DATA-DRIVEN SMOOTH TESTS OF THE PROPORTIONAL HAZARDS ASSUMPTION

DAVID KRAUS

Institute of Information Theory and Automation, Prague

&

Charles University in Prague, Department of Statistics

ABSTRACT. A new test of the proportional hazards assumption in the Cox model is proposed. The idea is based on Neyman's smooth tests. The Cox model with proportional hazards (i.e. time-constant covariate effects) is embedded in a model with a smoothly time-varying covariate effect that is expressed as a combination of some basis functions (e.g., Legendre polynomials, cosines). Then the smooth test is the score test for significance of these artificial covariates. Furthermore, we apply a modification of Schwarz's selection rule to choosing the dimension of the smooth model (the number of the basis functions). The score test is then used in the selected model. In a simulation study, we compare the proposed tests with standard tests based on the score process.

1. INTRODUCTION

We consider the Cox proportional hazards regression model (Cox, 1972) in the counting process formulation of Andersen and Gill (1982) (see also Andersen, Borgan, Gill and Keiding, 1993)

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{\beta^\mathsf{T} Z_i(t)\}.$$

Here $\lambda_i(t)$ is the intensity process of the i-th component of an n-variate counting process $N(t) = (N_1(t), \ldots, N_n(t))^\mathsf{T}$, $t \in [0, \tau]$, $Y_i(t)$ denotes the risk indicator process, $Z_i(t)$ is a p-dimensional covariate (predictable process), $\lambda_0(t)$ stands for an unknown function (baseline hazard) and β is a vector of unknown regression coefficients. Throughout this paper, we assume that the conditions of Andersen and Gill (1982) guaranteeing certain asymptotic

Key words and phrases. Cox model, Neyman's smooth test, proportional hazards assumption, Schwarz's selection rule.

Correspondence address. Institute of Information Theory and Automation, Pod Vodárenskou věží 4, CZ-182 08 Prague 8, Czechia; kraus@karlin.mff.cuni.cz; http://www.davidkraus.net/.

properties are satisfied. For simplicity, the time period is assumed to be finite (i.e., $\tau < \infty$); we refer to Andersen and Gill (1982, Section 4) for an extension to the whole line (see also Fleming and Harrington, 1991, Section 8.4).

The crucial assumption in the Cox model is proportionality of the effects of the covariates. This means that the hazard ratio for two individuals does not depend on time, or, when the covariates are time-dependent, it depends on time solely through the values of the covariates. The proportional hazards assumption can be violated in many ways. One is when some of the coefficients $\beta_1, \ldots, \beta_p$ vary with time. Another situation is when the regression model is misspecified (the true model can be, e.g., Aalen's additive regression) or when the supposed stochastic structure is incorrect (the counting processes can actually be, for instance, renewal processes) etc.

In the present paper, our aim is to test the proportional hazards assumption for the p-th (say) covariate against the alternative of time-varying coefficient $\beta_p(t)$. Various methods for detecting nonproportional hazards have been developed.

The most important (and most often used) inference tool is the score process

$$U_1(t;\hat{\beta}) = \sum_{i=1}^{n} \int_0^t Z_i(s) dN_i(s) - \int_0^t \frac{\sum_{i=1}^n Y_i(s) Z_i(s) \exp\{\hat{\beta}^\mathsf{T} Z_i(s)\}}{\sum_{i=1}^n Y_i(s) \exp\{\hat{\beta}^\mathsf{T} Z_i(s)\}} d\bar{N}(s)$$

(where $\bar{N} = \sum_{i=1}^n N_i$). Each of its p components reflects deviations from proportionality of the respective covariate. Lin, Wei and Ying (1993) use tests of the Kolmogorov–Smirnov type based both on components of the score process (for testing effects of individual covariates) and on the whole vector of processes (overall assessment of fit). Other functionals of the components, namely those leading to the test of the Anderson–Darling and Cramér–von Mises type, are studied by Kvaløy and Neef (2004).

The test based on the score process is a test of time-constancy of the effect β_p against a general unspecified alternative of time-varying $\beta_p(t)$. Another approach is to test against specific departures from proportionality (Cox, 1972; Andersen et al., 1993, Sec. VII.3.3). Recall that we wish to test whether the effect of the p-th covariate is constant. Then we may include a new time-dependent covariate $g(t)Z_{ip}(t)$ (with $g(t)$ being a nonrandom function) into the model as follows

$$\lambda_i(t) = Y_i(t)\lambda_0(t) \exp\{\beta^\mathsf{T} Z_i(t) + \gamma g(t) Z_{ip}(t)\},$$

and test its significance ($\gamma = 0$ against $\gamma \neq 0$) by standard (partial likelihood based) methods. Some frequent choices are $g(t) = t$ or $g(t) = \log t$.

A compromise between the two classic tests (global and directional) is represented by Neyman's smooth tests, which are the theme of this paper. The idea consists of testing the null hypothesis against an alternative with a smoothly time-varying coefficient for the covariate $Z_{ip}(t)$. This means that under the alternative the effect of the covariate $Z_{ip}(t)$ can be expressed as a combination of several (say k) smooth functions $\psi_1(t), \ldots, \psi_k(t)$ (and an intercept, of course). (The choice of the smooth functions is discussed later on.) Thus, we consider an alternative Cox model with k time-dependent covariates in the form

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{\beta^\mathsf{T} Z_i(t) + \theta^\mathsf{T}\psi(t)Z_{ip}(t)\}$$

and test significance of the covariates $\psi(t)Z_{ip}(t)$ (here $\psi(t) = (\psi_1(t), \ldots, \psi_k(t))^\mathsf{T}$). Explicitly, our smooth test is the score test of

$$H_0 : \theta = 0 \quad \text{against} \quad H_1 : \theta \neq 0.$$

Next, we address the issue of choosing k, the dimension of the smooth alternative. We follow the idea of data-driven smooth tests that is due to Ledwina and coauthors; see, for instance, Inglot, Kallenberg and Ledwina (1997), Kallenberg and Ledwina (1997) and references therein. In their situation of testing goodness of fit of a parametric family, they consider models with dimensions $1, \ldots, d$ (for a chosen integer d) and use a modification of Schwarz's selection rule for selecting one of them. The test is then based on the score statistic for the selected likely model. A similar approach is applied in our situation.

Before closing the introductory section, we must mention a completely different approach to testing proportionality that was proposed by Martinussen, Scheike and Skovgaard (2002) (see also Scheike and Martinussen, 2004). They consider an extended Cox model with time-varying coefficients. Their test is a test of possibility of reduction of a nonparametric time-varying Cox model to a semiparametric model with some effects being constant in time.

The structure of the paper is as follows. In Section 2 we develop the smooth test of proportionality and establish asymptotic properties of the test statistic. Section 3 deals with the data-driven version of the test based on Schwarz's selection rule. In Section 4, our tests are compared through simulations with tests based on the score process in various situations. Results are summarised in Section 5, which closes the paper.

2. Smooth tests

As mentioned in Introduction, the null model

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{\beta^{\mathsf{T}} Z_i(t)\} \tag{1}$$

is embedded in

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{\beta^{\mathsf{T}} Z_i(t) + \theta^{\mathsf{T}}\xi_i(t)\}, \tag{2}$$

where

$$\xi_i(t) = \psi(t)Z_{ip}(t), \qquad \psi(t) = (\psi_1(t), \ldots, \psi_k(t))^{\mathsf{T}}.$$

The functions representing smooth alternatives are chosen as some basis functions in transformed (standardised, uniformised) time, i.e. in the form

$$\psi_j(t) = \varphi_j(\Lambda_0(t)/\Lambda_0(\tau)), \qquad j = 1, \ldots, k \tag{3}$$

or

$$\psi_j(t) = \varphi_j(F_0(t)/F_0(\tau)), \qquad j = 1, \ldots, k. \tag{4}$$

Here $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ is the cumulative baseline hazard function and $F_0(t) = 1-\exp\{-\Lambda_0(t)\}$ the corresponding distribution function. The smooth functions φ_j are some bounded functions in $L_2[0,1]$ such that $\{1, \varphi_1, \ldots, \varphi_k\}$ is a set of linearly independent functions. Most popular examples are the orthonormal Legendre polynomials on $[0,1]$ and the cosine basis $\varphi_j(u) = \sqrt{2}\cos(\pi j u)$. There are many other possibilities, such as various spline bases, cell indicators, $\varphi_j(u) = u^j$, or other right-continuous functions with left-hand limits. For a discussion of the choice of the basis functions see, for instance, Inglot et al. (1997, p. 1227) in the traditional goodness-of-fit framework, or Peña (1998a,b) for hazard based models.

Before developing the score test of $\theta = 0$, we need to introduce the following basic notation. Let us denote

$$S^{(j,k)}(t;\beta,\theta) = \sum_{i=1}^{n} Y_i(t)Z_i(t)^{\otimes j}\xi_i(t)^{\otimes k}\exp\{\beta^{\mathsf{T}} Z_i(t) + \theta^{\mathsf{T}}\xi_i(t)\}$$

for $j = 0, 1, 2$, $k = 0, 1, 2$, $j + k \leq 2$. As we are mainly interested in situations with $\theta = 0$, we simplify the notation and use $S^{(j,k)}(t;\beta) = S^{(j,k)}(t;\beta,0)$. The same applies to other functions (processes) introduced later: whenever θ is dropped, it means that the function is evaluated

at $\theta = 0$. Furthermore, we set $S^{(j)} = S^{(j,0)}$ (this notation agrees with that introduced by Andersen and Gill, 1982).

Denote

$$C(t;\beta,\theta) = \sum_{i=1}^{n} \int_0^t [\beta^\mathsf{T} Z_i(s) + \theta^\mathsf{T} \xi_i(s)] dN_i(s) - \int_0^t \log\{S^{(0)}(s;\beta,\theta)\} d\bar{N}(s),$$

the logarithm of the partial likelihood in the k-dimensional model (2). Then $C(\tau;\beta) := C(\tau;\beta,0)$ is the log partial likelihood for the Cox model (1). The score process for this model is

$$U_1(t;\beta) = \frac{\partial}{\partial\beta} C(t;\beta) = \sum_{i=1}^{n} \int_0^t Z_i(s) dN_i(s) - \int_0^t \frac{S^{(1)}(s;\beta)}{S^{(0)}(s;\beta)} d\bar{N}(s).$$

The estimate $\hat{\beta}$ defined as the solution to

$$U_1(\tau;\beta) = 0$$

is the maximum partial likelihood estimate in the null model (1) (or the restricted maximum partial likelihood estimate in (2) under $\theta = 0$). The score process for θ in the model (2) is

$$U_2(t;\beta,\theta) = \frac{\partial}{\partial\theta} C(t;\beta,\theta) = \sum_{i=1}^{n} \int_0^t \xi_i(s) dN_i(s) - \int_0^t \frac{S^{(0,1)}(s;\beta,\theta)}{S^{(0)}(s;\beta,\theta)} d\bar{N}(s).$$

The score test for $\theta = 0$ is based on the quantity $U_2(\tau;\hat{\beta}) := U_2(\tau;\hat{\beta},0)$. Asymptotic properties of the score test in the Cox model are well-known (Andersen and Gill, 1982): the score $U_2(\tau;\hat{\beta})$ turns out to be asymptotically normal.

We need to investigate its asymptotic variance in order to be able to form a quadratic χ^2 statistic. By Taylor's expansion around the true value β_0, $U_2(\tau;\hat{\beta})$ may be written as

$$U_2(\tau;\hat{\beta}) = U_2(\tau;\beta_0) - D(\tau;\beta^*)(\hat{\beta} - \beta_0), \tag{5}$$

where $D(t;\beta) = -\frac{\partial}{\partial\beta^\mathsf{T}} U_2(t;\beta)$ and β^* lies on the line segment between β_0 and $\hat{\beta}$. Next we may use the identity $\hat{\beta} - \beta_0 = J(\tau;\tilde{\beta})^- U_1(\tau;\beta_0)$, which follows from Taylor's expansion $U_1(\tau;\hat{\beta}) - U_1(\tau;\beta_0) = -J(\tau;\tilde{\beta})(\hat{\beta} - \beta_0)$ and the fact $U_1(\tau;\hat{\beta}) = 0$; here $J(\tau;\beta) = -\frac{\partial}{\partial\beta^\mathsf{T}} U_1(\tau;\beta)$ stands for the information matrix and $\tilde{\beta}$ is again on the line segment between β_0 and $\hat{\beta}$. Inserting this into (5) we obtain

$$n^{-1/2} U_2(\tau;\hat{\beta}) = n^{-1/2} U_2(\tau;\beta_0) - \{n^{-1} D(\tau;\beta^*)\}\{n J(\tau;\tilde{\beta})^-\}\{n^{-1/2} U_1(\tau;\beta_0)\}. \tag{6}$$

Consequently, the key step is to study weak convergence of the martingale $n^{-1/2} U(t;\beta_0) = n^{-1/2}(U_1(t;\beta_0), U_2(t;\beta_0))^\mathsf{T}$ and convergence in probability of the other quantities in (6). It

may be shown that $n^{-1/2}U(t;\beta_0)$ converges weakly to a continuous zero-mean Gaussian martingale with covariance matrix denoted

$$\sigma(t;\beta_0) = \begin{pmatrix} \sigma_{11}(t;\beta_0) & \sigma_{12}(t;\beta_0) \\ \sigma_{21}(t;\beta_0) & \sigma_{22}(t;\beta_0) \end{pmatrix}.$$

Besides, the matrices $n^{-1}D(\tau;\beta^*)$ and $n^{-1}J(\tau;\tilde{\beta})$ converge in probability to $\sigma_{21}(\tau;\beta_0)$ and $\sigma_{11}(\tau;\beta_0)$, respectively. Therefore, $n^{-1/2}U_2(\tau;\hat{\beta})$ is asymptotically normal with zero mean and variance

$$v(\tau;\beta_0) = \sigma_{22}(\tau;\beta_0) - \sigma_{21}(\tau;\beta_0)\sigma_{11}(\tau;\beta_0)^{-1}\sigma_{12}(\tau;\beta_0).$$

Let

$$V(\tau;\hat{\beta}) = \Sigma_{22}(\tau;\hat{\beta}) - \Sigma_{21}(\tau;\hat{\beta})\Sigma_{11}(\tau;\hat{\beta})^{-}\Sigma_{12}(\tau;\hat{\beta}),$$

where $\frac{1}{n}\Sigma(\tau;\hat{\beta})$ (with corresponding submatrices) is a consistent estimator of $\sigma(\tau;\beta_0)$. Finally, the score statistic for testing $\theta = 0$ is

$$T_k = U_2(\tau;\hat{\beta})^{\mathsf{T}} V(\tau;\hat{\beta})^{-} U_2(\tau;\hat{\beta}), \tag{7}$$

which is asymptotically χ^2_k-distributed as $n \to \infty$. Obviously, the null hypothesis is rejected if T_k is significantly large. The number of degrees of freedom equals the rank of the limiting covariance matrix which is k by the assumptions of Andersen and Gill (1982) and by linear independence of the basis functions, see also Andersen et al. (1993, p. 503).

The estimator $\frac{1}{n}\Sigma(\tau;\hat{\beta})$ of $\sigma(\tau;\beta_0)$ is obtained by plugging $\hat{\beta}$ into the quadratic variation of $n^{-1/2}U(\cdot;\beta_0)$. Explicitly,

$$\begin{aligned}
\Sigma_{11}(t;\beta) &= [U_1(\cdot;\beta)](t) = \sum_{i=1}^{n} \int_0^t \left[Z_i(s) - \frac{S^{(1)}(s;\beta)}{S^{(0)}(s;\beta)} \right]^{\otimes 2} dN_i(s), \\
\Sigma_{22}(t;\beta) &= [U_2(\cdot;\beta)](t) = \sum_{i=1}^{n} \int_0^t \left[\xi_i(s) - \frac{S^{(0,1)}(s;\beta)}{S^{(0)}(s;\beta)} \right]^{\otimes 2} dN_i(s), \\
\Sigma_{21}(t;\beta) &= [U_2(\cdot;\beta), U_1(\cdot;\beta)](t) \\
&= \sum_{i=1}^{n} \int_0^t \left[\xi_i(s) - \frac{S^{(0,1)}(s;\beta)}{S^{(0)}(s;\beta)} \right] \left[Z_i(s) - \frac{S^{(1)}(s;\beta)}{S^{(0)}(s;\beta)} \right]^{\mathsf{T}} dN_i(s).
\end{aligned}$$

The score test statistic T_k can be easily computed in existing software (like the `survival` package in R).

The time transformation in the smooth functions $\psi_j(t)$ (in (3) or (4)) depends on the unknown cumulative baseline hazard function $\Lambda_0(t)$. In practice, we have to estimate it. The Breslow estimator is

$$\hat{\Lambda}_0(t) = \int_0^t \frac{d\bar{N}(s)}{S^{(0)}(s;\hat{\beta})}.$$

By uniform consistency of this estimator it follows that the weak limit of the score is the same as if we knew Λ_0.

Which transformation ((3) or (4)) should we use? For survival data (i.e., for counting processes with at most one jump) I prefer the transformation (4) based on the baseline distribution function F_0. The reason is as follows. If we use the transformation (3), periods with highly increasing $\Lambda_0(t)$ (i.e., high $\lambda_0(t)$) are mapped to larger periods in $[0,1]$ than periods with moderate increase of $\Lambda_0(t)$. This is reasonable, and it is the purpose of the time transformations. However, if such a period with high $\lambda_0(t)$ occurs late on the time line (where 'late' means that the cumulative intensity $\Lambda_0(t)$ is large, i.e., there is only a small probability of surviving so long), then the actual proportion of observations in such a period will be much lower than the proportion of the corresponding period in $[0,1]$. In other words, late periods with high $\lambda_0(t)$ may be overrepresented in the domain of the smooth functions. Moreover, a typical feature of the Breslow estimator is that it has several large jumps at the end, and thus again the end of the time period may receive much larger weight in $[0,1]$ than is adequate. Consequently, the shape of the smooth functions may not be fully exploited with the time transformation (3), and it is better to use (4). On the other hand, however, if the data consist of repeated events (such as observations of (possibly nonhomogeneous) Poisson processes), one may consider using the transformation (3) because the intensity Λ_0 is a more proper characteristic of the stochastic structure than the distribution function F_0.

We close this section by a practical comment. If the covariates are time independent, it is suitable to compute the baseline distribution at the covariate means. It then describes the behaviour of a typical observation.

3. Data-driven version of the test

Smooth tests presented up to now were score tests of $\theta = 0$ against $\theta \neq 0$ in the k-dimensional model (2), where k was fixed (chosen prior to testing). Simulations (reported in Section 4) show that the proper choice of k plays an important role. If we choose k too

large, we test against a superfluously complex alternative. It contains redundant covariates which do not contribute to the test statistic markedly but increase the number of degrees of freedom (and, hence, critical values). This causes a loss of power.

The idea of data-driven tests consists of choosing out of d alternative models (with increasing dimensions) one that describes the data well but is not too large. Then the smooth test is performed in this model.

The idea dates back to Ledwina (1994) who applied the Bayesian information criterion (BIC, Schwarz's selection rule) to the task of testing uniformity (or other single distribution). Later, Inglot et al. (1997) and Kallenberg and Ledwina (1997) extended this method to composite hypotheses. In the Cox model, Abrahamowicz, MacKenzie and Esdaile (1996) employed the Akaike information criterion (AIC) for choosing the dimension. However, they did not investigate asymptotic distribution of the test statistic when the dimension was selected by the AIC. In Peña (2003), a modification of Schwarz's rule was considered in a different hazard based model.

Let d be the maximal dimension of the alternative model. The considered models are

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{\beta^{\mathsf{T}} Z_i(t) + \theta_1\xi_{i1}(t) + \cdots + \theta_k\xi_{ik}(t)\}, \qquad k = 1, \ldots, d.$$

Schwarz's rule in its traditional form selects among the d models the one whose penalised (partial) log-likelihood is largest. The log partial likelihood is penalised by subtracting $\frac{k}{2}\log n$. Since the rule based on the partial likelihood requires optimisation of the partial likelihood function for all d models, it may be computationally inconvenient. Instead, we will use a modified rule based on the score statistic. Let T_k be the score statistic defined in (7) for the k-dimensional alternative. Then the selection rule is

$$S = \underset{k\in\{1,\ldots,d\}}{\arg\max}\{T_k - k\log n\}. \tag{8}$$

The statistic of the data-driven test is T_S.

For a fixed dimension k, we have seen that the statistic T_k of the smooth test is approximately χ^2_k-distributed. Now we find asymptotic distribution of the statistic with dimension selected by Schwarz's rule. The lemma that follows states that under the null the selection rule is asymptotically concentrated in 1, the smallest possible dimension.

Lemma 1. *Under* H_0, $\Pr[S = 1] \xrightarrow[n\to\infty]{} 1$.

Proof. It suffices to show that $\Pr[S = k] \xrightarrow[n\to\infty]{} 0$ for $k = 2, \ldots, d$ (because $\Pr[S = 1] = 1 - \sum_{k=2}^{d} \Pr[S = k]$). This is apparent from

$$\Pr[S = k] \le \Pr[T_k - k \log n \ge T_1 - \log n] = \Pr[T_k / \log n - T_1 / \log n \ge k - 1] \xrightarrow[n\to\infty]{} 0,$$

where the convergence holds because of the weak convergence of T_j to a nondegenerate (χ^2_j-distributed) variable for any j (and, hence, convergence in probability of $T_j / \log n$ to 0). □

Consequently, we have the following result.

Theorem 1. *Under H_0, $T_S \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2_1$.*

Proof. Immediately follows from Lemma 1. □

Kallenberg and Ledwina (1997) point out that the χ^2_1 approximation of the null distribution of T_S is often inaccurate. Typically, when this approximation is used, the test considerably exceeds its prescribed nominal level. The same problem is present in our situation, see simulations in Section 4. Kallenberg and Ledwina (1997, p. 1097) (see also Kallenberg and Ledwina, 1995) derived a much more accurate approximation. Here we adapt their ideas to our setting.

First, we write

$$H(x) = \Pr[T_S \le x] = \Pr[T_1 \le x, S = 1] + \Pr[T_2 \le x, S = 2] + \Pr[T_S \le x, S \ge 3],$$

where the third term on the right-hand side can be neglected under H_0. The event $[S = 1]$ is approximated by $[T_1 - \log n \ge T_2 - 2 \log n] = [T_2 - T_1 \le \log n]$. Similarly, $[S = 2]$ is approximated by $[T_2 - T_1 \ge \log n]$.

Before we proceed, we need to investigate the asymptotic distribution of $(T_1, T_2 - T_1)^\mathsf{T}$. We do it under the assumption that the bases are nested in the sense that the $(k+1)$-dimensional basis contains the k-dimensional one and one more function. The variables T_1, T_2 are functions of the score $U_2(\tau; \hat{\beta})$ that is asymptotically distributed as a bivariate normal vector $(R_1, R_2)^\mathsf{T}$ with variance matrix $v = v(\tau; \beta_0)$. Denote elements of v as $\left(\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right)$ and $\rho = b/\sqrt{ac}$. The distribution $\mathrm{N}(0, v)$ of $(R_1, R_2)^\mathsf{T}$ can be obtained from two independent standard normal variables G_1, G_2: if $\tilde{R}_1 = \sqrt{a}[\sqrt{1-\rho^2} G_1 + \rho G_2]$ and $\tilde{R}_2 = \sqrt{c} G_2$, then $(\tilde{R}_1, \tilde{R}_2) \sim (R_1, R_2)$. Thus, T_1 is asymptotically distributed as $R_1^2/a \sim \tilde{R}_1^2/a = [\sqrt{1-\rho^2} G_1 + \rho G_2]^2 =: T_1^\infty$. Similarly, asymptotic distribution of T_2 is that of $(R_1, R_2) v^{-1} (R_1, R_2)^\mathsf{T} \sim$

$(\tilde{R}_1, \tilde{R}_2)v^{-1}(\tilde{R}_1, \tilde{R}_2)^\mathsf{T} =: T_2^\infty$. Straightforward but tedious computations yield that $T_2^\infty - T_1^\infty = [\rho G_1 - \sqrt{1-\rho^2}G_2]^2$. Finally, since $\rho G_1 + \sqrt{1-\rho^2}G_2$ and $\rho G_1 - \sqrt{1-\rho^2}G_2$ are independent standard normal, we obtain that $(T_1, T_2 - T_1)^\mathsf{T}$ is asymptotically distributed as a vector of two independent χ_1^2 variables.

Now we can study $H(x)$. We will treat $H(x)$ separately for $x \le \log n$, $\log n < x < 2\log n$ and $x \ge 2\log n$.

For $x \le \log n$,

$$\Pr[T_2 \le x, S = 2] \doteq \Pr[T_2 \le x, T_2 - T_1 \ge \log n] = 0,$$

because $T_1 \ge 0$ a.s. Thus

$$H(x) \doteq \Pr[T_1 \le x, T_2 - T_1 \le \log n] \doteq [2\Phi(\sqrt{x}) - 1][2\Phi(\sqrt{\log n}) - 1], \quad x \le \log n.$$

If $x \ge 2\log n$,

$$\Pr[T_2 \le x, S = 2] \doteq \Pr[T_2 \le x, T_2 - T_1 \ge \log n] \doteq \Pr[T_2 - T_1 \ge \log n].$$

Motivation for the latter approximation is as follows. Rewrite

$$\Pr[T_2 \le x, T_2 - T_1 \ge \log n] = \Pr[T_2 - T_1 \ge \log n] - \Pr[T_2 > x, T_2 - T_1 \ge \log n]. \tag{9}$$

As $T_2 - T_1$ is approximately χ_1^2 distributed, we have

$$\Pr[T_2 - T_1 \ge \log n] \doteq 2(1 - \Phi(\sqrt{\log n})) \doteq 2\frac{\varphi(\sqrt{\log n})}{\sqrt{\log n}} = \frac{2}{\sqrt{2\pi}}\frac{n^{-1/2}}{\sqrt{\log n}} \tag{10}$$

(here we use the well-known fact $1 - \Phi(t) \sim \varphi(t)/t$ for $t \to \infty$, where Φ and φ stand for the standard normal distribution function and density, respectively). Similarly

$$\Pr[T_2 > x, T_2 - T_1 \ge \log n] \le \Pr[T_2 > x] \le \Pr[T_2 > 2\log n] \doteq \exp\{-\tfrac{1}{2}2\log n\} = n^{-1}. \tag{11}$$

In (10) and (11), the use of the approximations of the tail probabilities by the tail probabilities of the limiting χ^2 distributions is correct, see Woodroofe (1978). Hence $\Pr[T_2 - T_1 \ge \log n]$ converges to zero much slower than $\Pr[T_2 > x, T_2 - T_1 \ge \log n]$, and thus the latter probability may be neglected in (9). Therefore, finally,

$$\begin{aligned} H(x) &\doteq \Pr[T_1 \le x, T_2 - T_1 \le \log n] + \Pr[T_2 - T_1 \ge \log n] \\ &\doteq [2\Phi(\sqrt{x}) - 1][2\Phi(\sqrt{\log n}) - 1] + 2[1 - \Phi(\sqrt{\log n})], \quad x \ge 2\log n. \end{aligned}$$

For x between $\log n$ and $2\log n$ Kallenberg and Ledwina (1995) suggested to linearize as follows

$$H(x) \doteq H(\log n) + \frac{x - \log n}{\log n}[H(2\log n) - H(\log n)], \quad \log n < x < 2\log n.$$

Let us summarise the results: for the distribution function of the test statistic T_S we use the approximation

$$(12) \quad H(x) = \Pr[T_S \le x]$$

$$\doteq \begin{cases} [2\Phi(\sqrt{x}) - 1][2\Phi(\sqrt{\log n}) - 1], & x \le \log n, \\ H(\log n) + \frac{x-\log n}{\log n}[H(2\log n) - H(\log n)], & x \in (\log n, 2\log n), \\ [2\Phi(\sqrt{x}) - 1][2\Phi(\sqrt{\log n}) - 1] + 2[1 - \Phi(\sqrt{\log n})], & x \ge 2\log n. \end{cases}$$

4. Simulation study

We investigate performance of the proposed tests through simulations. Our tests are compared with standard tests based on various functionals of the score process (of the Kolmogorov–Smirnov, hereafter KS, Cramér–von Mises, CM, and Anderson–Darling, AD, type). The standard tests were thoroughly examined by Kvaløy and Neef (2004) (throughout this section refered to as KN) in an extensive simulation study. For the sake of ease of comparison, we illustrate the behaviour of our tests in the same models. These include survival data where the proportionality assumption is satisfied, as well as models with both monotonic and nonmonotonic hazard ratios.

To clarify the terminology, we recall that by the 'score tests' we mean the smooth tests based on the score vector $U_2(\tau; \hat{\beta})$ (with the test statistic T_k or T_S). On the contrary, by the term 'score process based tests' we mean the tests of the KS, CM and AD type. By using the word 'process' we stress that the test employs the whole path of the score process $U_1(\cdot; \hat{\beta})$.

The smooth tests with both a fixed and data-driven choice of the number of basis functions are compared (we consider $d = 3, 4, 5, 6$, which is either the dimension for the smooth test or the maximum dimension for the data-driven version). The choice of the basis of functions does not seem to be of great importance; the Legendre polynomial basis leads to slightly higher power in some cases and is therefore used in all simulations. The time transformation for the basis functions is in the form (4).

For the null distribution of the data-driven test statistic T_S we consider both the χ^2_1 approximation and the improved approximation (12) presented in the previous section.

For the tests based on the whole score process, the simulation technique of Lin et al. (1993) is used (the method consists of generating a sample from the asymptotic distribution of the score process; the size of the sample is 1000 everywhere).

All tests are performed on a nominal level of 5 %. The number of repetitions of Monte Carlo simulations is 20 000 under the null hypothesis and 5000 under alternatives and in models with two covariates. Thus the standard deviations of the estimated rejection probabilities in Table 1 are about $\sqrt{0.05 \times 0.95/20000} \doteq 0.002$ (at most $\sqrt{0.5 \times 0.5/20000} \doteq 0.004$). The standard deviations of the estimates in the other tables are at most $\sqrt{0.5 \times 0.5/5000} \doteq 0.007$. The simulations and computations are carried out in R. We use the default random number generator which is 'Mersenne Twister'.

First, we consider a model with one covariate whose effect is proportional. The hazard function follows the form $\lambda(t) = 2\exp(Z)$, the covariate Z is $\mathrm{U}(0,1)$ distributed, both without censoring and with $\mathrm{U}(0,1)$ censoring times. The model is the same as in Case 1 of KN. Results are reported in Table 1. The fixed-dimension test preserves the prescribed level. For the data-driven test, the χ^2_1 approximation cannot be used since the nominal level is highly exceeded. The improved approximation based on H of (12) works quite satisfactorily for the sample size $n = 100$. Therefore, in the remaining simulations we use only this approximation and not the χ^2_1 one. The tests based on the score process with simulated critical values preserve the level as well. Note that in models with one covariate we could use asymptotic critical values (based on the corresponding functionals of the Brownian bridge which is the weak limit of the score process in models with one covariate), however in that case particularly the Kolmogorov–Smirnov type test is too conservative; see KN for details.

Now we proceed to two models not satisfying the proportional hazards assumption. In the first one the effect of the covariate varies monotonically in time, the hazard function is $\lambda(t) = 2\exp(4tZ)$, with Z uniformly distributed on $(0,1)$, without censoring (Case 1 of KN) as well as with $\mathrm{U}(0,1)$ censoring. The model with nonmonotonic hazard ratios has the hazard function $\lambda(t) = 2\exp(\beta(t)Z)$ with $\beta(t) = -\log 4 + 1_{[0.3,0.6]}(t)\log 4$, Z is $\mathrm{U}(0,2)$ distributed, without censoring and with censoring at 1.2 (Case 2 of KN). Estimated rejection probabilities are given in Table 2.

TABLE 1. Estimated sizes of the tests in the model $\lambda(t) = 2\exp(Z)$ with Z being $\mathrm{U}(0,1)$ distributed, without censoring and with $\mathrm{U}(0,1)$ censoring (giving a 30 % censoring rate). Figures based on 20 000 Monte Carlo repetitions (standard deviation of the estimates about 0.002).

		No censoring		Censoring $\mathrm{U}(0,1)$	
		$n = 50$	$n = 100$	$n = 50$	$n = 100$
$d = 3$	T_S (H)	0.060	0.050	0.057	0.050
	T_S (χ^2_1)	0.117	0.088	0.117	0.091
	T_d	0.057	0.057	0.052	0.054
$d = 4$	T_S (H)	0.063	0.051	0.058	0.051
	T_S (χ^2_1)	0.120	0.089	0.119	0.091
	T_d	0.056	0.057	0.049	0.053
$d = 5$	T_S (H)	0.064	0.052	0.059	0.051
	T_S (χ^2_1)	0.120	0.089	0.119	0.091
	T_d	0.056	0.057	0.045	0.052
$d = 6$	T_S (H)	0.064	0.052	0.059	0.051
	T_S (χ^2_1)	0.120	0.090	0.119	0.092
	T_d	0.054	0.054	0.042	0.049
	KS	0.052	0.051	0.053	0.057
	CM	0.048	0.047	0.050	0.050
	AD	0.044	0.046	0.044	0.047

In Table 2 we can see the importance of choosing the number of the basis functions properly. If Neyman's tests with fixed dimensions $d = 3, 4, 5, 6$ are used, we observe that the power typically decays as the dimension increases. The reason is obvious: Since the model is well described with one or two basis functions, including additional redundant basis functions (artificial covariates) does not increase the score test statistic dramatically, but, on the other hand, increases critical values (degrees of freedom increase). The results show that the data-driven choice of the dimension based on the modification of Schwarz's selection is a suitable remedy that is worthwhile. The power is stable for various values of the maximal dimension d.

TABLE 2. Estimated powers of the tests in the model $\lambda(t) = 2\exp(4tZ)$ (monotonic HR), where Z is $\mathrm{U}(0,1)$ distributed, without censoring and with $\mathrm{U}(0,1)$ censoring (leading to a 31 % censoring percentage), and in the model $\lambda(t) = 2\exp(\beta(t)Z)$ with $\beta(t) = -\log 4 + 1_{[0.3,0.6]}(t)\log 4$ (nonmonotonic HR), where Z is $\mathrm{U}(0,2)$ distributed, without censoring and with censoring at 1.2 (33 %). Sample size $n = 100$. Figures based on 5000 Monte Carlo repetitions (standard deviation of the estimates about 0.007).

		Monotonic hazard ratio		Nonmonotonic hazard ratio	
		No censoring	Censoring $\mathrm{U}(0,1)$	No censoring	Censoring at 1.2
$d=3$	T_S	0.369	0.194	0.622	0.619
	T_d	0.353	0.192	0.695	0.569
$d=4$	T_S	0.370	0.195	0.628	0.622
	T_d	0.316	0.168	0.665	0.542
$d=5$	T_S	0.370	0.195	0.632	0.623
	T_d	0.289	0.155	0.679	0.503
$d=6$	T_S	0.370	0.195	0.632	0.623
	T_d	0.272	0.143	0.648	0.472
	KS	0.378	0.211	0.470	0.288
	CM	0.432	0.234	0.411	0.240
	AD	0.432	0.233	0.444	0.296

In comparison to the score process based tests, our test is less powerful for detecting monotonic deviations from proportionality but more powerful for detecting nonmonotonic hazard ratios.

Now we examine models with two covariates such that one covariate (Z_1) has a nonproportional effect (both monotonic and nonmonotonic) while the effect of the other covariate Z_2 is proportional.

The model with a monotonic coefficient of Z_1 follows the form $\lambda(t) = \exp(0.5tZ_1 + Z_2 - 8)$. The covariates Z_1, Z_2 are jointly normally distributed, both have expectation 4 and variance 1, their correlation is ρ (various values are considered). Censoring times are drawn from the $\mathrm{U}(0,5)$ distribution. This model corresponds to Case 4 of KN. The second model is $\lambda(t) =$

TABLE 3. Estimated rejection probabilities for both covariates in the model $\lambda(t) = \exp(0.5tZ_1 + Z_2 - 8)$, where Z_1, Z_2 are jointly normal with expectation 4, variance 1 and correlation ρ, censoring times with the $U(0, 5)$ distribution (about 45 % censoring). Sample size $n = 100$. Figures obtained from 5000 Monte Carlo simulations (standard deviation of the estimates about 0.007).

		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
		Z_1	Z_2	Z_1	Z_2	Z_1	Z_2
$d = 3$	T_S	0.344	0.056	0.320	0.055	0.271	0.066
	T_d	0.346	0.055	0.323	0.056	0.262	0.072
$d = 4$	T_S	0.345	0.057	0.320	0.058	0.272	0.068
	T_d	0.306	0.054	0.280	0.057	0.228	0.070
$d = 5$	T_S	0.345	0.058	0.321	0.058	0.273	0.068
	T_d	0.281	0.054	0.261	0.056	0.204	0.064
$d = 6$	T_S	0.345	0.058	0.321	0.058	0.273	0.068
	T_d	0.251	0.056	0.237	0.057	0.187	0.058
	KS	0.409	0.051	0.391	0.070	0.348	0.105
	CM	0.471	0.047	0.452	0.069	0.398	0.115
	AD	0.466	0.040	0.442	0.059	0.387	0.106

$\exp(\beta(t)Z_1 + Z_2 - 8)$, where the nonmonotonic effect of Z_1 is of the form $\beta(t) = 0.4 + 0.7 \times 1_{[1.2,2]}(t)$. The covariates Z_1, Z_2 have the same distribution as in the previous model. Results of testing proportionality for both of the covariates in the two models are displayed in Tables 3 and 4.

Let us notice the behaviour of the tests for Z_2 (whose effect is proportional). Generally, the smooth tests (both data-driven and fixed-dimension) seem to preserve the level better than the score process based tests. When the proportional covariate Z_2 is highly correlated with the nonproportional covariate Z_1, the score process based tests exceed the prescribed nominal level. This behaviour occurs starting with $\rho = 0.5$. Concerning the smooth tests, this problem is apparent too, but rather for very high $\rho = 0.7$ when the effect of Z_1 is nonmonotonic. Therefore, the power results for Z_1 in Tables 3 and 4 should be looked at with caution especially for high values of ρ.

TABLE 4. Estimated rejection probabilities for both covariates in the model $\lambda(t) = \exp(\beta(t)Z_1 + Z_2 - 8)$ with $\beta(t) = 0.4 + 0.7 \times 1_{[1.2,2]}(t)$, where Z_1, Z_2 are jointly normal with expectation 4, variance 1 and correlation ρ, with constant censoring at 5 (giving about 31 % censoring). Sample size $n = 100$. Figures based on 5000 Monte Carlo repetitions (standard deviation of the estimates about 0.007).

		$\rho = 0.3$		$\rho = 0.5$		$\rho = 0.7$	
		Z_1	Z_2	Z_1	Z_2	Z_1	Z_2
$d = 3$	T_S	0.344	0.052	0.322	0.059	0.288	0.097
	T_d	0.336	0.057	0.319	0.068	0.275	0.095
$d = 4$	T_S	0.348	0.053	0.327	0.060	0.290	0.098
	T_d	0.330	0.058	0.309	0.062	0.257	0.091
$d = 5$	T_S	0.349	0.054	0.328	0.060	0.291	0.098
	T_d	0.312	0.057	0.294	0.057	0.242	0.082
$d = 6$	T_S	0.349	0.054	0.328	0.060	0.292	0.098
	T_d	0.293	0.056	0.283	0.058	0.234	0.078
	KS	0.330	0.055	0.336	0.074	0.307	0.124
	CM	0.298	0.055	0.309	0.068	0.301	0.124
	AD	0.277	0.050	0.284	0.065	0.274	0.110

As for Z_1, a similar behaviour as in the one covariate situation of Table 2 is observed: in some cases the power of the fixed-dimension smooth test slightly decays as the dimension increases (however, one has to take into account the standard deviations of the estimates) whereas the power of the data-driven test is stable. Since the nominal level is not satisfied when there is high association, comparison of powers of the smooth tests and the score process based tests is possible only for low correlation between the covariates ($\rho = 0.3$). The standard tests based on the score process are more powerful for detecting monotonically time-varying effects (see Table 3). In the nonmonotonic situation of Table 4, both kinds of tests behave similarly.

5. Conclusion and discussion

The simulation study showed that the proposed smooth test and its data-driven version could be a reasonable alternative to the traditional tests of the proportional hazards assumption based on functionals of the score process. In our simulations, the new test was typically more powerful against alternatives with nonmonotonic hazard ratios. Monotonic hazard ratios were, per contra, better detected by the score process based tests. Although the new procedure does not universally dominate the standard methods, I believe that the proposed approach is worth studying.

For the traditional goodness-of-fit problem there have been proposed various modifications of data-driven tests. Janssen (2003, see further references therein) points out that the penalty with $\log n$ may be too heavy under local alternatives despite certain optimality properties under intermediate alternatives shown in several papers by Ledwina and coauthors. He suggested to use a selection rule equal to the maximum of Schwarz's rule and a prescribed minimal dimension of the alternative. Claeskens and Hjort (2004) considered tests based on a selection rule that chooses the alternative among all possible nonempty subsets of $\{\varphi_1, \ldots, \varphi_d\}$ (in contrast to the standard BIC which chooses one of the d nested alternatives). Inglot and Ledwina (2004) used a combination of the BIC and AIC based on a threshold rule. Extensions of these approaches to censored lifetime data problems may be investigated further.

Another challenge is to improve smooth tests to be more capable to distinguish which covariates are proportional and which not. In Tables 3 and 4 we have seen that in some situations our smooth tests as well as the score process tests are not reliable for such a discrimination. Research on this issue is in progress, some recent results are presented in Kraus (2006).

Acknowledgement

I would like to thank two anonymous reviewers for their detailed insightful comments that lead to a significant improvement of the paper and inspired me for further research. I gratefully acknowledge that the work has been supported by the GAČR Grant No. 201/05/H007. Computations have been carried out in METACentrum (Czech academic supercomputer network).

References

Abrahamowicz, M., MacKenzie, T. and Esdaile, J. M. (1996). Time-dependent hazard ratio: Modeling and hypothesis testing with application in lupus nephritis. *J. Amer. Statist. Assoc.*, 91, 1432–1439.

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.*, 10, 1100–1120.

Claeskens, G. and Hjort, N. L. (2004). Goodness of fit via non-parametric likelihood ratios. *Scand. J. Statist.*, 31, 487–513.

Cox, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, 34, 187–220.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.

Inglot, T., Kallenberg, W. C. M. and Ledwina, T. (1997). Data driven smooth tests for composite hypotheses. *Ann. Statist.*, 25, 1222–1250.

Inglot, T. and Ledwina, T. (2004). Refined version of data driven Neyman's test. Preprint 001, Institute of Mathematics, Wrocław University of Technology.

Janssen, A. (2003). Which power of goodness of fit tests can really be expected: intermediate versus contiguous alternatives. *Statist. Decisions*, 21, 301–325.

Kallenberg, W. C. M. and Ledwina, T. (1995). On data driven Neyman's tests. *Probab. Math. Statist.*, 15, 409–426.

Kallenberg, W. C. M. and Ledwina, T. (1997). Data-driven smooth tests when the hypothesis is composite. *J. Amer. Statist. Assoc.*, 92, 1094–1104.

Kraus, D. (2006). Identifying nonproportional covariates in the Cox model. Research Report 2170, Institute of Information Theory and Automation, Prague.

Kvaløy, J. T. and Neef, L. R. (2004). Tests for the proportional intensity assumption based on the score process. *Lifetime Data Anal.*, 10, 139–157.

Ledwina, T. (1994). Data-driven version of Neyman's smooth test of fit. *J. Amer. Statist. Assoc.*, 89, 1000–1005.

Lin, D. Y., Wei, L. J. and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80, 557–572.

Martinussen, T., Scheike, T. H. and Skovgaard, I. M. (2002). Efficient estimation of fixed and time-varying covariate effects in multiplicative intensity models. *Scand. J. Statist.*, 29, 57–74.

Peña, E. A. (1998a). Smooth goodness-of-fit tests for composite hypothesis in hazard based models. *Ann. Statist.*, 26, 1935–1971.

Peña, E. A. (1998b). Smooth goodness-of-fit tests for the baseline hazard in Cox's proportional hazards model. *J. Amer. Statist. Assoc.*, 93, 673–692.

Peña, E. A. (2003). Classes of fixed-order and adaptive smooth goodness-of-fit tests with discrete right-censored data. In *Mathematical and statistical methods in reliability (Trondheim, 2002)*. World Sci. Publishing, River Edge.

Scheike, T. H. and Martinussen, T. (2004). On estimation and tests of time-varying effects in the proportional hazards model. *Scand. J. Statist.*, 31, 51–62.

Woodroofe, M. (1978). Large deviations of likelihood ratio statistics with applications to sequential testing. *Ann. Statist.*, 6, 72–84.