{u}} - 1})}{a} \quad (u \geq 0, a > 0).$$

The following facts hold:

- i) for each  $u \in [0, \infty)$ ,  $h_a(u)$  increases with  $a$ ;
- ii) for each  $u \in [0, \infty)$ ,

$$\lim_{a \rightarrow 0^+} h_a(u) = \sqrt{u} \quad \text{and} \quad \lim_{a \rightarrow \infty} h_a(u) = \infty;$$

- iii) for each  $a > 0$ ,

$$h_a(u) = \frac{a}{2} + \frac{\ln(2)}{a}u + \frac{O(e^{-\frac{a^2}{2u}})}{u} \quad \text{as } u \rightarrow 0$$

and

$$h_a(u) = \sqrt{u} + \frac{a^2}{12} \frac{1}{\sqrt{u}} + O\left(\frac{1}{u}\right) \quad \text{as } u \rightarrow \infty.$$

- iv) for latter reference, note that  $h_a(u) = \sqrt{uy} \left(\frac{a}{\sqrt{u}}\right)$ , where

$$y(z) = \frac{z}{2} + \frac{\ln\left(1 + \sqrt{1 - e^{-z^2}}\right)}{z}$$

is an increasing function of  $z$ , with range  $[1, \infty)$ .

Given a process  $\{X(u)\}_{0 \leq u \leq \alpha}$  with continuous paths, where  $0 < \alpha < \infty$ , let

$$a^* = \inf\{a \mid |X(u)| \leq h_a(u) \text{ for all } u \in [0, \alpha]\}$$

i.e., the region  $\{(u, v) \mid |u| \leq h_{a^*}(v), u \in [0, \alpha]\}$  is the tightest region bounded by the graphs of  $\pm h_a$  that contains the given path  $X$ . An equivalent formula for  $a^*$ , which we use for computations, is:

$$a^* = \sup\left\{ \sqrt{uy}^{-1} \left( \frac{|X(u)|}{\sqrt{u}} \vee 1 \right) \mid u \in (0, \alpha) \right\}$$

**Lemma.** Let  $W$  be a standard brownian motion defined on the interval  $[0, s]$  ( $0 < s < \infty$ ), let  $a > 0$ . Then:

$$P(|W_u| \geq h_a(u) \text{ for some } u \in (0, s]) =$$

$$2 \left( 1 - \Phi\left(\frac{h_a(s)}{\sqrt{s}}\right) \right) + \left( \Phi\left(\frac{a+h_a(s)}{\sqrt{s}}\right) - \Phi\left(\frac{a-h_a(s)}{\sqrt{s}}\right) \right).$$

**Proof.** The proof of this lemma is a lengthy computation based on well known facts about brownian motions. We refer to [Karatzas, Shreve], Chapter 4, section 4.3.C for details. The idea is to use the following non-negative solution of the backward heat equation:

$$v(t, x) = e^{-\frac{ta^2}{2}} \cosh(xa). \quad \text{The definition of } h_a \text{ is such that } |W_u| \leq h_a(u) \text{ if and only if } v(t, tW_1) \geq 1 \text{ with } t = 1/u.$$

But  $W_t^* = tW_1$  is a standard brownian motion,

$Z_t = v(t, tW_1)$  is a non-negative martingale, and  $Z_t \rightarrow 0$  in

probability as  $t \rightarrow \infty$ . To get the final result, use [Karatzas, Shreve], problem 3.28 and the fact that  $W_{1/s}^*$  has a normal distribution.

To assess how extreme an observation of  $H_n$  is, we evaluate

$$a_n^* = \sup_{0 < t \leq T} \sqrt{a(t)} y^{-1} \left( \frac{|H_n(t)|}{\sqrt{a(t)}} \vee 1 \right).$$

Weak convergence of  $\{H_n\}_n$  imply weak convergence of  $\{a_n^*\}_n$ , thus, we approximate the distribution of  $a_n^*$  with the distribution of  $a^*$ . We use (A.2.2)

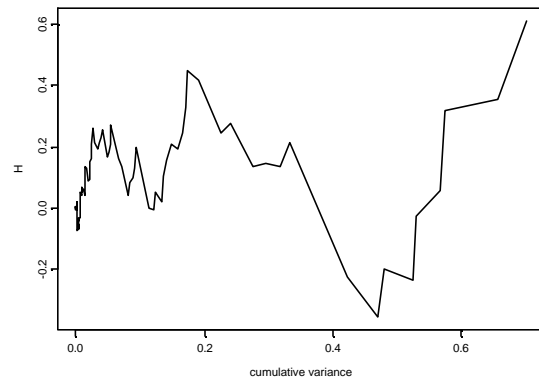
$$p^* = 2 \left( 1 - \Phi\left(\frac{h_a(s)}{\sqrt{s}}\right) \right) + \left( \Phi\left(\frac{a+h_a(s)}{\sqrt{s}}\right) - \Phi\left(\frac{a-h_a(s)}{\sqrt{s}}\right) \right)$$

with  $a = a_n^*$ , and  $s = \alpha_n(T)$  as p-value.

**Example:** A simulation using the sequential.test.Poisson function given in Appendix 1:

```
> set.seed(28)
> X2_rexp(100)
> test2_sequential.test.Poisson(X2)
> plot(test2$alpha, test2$H, type='l', ylab='H',
xlab='cumulative variance')
```

Here is the graph of the sequential residual process:



For this simulation,  
p-value using (A.2.1) = 0.8713729  
p-value using (A.2.2) = 0.9368366 .

**References:**

- Andersen, P. K. ; Borgan, Ø. ; Gill, R. D. , Keiding, N. : *Statistical models based on counting processes*, 1993, Springer-Verlag.
- Durbin, J. Weak convergence of the sample distribution function when parameters are estimated, *Ann. Stat.*, vol 1(1973), No. 2, 279-290
- Ferrenberg, A. M.; Landau, D. P.; Wong, Y. J. Monte Carlo simulations: hidden errors from "good" random number generators, *Phys. rev. lett.*, vol 69, no 23, December 1992.
- Fleming, T. R.; Harrington, D. P. *Counting processes and survival analysis*, 1991, Wiley
- Hida, Takeyuki ; Hitsuda, Masuyuki "*Gaussian Processes*", Translation of mathematical monographs, v. 120, 1993, American Mathematical Society.
- Hjort, N. L. Goodness of fit tests in models for life history data based on cumulative hazard rates, *Ann. Stat.*, vol 18(1990), No. 3, 1221-1258
- Karatzas, I.; Shreve, S. E. *Brownian motion and stochastic calculus*, 1988, Springer-Verlag.
- Khmaladze, E. V. Martingale approach in the theory of goodness-of-fit tests, *Theory Prob. Appl.*, 26 (1981), no 2, pp 240-257.
- Khmaladze, E. V. Goodness of fit problems and scanning innovation martingales, *Ann. Stat.*, v. 21 (1993), no. 2, pp 798-829
- Li, G.; Sun, Y. A simulation-based goodness-of-fit test for survival data, *Stat. Prob. Lett.* (2000), vol. 47, pp. 403-410
- Lin, Chin-Yu ; Kosorok, Michael R. A general class of function-indexed non-parametric tests for survival analysis, *Ann. Stat.*, v. 27 (1999), no. 5, pp. 1722-1744
- Lin, D. Y.; Fleming, T. R.; Wei, L. J. Confidence bands for survival curves under the proportional hazards model, *Biometrika* (1994), vol 81, no 1, pp 73-81.
- Lin, D. Y. ; Wei, L. J. ; Ying, Z. Checking the Cox model with cumulative sums of martingale-based residuals, *Biometrika* (1993), vol. 80, no. 3, pp. 557-72.
- Zurkowski, V. D. to appear.