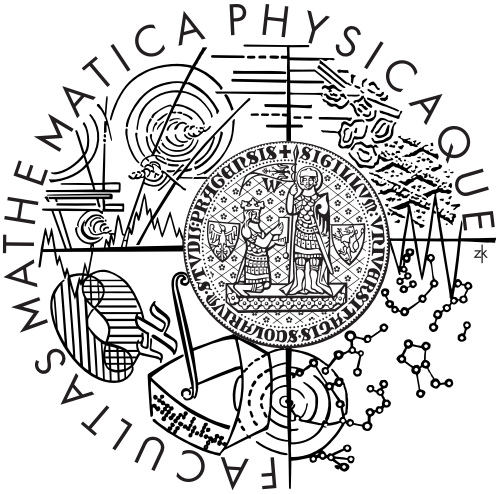
Charles University in Prague
Faculty of Mathematics and Physics



Neyman's smooth tests in survival analysis

by

David Kraus

A dissertation submitted in partial fulfilment
of the requirements for the degree of

Doctor of Philosophy

in the Department of Probability
and Mathematical Statistics

Advisor: Doc. Petr Volf, CSc.
Branch: M4 Probability and Mathematical Statistics

December 2007

To my parents

Acknowledgements

First and foremost, I am deeply indebted to my parents for their love, patience, care and support during my studies.

I would like to thank my thesis advisor Petr Volf for introducing me to event history analysis and for his support and confidence in me. I am grateful to many people in and around the Department of Statistics. I would especially like to appreciate Professor Marie Hušková for her service as the chair of the graduate studies in Statistics. I thank Marek Omelka for advices in topics I was not proficient in.

I was supported by the GAČR grant 201/05/H007, GAAV grant IAA101120604, MŠMT project 1M06047, and by the Institute of Information Theory and Automation, Prague. Most computations were carried out in METACentrum (Czech academic supercomputer network).

Contents

Introduction

Smooth tests

The idea of smooth tests of goodness of fit is due to Neyman (1937). Consider a sample from a continuous distribution with density f and distribution function F. To test the simple null hypothesis H_0: $f = f_0$ (for some specified density f_0), Neyman suggested to replace the general alternative $f \neq f_0$ by an alternative of order k with density

$$f(x;\theta) = f_0(x) \exp\{\theta^{\mathsf{T}} \varphi(F_0(x)) - c(\theta)\}, \quad x \in \mathbb{R},$$

where $\varphi(u) = (\varphi_1(u), \ldots, \varphi_k(u))^{\mathsf{T}}, u \in [0,1]$ are functions describing departures from the null density f_0 and $c(\theta)$ is a normalising constant. The problem then translates to the task of testing $\theta = 0$ versus $\theta \neq 0$. The smooth test is the Rao score test (or likelihood ratio or Wald test) of significance of θ.

Neyman used $\varphi_1, \ldots, \varphi_k$ equal to the orthonormal Legendre polynomials of degree $1, \ldots, k$. Other linearly independent functions on $[0,1]$, independent of a constant function, are possible, for instance cosines with various frequencies. If φ_j are indicators of the intervals in a partition of $[0,1]$, the test becomes Pearson's X^2 test, which is what probably most people imagine when the term 'goodness of fit' is mentioned.

Smooth tests are a compromise between directional and omnibus tests. Directional tests are designed to have high power against a specific departure from the null hypothesis (e.g., a location-shift alternative). Omnibus tests (e.g., the Kolmogorov–Smirnov test) are constructed without any specific alternative in mind, they should be able to detect the deviation from the hypothesis in any direction. Smooth tests are somewhere on the way between these two approaches. They focus on a finite number of directions in the space of alternatives. They are intended to detect a wider spectrum of alternatives than directional tests but they leave aside very complicated departures from the hypothesis.

Smooth tests may also be used when the hypothesis is composite, i.e., for testing fit of a parametric family. The method is applicable for discrete distributions as well. A detailed account of the theory and applications of smooth tests can be found in the book by Rayner and Best (1989).

Besides the embedding, there are at least two other ways of deriving smooth tests in the simple goodness-of-fit setting: they may be motivated as combinations of rank tests, or as truncated series of L^2 integral statistics. These relationships will be seen in the context of this thesis in some parts. However, mainly the embedding point of view is of interest because it opens ways to extensions to much more situations than the classical goodness of fit with simple or composite hypotheses. The embedding of the null hypothesis in an alternative of order k is a general way of constructing statistical tests. Here I develop such smooth tests for some two-sample problems and some tests of fit in regression models.

Neyman's embedding idea is one of two main ingredients applied in this thesis. The second one is the idea of data-driven tests which is due to Ledwina (1994). Data-driven tests were

developed to address the important issue of the choice of the number k of the basis functions used in the smooth test. Ledwina's approach is based on Schwarz's selection rule which is the value of k which maximises the penalised score statistic (or likelihood). When the suitable alternative is selected, the smooth test is applied against this alternative. Thus, the idea is to let the test adapt to the data. Further variants of this strategy were developed, and some of them will be considered to some extent in some passages of this thesis.

Thesis outline

As the title of the thesis suggests, I study Neyman's smooth tests and their data-driven versions in the context of survival analysis. Surprisingly, not much work has been done in the development of smooth tests for survival data. The main contributions seem to be those of Peña (1998a,b) who proposed tests for some parametric survival models. The problems dealt with in this thesis are nonparametric or semiparametric. Two main areas are covered: Part I (Chapters 1, 2, 3) investigates various two-sample problems, Part II (Chapters 4, 5, 6) develops methods for Cox's proportional hazards regression model.

Chapter 1 deals with the simplest task of the comparison of the survival distributions in two samples of right-censored data. The hypothesis is most conveniently formulated in terms of hazard rates which enables us to apply Neyman's embedding idea. The smooth test is derived and various variants of the data-driven selection procedure are discussed. Their asymptotic behaviour is studied, the performance of the asymptotic approximations is investigated and some improvements provided. A Monte Carlo study explores the behaviour of various selection rules and compares these tests with standard two-sample tests.

In Chapter 2, the attention turns to a two-sample problem for competing risks data. In survival data with competing risks, not only the failure time is recorded but also the cause or type of failure is available (for uncensored observations which actually experience the failure). The chapter focuses on the comparison of the cumulative incidence functions. The embedding construction is accomplished in terms of subdistribution hazard rates. Some asymptotic results are derived and simulations carried out.

Chapter 3 is devoted to the verification of the assumption of proportional rates in two samples, with main examples being proportional hazards and proportional odds. Unlike in the first two chapters, here the null hypothesis model is semiparametric (a linear transformation model). The smooth test is obtained by embedding the transformation rate into a k-dimensional alternative. Asymptotic properties derived and the proposed tests are compared with some existing methods through simulations.

In Chapter 4, I develop data-driven smooth tests of the proportional hazards assumption in the Cox regression model. The chapter is focused on the testing of the hypothesis that the effect of a covariate is constant in time versus its time dependence. The smooth modelling of the possibly time-varying effect gives rise to the smooth test. The relationship between the smooth tests and the integral tests is demonstrated.

Chapter 5 studies the use of smooth tests and some other existing methods in Cox models with multiple covariates where the proportionality assumption may be violated for several covariates. It is shown that the initial naïve approch may be misleading when one wishes to identify which covariate satisfies the assumption and which not. An improvement is proposed and investigated via simulations.

Chapter 6 contains a short suggestion of global tests of the proportional hazards assumption

which complements the previous topic of covariate-specific tests.

The thesis comes with two software packages which are briefly described in Appendix A.

The first five chapters approximately correspond to five papers. Their content has been partly extended, partly reduced and partly reorganised because the chronological order of their creation differs from the order of the chapters. Nonetheless, the chapters are rather self-contained.

Chapter 1: Kraus (2007a). Adaptive Neyman's smooth tests of homogeneity of two samples of survival data. Research Report 2187, Institute of Information Theory and Automation, Prague. Submitted.

Chapter 2: Kraus (2007d). Smooth tests of equality of cumulative incidence functions in two samples. Research Report 2197, Institute of Information Theory and Automation, Prague. Submitted.

Chapter 3: Kraus (2007b). Checking proportional rates in the two-sample transformation model. Research Report 2203, Institute of Information Theory and Automation, Prague. Submitted.

Chapter 4: Kraus (2007c). Data-driven smooth tests of the proportional hazards assumption. *Lifetime Data Anal.*, 13, 1–16.

Chapter 5: Kraus (2008). Identifying nonproportional covariates in the Cox model. *Comm. Statist. Theory Methods*, 37. To appear.

Part I

Two-sample tests

1 Homogeneity of two samples of censored survival data

Summary

The problem of testing whether two samples of possibly right-censored survival data come from the same distribution is considered. The aim is to develop a test which is capable of detection of a wide spectrum of alternatives. A new class of tests based on Neyman's embedding idea is proposed. The null hypothesis is tested against a model where the hazard ratio of the two survival distributions is expressed by several smooth functions. A data-driven approach to the selection of these functions is studied. Asymptotic properties of the proposed procedures under alternatives are discussed. Small-sample performance is explored via simulations which show that the power of the proposed tests appears to be more robust than the power of some versatile tests previously proposed in the literature (such as combinations of weighted logrank tests, or Kolmogorov–Smirnov tests).

1.1 Introduction

This chapter presents a new approach to testing homogeneity of two samples of right censored survival data. The goal is to provide an 'omnibus' test procedure sensitive against a range of alternatives. Such a test is useful, for instance, in situations when the Kaplan–Meier curves for the two samples do not suggest an alternative against which one should test (e.g, when the curves cross, as in the example in Section 1.8), or in situations when the visual inspection of the Kaplan–Meier plots is impossible (e.g., when data must be analysed automatically).

Omnibus tests cannot have power superior to all tests in all situations. Therefore, the objective is different: it is desirable to have a test which should not fail against a rather broad spectrum of alternatives. The method proposed in this chapter achieves this goal.

Consider two samples of survival data. The jth sample consists of observations $(T_{j,i}, \delta_{j,i})$, $i = 1, \ldots, n_j$, $j = 1, 2$, where $T_{j,i} = R_{j,i} \wedge C_{j,i}$ is the possibly censored survival time, $\delta_{j,i} = 1_{[R_{j,i} \le C_{j,i}]}$ is the failure indicator, $R_{j,i}$ is the unobserved survival time and $C_{j,i}$ is the unobserved censoring time. The survival time and censoring time are assumed to be independent. All $n = n_1 + n_2$ observations $(T_{j,i}, \delta_{j,i})$ are mutually independent. The times $R_{j,i}$ come from a distribution with hazard function $\alpha_j(t)$. The aim is to test the hypothesis H_0: $\alpha_1 = \alpha_2$ without any specific alternative in mind.

The standard counting process notation is adopted (which makes it possible to extend the results of the chapter to a broader range of situations, e.g., to data with recurrent events; however, survival data will be of primary interest here). Consider an n-variate counting process $N(t) = (N_{1,1}(t), \ldots, N_{1,n_1}(t), N_{2,1}(t), \ldots, N_{2,n_2}(t))^{\mathsf{T}}$ observed on a finite interval $[0, \tau]$. Denote its cumulative intensity process $\Lambda(t)$ with components $\Lambda_{j,i}(t) = \int_0^t \lambda_{i,j}(s)ds$ which are compensators of $N_{j,i}(t)$, $j = 1, 2$, $i = 1, \ldots, n_j$. The intensity processes fol-

low the form $\lambda_{j,i}(t) = Y_{j,i}(t)\alpha_j(t)$ where $Y_{j,i}(t)$ are the at-risk indicator processes. Denote $\bar{N}_j(t) = \sum_{i=1}^{n_j} N_{j,i}(t)$, $\bar{N}(t) = \bar{N}_1(t) + \bar{N}_2(t)$, $\bar{Y}_j(t) = \sum_{i=1}^{n_j} Y_{j,i}(t)$ and $\bar{Y}(t) = \bar{Y}_1(t) + \bar{Y}_2(t)$.

The traditional approach is to use a weighted logrank test statistic $\int_0^\tau L(t)dU_0(t)$, where the logrank process equals

$$U_0(t) = \int_0^t \frac{\bar{Y}_1(s)\bar{Y}_2(s)}{\bar{Y}(s)} \left[\frac{d\bar{N}_2(s)}{\bar{Y}_2(s)} - \frac{d\bar{N}_1(s)}{\bar{Y}_1(s)} \right].$$

Harrington and Fleming (1982) proposed to use weight functions from the $G^{\rho,\gamma}$ class of the form $L(t) = K(\hat{S}(t-))$ with $K(u) = u^\rho(1-u)^\gamma$, $\rho, \gamma \geq 0$ and $\hat{S}$ being an estimator of the survival function computed from the pooled sample (e.g., the Kaplan–Meier estimator or the exponential of minus the Nelson–Aalen estimator). Various members of this class are suitable for discovering various departures from the null hypothesis. Obviously, tests with $\rho > 0$ and $\gamma = 0$ are sensitive against early differences in hazard functions, tests with $\rho = 0$ and $\gamma > 0$ are powerful against late differences, a choice with $\rho > 0$ and $\gamma > 0$ yields a test good at detecting middle differences and the logrank test $G^{0,0}$ does well under proportional hazards. More precise results on performance of $G^{\rho,\gamma}$ tests are can be found in Fleming and Harrington (1991, Chapter 7) or Andersen, Borgan, Gill and Keiding (1993, Section V.2). For instance, the logrank test $G^{0,0}$ is optimal (locally efficient) against the proportional hazards alternative (as it is the partial likelihood score test in a Cox model with a group indicator covariate) and the Prentice–Wilcoxon statistic $G^{1,0}$ is optimal against shift alternatives in the logistic distribution.

The $G^{\rho,\gamma}$ tests are directed against specific alternatives. While such a test is highly sensitive (often optimal) against the particular direction in the space of alternatives, it may fail to detect different kinds of alternatives. One often does not have a clear advance idea of the nature of heterogeneity of the samples. Therefore, more omnibus tests were developed. Fleming and Harrington (1991, Section 7.5) describe two classes of such tests: tests using the whole path of the logrank process U_0 and tests combining several statistics of the $G^{\rho,\gamma}$ type. The former include supremum (Kolmogorov–Smirnov) tests and integral tests (of the Cramér–von Mises and Anderson–Darling type). See also Gill (1980, Section 5.4) and Schumacher (1984). The latter class uses the maximum or sum of a finite cluster of weighted logrank statistics. Yet another procedure has been proposed by Pecková and Fleming (2003) who select a statistic from this cluster on the basis of estimated asymptotic relative efficiencies (within the cluster) against location shift alternatives.

Here I make a further step towards versatile tests with robust power, that is towards tests which on one hand do not collapse against a wide range of alternatives and on the other hand do not lose much compared to optimal directional tests.

My approach is based on Neyman's embedding idea combined with Schwarz's selection rule. The null nonparametric model of homogeneous samples is viewed as a submodel of a larger semiparametric model in which the hazard ratio of the two samples is expressed in terms of several smooth functions. A score test is applied to testing the null model versus the smooth model. Furthermore, selection criteria are used for choosing the smooth model. This data-driven strategy is inspired by the approach of Ledwina (1994) and Inglot, Kallenberg and Ledwina (1997). Smooth tests in the context of event history analysis were previously considered by Peña (1998a,b).

In Section 1.2 the smooth test is constructed. Section 1.3 provides its data-driven version. In Section 1.4 an approximation for the null distribution of one type of the data-driven test

is derived. The behaviour of the proposed procedures under alternatives is investigated in Section 1.5. The Monte Carlo study of Section 1.6 explores level properties and power. Section 1.7 contains a remark on the principal components analysis of the Cramér–von Mises test and its relation to Neyman's test. The method is illustrated on a real data set in Section 1.8.

1.2 Construction of Neyman's test

Neyman's goodness-of-fit idea is here used as follows. The null model with $\alpha_1 = \alpha_2$ is embedded in a d-dimensional model

$$\alpha_2(t) = \alpha_1(t) \exp\{\theta^{\mathsf{T}} \psi(t)\}, \tag{1.1}$$

where $\theta = (\theta_1, \dots, \theta_d)^{\mathsf{T}}$ is a parameter and $\psi(t) = (\psi_1(t), \dots, \psi_d(t))^{\mathsf{T}}$ are some bounded functions modelling possible difference of α_2 from α_1. The functions $\psi_k(t)$ are taken in the form $\psi_k(t) = \varphi_k(g(t))$ where $\{\varphi_1, \dots, \varphi_d\}$ forms a set of linearly independent bounded functions in $L^2[0,1]$ and g is an increasing transformation that maps the time period $[0, \tau]$ to $[0,1]$.

The task of testing $\alpha_1 = \alpha_2$ versus (1.1) is equivalent to testing H_0: $\theta = 0$ versus H_d: $\theta \neq 0$. It is advantageous to introduce the group indicator variable $Z_{j,i} = 1_{[j=2]}$. With this notation the intensities admit the form

$$\lambda_{j,i}(t) = Y_{j,i}(t) \alpha(t) \exp\{\theta^{\mathsf{T}} \psi(t) Z_{j,i}\}. \tag{1.2}$$

Hence we arrive at a Cox proportional hazards model with d artificial time-dependent covariates $\psi_1(t)Z_{j,i}, \dots, \psi_d(t)Z_{j,i}$ whose significance is to be tested. To this end we may use well-known partial likelihood tools, of which the score test is particularly appealing as it does not involve estimation of θ.

The score process for the Cox model (1.2) under $\theta = 0$ takes the form

$$U(t) = \int_0^t \psi(s) \frac{\bar{Y}_1(s)\bar{Y}_2(s)}{\bar{Y}(s)} \left[\frac{d\bar{N}_2(s)}{\bar{Y}_2(s)} - \frac{d\bar{N}_1(s)}{\bar{Y}_1(s)} \right] = \int_0^t \psi(s) dU_0(s).$$

Denote by $\bar{y}_1, \bar{y}_2$ the uniform limits in probability of $n^{-1}\bar{Y}_1, n^{-1}\bar{Y}_2$, respectively (they exist by the Glivenko–Cantelli theorem if $n_j/n \to a_j$), and assume that the limit functions are bounded away from zero on $[0, \tau]$ (this holds if $a_j \in (0,1)$); let $\bar{y}$ stand for $\bar{y}_1 + \bar{y}_2$. It is known (Fleming and Harrington, 1991, Corollary 7.2.1; Andersen et al., 1993, Theorem V.2.1) that under the hypothesis $\alpha_1 = \alpha_2$ the logrank process $n^{-1/2}U_0$ converges weakly in $D[0, \tau]$ with Skorohod topology to a zero mean Gaussian martingale V_0 whose variance function and its uniformly consistent estimator are

$$\sigma_0(t) = \int_0^t \frac{\bar{y}_1(s)\bar{y}_2(s)}{\bar{y}(s)} dA(s), \qquad n^{-1}\hat{\sigma}_0(t) = n^{-1} \int_0^t \frac{\bar{Y}_1(s)\bar{Y}_2(s)}{\bar{Y}(s)} \frac{d\bar{N}(s)}{\bar{Y}(s)}.$$

Consequently, under the null (i.e., $\theta = 0$) the process $n^{-1/2}U$ is asymptotically distributed as a d-variate zero mean Gaussian martingale V with covariance matrix function and its estimator

$$\sigma(t) = \int_0^t \psi(s)^{\otimes 2} d\sigma_0(s), \qquad n^{-1}\hat{\sigma}(t) = n^{-1} \int_0^t \psi(s)^{\otimes 2} d\hat{\sigma}_0(s)$$

(that is, $\mathrm{cov}(V_k(s), V_l(t)) = \sigma_{k,l}(s \wedge t)$). (The standard notation $a^{\otimes k}$, where a is a column vector, means $1, a, aa^{\mathsf{T}}$ for $k = 0, 1, 2$, respectively.)

The partial likelihood score statistic

$$T_d = U(\tau)^{\mathsf{T}} \hat{\sigma}(\tau)^{-1} U(\tau)$$

used for testing $\theta = 0$ versus $\theta \neq 0$ is asymptotically chi-squared distributed with d degrees of freedom. The hypothesis is rejected for large values of T_d.

As mentioned before, the basis functions $\varphi_1, \ldots, \varphi_d$ are linearly independent. We can take several functions from a well-known orthonormal basis of $L^2[0,1]$. For instance, we can use the cosine basis $\varphi_k(u) = \sqrt{2}\cos(k\pi u)$, $k = 1, \ldots, d$, or orthonormal Legendre polynomials on $[0,1]$. It is natural (but not always necessary) to have the unity in the linear span of these functions in order to capture possible proportional hazards alternatives. We can set $\varphi_1 \equiv 1$ (note that a model of the form (1.2) containing an intercept is identifiable) and the other functions may be cosines or Legendre polynomials. Another option is to choose a partition $0 = u_0 < u_1 < \cdots < u_d = 1$ and let the basis functions be indicators of these intervals, i.e., $\varphi_k(u) = 1_{(u_{k-1}, u_k]}(u)$, $k = 1, \ldots, d$.

Modelling the logarithm of the hazard ratio by linear combinations of smooth functions is a flexible approach. For instance, consider $d = 3$ polynomials (of order $0, 1, 2$). Their linear span contains the weight functions $G^{0,0}$, $G^{1,0}$, $G^{0,1}$ and $G^{1,1}$. Hence $\varphi_1, \varphi_2, \varphi_3$ can capture the same alternatives as the four logrank weights (proportional hazards, early, middle and late differences). Moreover, also crossing hazards and hence non-location alternatives (crossing survival curves) can be expressed by combinations of $\varphi_1, \varphi_2, \varphi_3$.

The time-transformation $g : [0, \tau] \to [0, 1]$ may be simply $g(t) = t/\tau$. However, the purpose of the transformation is to standardise the speed of time so as to spread the observations evenely in $[0,1]$ and benefit from the flexibility of the basis functions. The aim is to avoid clustering many observations in certain parts of $[0,1]$ and leaving other parts not fully used. To this end we use $g(t) = F(t)/F(\tau)$ (with $F(t) = 1 - S(t)$ being the common distribution function of survival times). Alternatively, as discussed in Chapter 4, one may use $g(t) = A(t)/A(\tau)$ (with $A(t) = \int_0^t \alpha(s)ds$, the corresponding cumulative hazard) but, rather, this choice is suitable for data with recurrent events. Yet another possibility is $g(t) = \sigma_0(t)/\sigma_0(\tau)$. If the functions $\varphi_1, \ldots, \varphi_d$ are orthonormal this transformation yields a diagonal asymptotic covariance matrix (that is, the components of the score vector are asymptotically independent). In practice, a transformation depending on unknown quantities is replaced by a uniformly consistent estimator $\hat{g}$ computed from the pooled sample.

1.3 Selection rules and adaptive tests

The difference between α_1 and α_2 is often well described by less than all d smooth functions. However, one does not know which functions should be included in the model and which not. Omitting a function that highly contributes to the description of the data or including an improper function may result in bad performance of the test. Therefore, it is reasonable to let the test adapt to the data, make the test data-driven.

This is accomplished by means of Schwarz's selection rule (or Bayesian information criterion, BIC). The adaptive test consists of two steps. First, a subset of $\{\varphi_1, \ldots, \varphi_d\}$ is selected on the basis of Schwarz's rule. Once a subset is selected, the score test against this likely alternative is performed.

The idea of data-driven Neyman's smooth tests is due to Ledwina (1994). She applied Schwarz's selection rule to the task of testing uniformity (or other single distribution). Inglot et al. (1997) and Kallenberg and Ledwina (1997) extended this method to goodness-of-fit tests of composite hypotheses. Janic-Wróblewska and Ledwina (2000) developed data-driven rank tests for the classical two-sample problem and Antoch, Hušková, Janic and Ledwina (2007) proposed a test of this type for a change-point problem.

We must specify a class $\mathcal{S}$ of nonempty index subsets out of which the selection rule will pick the most suitable one. I consider two classes previously proposed in the literature. Ledwina (1994) used d nested subsets of the form $\mathcal{S}^{\text{nested}} = \{\{1\}, \{1,2\}, \ldots, \{1,\ldots,d\}\}$. This choice is reasonable when the basis functions are naturally ordered, e.g., according to increasing complexity (which is the case, for instance, for the cosine basis with increasing frequencies but hardly for the indicator basis). Claeskens and Hjort (2004) proposed to use all nonempty subsets of $\{1,\ldots,d\}$, that is $\mathcal{S}^{\text{all}} = 2^{\{1,\ldots,d\}} \setminus \{\emptyset\}$. I also consider the strategy proposed by Janssen (2003). He suggested to prescribe a set of basis functions of primary interest which are always included. Without loss of generality, let these functions be several first basis functions, i.e., let their indices be $C_0 = \{1,\ldots,d_0\}$ for some d_0 (with $d_0 = 0$ meaning $C_0 = \emptyset$). Then the class of subsets is $\{C \cup C_0 : C \in \mathcal{S}'\}$ (where $\mathcal{S}'$ may be $\mathcal{S}^{\text{nested}}$, $\mathcal{S}^{\text{all}}$, or some other class of nonempty sets).

Schwarz's criterion (a modification proposed by Ledwina, 1994) selects the set S maximising the penalised score statistic, i.e.,

$$S = \arg\max_{C\in\mathcal{S}} \{T_C - |C| \log n\},$$

where $|C|$ denotes the number of elements of C and T_C stands for the score statistic computed in the model with basis functions φ_k, $k \in C$. The adaptive test is based on T_S.

The asymptotic behaviour of the statistic T_S is given by the following theorem.

Theorem 1.1. *Denote $d^* = \min\{|C| : C \in \mathcal{S}\}$ (that is $d^* = \max(d_0, 1)$). Then, under the null hypothesis, the selection criterion asymptotically concentrates in sets of dimension d^*, i.e., $\Pr[|S| = d^*] \to 1$ as $n \to \infty$. Consequently, T_S is asymptotically distributed as*

$$\max\{V_C(\tau)^{\mathsf{T}} \sigma_{CC}(\tau)^{-1} V_C(\tau) : C \in \mathcal{S}, |C| = d^*\},$$

where $V_C(\tau)$ and $\sigma_{CC}(\tau)$ are, respectively, the subvector and submatrix of $V(\tau)$ and $\sigma(\tau)$ corresponding to the subset C.

Proof. Any d^*-dimensional set C asymptotically wins against any set $\tilde{C}$ of dimension $k > d^*$ because $\Pr[T_{\tilde{C}} - k \log n < T_C - d^* \log n] = \Pr[T_{\tilde{C}}/\log n - T_C/\log n < k - d^*] \to 1$. Among d^*-dimensional sets the one whose score statistic is maximal is selected. □

When no high priority directions are specified ($d_0 = 0$) the nested subsets test statistic is approximately χ^2_1-distributed. Although asymptotically valid the χ^2_1 approximation is inaccurate for small samples (it will be seen in simulations in Section 1.6). A two-term approximation taking into account the possibility of selection not only of the set $\{1\}$ but also $\{1,2\}$ is provided in the next section.

For the class of all subsets with $d_0 = 0$, the test statistic T_S converges to the variable $\max\{V_1(\tau)^2/\sigma_{11}(\tau), \ldots, V_d(\tau)^2/\sigma_{dd}(\tau)\}$, the maximum of generally dependent χ^2_1 variables. It may be easily approximated by simulation from the distribution of $V(\tau)$ (zero-mean normal with variance matrix estimated by $n^{-1}\hat{\sigma}(\tau)$).

With $d_0 > 0$ both nested and all subsets criterion gives a statistic with asymptotic χ^2 distribution with d_0 degrees of freedom.

Small-sample accuracy of asymptotic approximations is investigated via simulations in Section 1.6.

1.4 Two-term approximation of the null distribution

Let us focus on the data-driven test with the class of nested subsets statistic with $d_0 = 0$. Kallenberg and Ledwina (1997) (in their classical goodness-of-fit context) 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, as will be seen in simulations in Section 1.6. Kallenberg and Ledwina (1997, p. 1097) (see also Kallenberg and Ledwina, 1995) derived a much more accurate approximation. Here I adapt their ideas to the present setting.

First, we write

$$\Pr[T_S \le x] = \Pr[T_1 \le x, |S| = 1] + \Pr[T_2 \le x, |S| = 2] + \Pr[T_S \le x, |S| \ge 3].$$

Under the null, the second and third term on the right-hand side asymptotically vanish, and the first term converges to the χ^2_1 distribution function. However, the convergence of S to the smallest is not fast enough. To improve accuracy of the approximation, only the third (and not the second) term will be neglected. The event $[|S| = 1]$ is approximated by $[T_1 - \log n \ge T_2 - 2\log n] = [T_2 - T_1 \le \log n]$ (in words, the event "dimension 1 wins over all the other dimensions" is approximated by "dimension 1 wins over the dimension 2"). Similarly, $[|S| = 2]$ is approximated by $[T_2 - T_1 \ge \log n]$.

Before we proceed, we need to investigate the asymptotic distribution of $(T_1, T_2 - T_1)^\mathsf{T}$. The variables T_1, T_2 are functions of the score $U(\tau)$ that is asymptotically distributed as a bivariate normal vector $(R_1, R_2)^\mathsf{T}$ with variance matrix $\sigma = \sigma(\tau)$. Denote elements of σ as $\left(\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right)$ and $\rho = b/\sqrt{ac}$. The distribution $\mathrm{N}(0, \sigma)$ 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 $(R_1, R_2)\sigma^{-1}(R_1, R_2)^\mathsf{T} \sim (\tilde{R}_1, \tilde{R}_2)\sigma^{-1}(\tilde{R}_1, \tilde{R}_2)^\mathsf{T} =: T_2^\infty$. 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^2_1 variables.

Now we can study $\Pr[T_S \le x]$. We will treat $\Pr[T_S \le x]$ 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

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

If $x \geq 2\log n$,

$$\Pr[T_2 \leq x, |S| = 2] \doteq \Pr[T_2 \leq x, T_2 - T_1 \geq \log n] \doteq \Pr[T_2 - T_1 \geq \log n].$$

Motivation for the latter approximation is as follows. Rewrite

$$\Pr[T_2 \leq x, T_2 - T_1 \geq \log n] = \Pr[T_2 - T_1 \geq \log n] - \Pr[T_2 > x, T_2 - T_1 \geq \log n]. \tag{1.3}$$

As $T_2 - T_1$ is approximately χ^2_1 distributed, we have

$$\Pr[T_2 - T_1 \geq \log n] \doteq 2(1 - \Phi(\sqrt{\log n})) \doteq 2\frac{\varphi(\sqrt{\log n})}{\sqrt{\log n}} = \frac{2}{\sqrt{2\pi}}\frac{n^{-1/2}}{\sqrt{\log n}} \tag{1.4}$$

(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

$$\Pr[T_2 > x, T_2 - T_1 \geq \log n] \leq \Pr[T_2 > x] \leq \Pr[T_2 > 2\log n] \doteq \exp\{-\tfrac{1}{2}2\log n\} = n^{-1}. \tag{1.5}$$

In (1.4) and (1.5), 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 $\Pr[T_2 - T_1 \geq \log n]$ converges to zero much slower than $\Pr[T_2 > x, T_2 - T_1 \geq \log n]$, and thus the latter probability may be neglected in (1.3). Therefore, finally,

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

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

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

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

$$\begin{aligned}\Pr[T_S \leq x] &\doteq H(x) \\ &= \begin{cases} [2\Phi(\sqrt{x}) - 1][2\Phi(\sqrt{\log n}) - 1], & x \leq \log n, \\ H(\log n) + \frac{x - \log n}{\log n}[H(2\log n) - H(\log n)], & x \in (\log n, 2\log n), \\ [2\Phi(\sqrt{x}) - 1][2\Phi(\sqrt{\log n}) - 1] + 2[1 - \Phi(\sqrt{\log n})], & x \geq 2\log n. \end{cases}\end{aligned} \tag{1.6}$$

1.5 Behaviour under alternatives

Let us investigate when the smooth tests and their data-driven versions are consistent. Consider a fixed general alternative of different hazards in the two samples, i.e., $\alpha_1(t) \neq \alpha_2(t)$ on a non-null set.

Theorem 1.2. *Denote by $\bar{y}_1^*, \bar{y}_2^*, g^*$ functions to which $n^{-1}\bar{Y}_1, n^{-1}\bar{Y}_2, \hat{g}$ converge in probability under the fixed alternative and let $\psi^*(t) = \varphi(g^*(t))$. Then smooth tests (both fixed-dimensional and data-driven) are consistent against any alternative satisfying*

$$\int_0^\tau \psi^*(t)\frac{\bar{y}_1^*(t)\bar{y}_2^*(t)}{\bar{y}^*(t)}(\alpha_2(t) - \alpha_1(t))dt \neq 0 \tag{1.7}$$

(i.e., at least one component is nonzero).

Proof. The left-hand side in (1.7) is the limit in probability of $n^{-1}U(\tau)$ under the alternative. The variance estimator $n^{-1}\hat{\sigma}(\tau)$ converges under the alternative to a finite matrix. Therefore, the limit of $n^{-1}T_d$ is nonzero and consistency of the fixed-dimensional test follows. To see consistency of data-driven tests it remains to realise that for any subset $C \in \mathcal{S}$ containing at least one index corresponding to a nonzero component of (1.7) it holds that $T_C - |C| \log n \to \infty$ in probability. Hence some of subsets with the score statistic converging to infinity will be selected with probability converging to 1 which proves the assertion. □

The condition (1.7) may be interpreted as follows. Our working model is (1.2). The true form of the hazard functions is, however, more general: it may be rewritten as $\lambda_{j,i}(t) = Y_{j,i}(t)\alpha(t)\exp\{\eta(t)Z_{j,i}\}$, where the function η is nonzero on a non-null set. Thus we work with a (possibly) misspecified Cox model. Struthers and Kalbfleisch (1986, Theorem 2.1) (see also Lin and Wei, 1989) show that the maximum partial likelihood estimator in a misspecified proportional hazards model converges to the solution to a limiting estimating equation. In our situation this limiting equation for θ is

$$\int_0^\tau \psi^*(t)\frac{\bar{y}_1^*(t)\bar{y}_2^*(t)}{\bar{y}_1^*(t)+\bar{y}_2^*(t)\exp\{\theta^\mathsf{T}\psi^*(t)\}}(\alpha_2(t)-\alpha_1(t)\exp\{\theta^\mathsf{T}\psi^*(t)\})dt = 0.$$

The condition (1.7) just means that $\theta = 0$ is not the solution to the limiting estimating equation, i.e., the estimate in the smooth model does not asymptotically fall to the null model. In other words, (1.7) says that the choice of the basis functions $\varphi_1, \ldots, \varphi_d$ is not completely wrong in the sense that at least some of them contributes to the approximation of η.

Under the nested subsets search, it would be possible to let the maximum dimension d tend to infinity at a suitable rate and obtain the consistency against arbitrary alternatives. Note that for the all subsets procedure with $d_0 = 0$, it is necessary to fix d, see a thorough analysis of Claeskens and Hjort (2004) in the classical goodness-of-fit setting.

Next, the limit distribution of the test statistics is investigated under a sequence of local alternatives. Consider local alternatives of the form $\alpha_2(t) = \alpha_1(t)\exp\{n^{-1/2}\eta(t)\}$, where η is a bounded function.

Theorem 1.3. *Under the sequence of local alternatives*

$$\lambda_{j,i}(t) = Y_{j,i}(t)\alpha(t)\exp\{n^{-1/2}\eta(t)Z_{j,i}\}$$

the logrank process $n^{-1/2}U_0(t)$ converges weakly in $D[0,\tau]$ to the Gaussian process $\mu_0(t)+V_0(t)$, where the martingale V_0 is given in Section 1.2 and the mean function is

$$\mu_0(t) = \int_0^t \eta(s)\frac{\bar{y}_1(s)\bar{y}_2(s)}{\bar{y}(s)}\alpha(s)ds = \int_0^t \eta(s)d\sigma_0(s).$$

The process $n^{-1/2}U(t)$ converges to $\mu(t)+V(t)$ with $\mu(t) = \int_0^t \psi(s)d\mu_0(s)$, and, consequently, the statistic T_d is asymptotically distributed as a chi-squared variable with d degrees of freedom and noncentrality parameter $\mu(\tau)^\mathsf{T}\sigma(\tau)^{-1}\mu(\tau)$. The statistic T_S of the adaptive test converges weakly to

$$\max\{(\mu_C(\tau)+V_C(\tau))^\mathsf{T}\sigma_{CC}(\tau)^{-1}(\mu_C(\tau)+V_C(\tau)) : C \in \mathcal{S}, |C| = d^*\}.$$

Proof. The convergence of the logrank process is shown in Andersen et al. (1993, Section V.2.3). The convergence of $n^{-1/2}U$ and T_d is an immediate consequence. The results for data-driven tests follow from the fact that also along the sequence of local alternatives all variants of Schwarz's rule asymptotically concentrate in sets of the minimal dimension d^*. □

For nested subsets with $d_0 = 0$ the test behaves asymptotically under local alternatives like the directional test based on the first basis functions (the logrank test). Nevertheless, simulations in Section 1.6 show that in finite samples the performance of this data-driven test is much better than the performance of the logrank test in situations not suitable for the logrank test (nonproportional hazards). (Note also that the data-driven test is consistent against the same alternatives as tests with all d functions.) The local behaviour was the motivation of Janssen (2003) for including high priority basis functions. Such tests (both with nested subsets and all subsets) behave asymptotically like the smooth test with d_0 basis functions.

For the class of nested subsets, Ducharme and Ledwina (2003) were able to derive a deep efficiency result for the data-driven rank test of Janic-Wróblewska and Ledwina (2000) for the classical (uncensored) two-sample problem. Their study was motivated by the phenomenon that the behaviour of the data-driven test empirically appears quite different from what would be expected from the $n^{-1/2}$ asymptotics (as the behaviour of the test based on the first direction). They let d converge to infinity suitably and consider sequences of alternatives converging to the hypothesis at rates slower than $n^{-1/2}$ while at the same time the significance level converges to zero as n grows in such a way that the limit of the power is nontrivial. Then they prove that the asymptotic power of the data-driven rank test is the same as the limiting power of Neyman–Pearson tests constructed for this sequence of alternatives and significance levels. I do not have such a result for the context of survival analysis but it will be seen in simulations in the next section (and for other problems throughout this thesis) that the behaviour of data-driven tests with nested subsets in finite samples is remarkably different from the performance of directional tests.

1.6 Numerical study

1.6.1 General information

I conducted simulations to examine the behaviour of the proposed tests and compare them with some of existing two-sample procedures. I considered one situation satisfying the null hypothesis and several alternative configurations with hazard differences of various kind.

Random numbers were generated using the Mersenne–Twister generator implemented in R. 20 000 Monte Carlo runs were performed under the null hypothesis, and 5000 for alternative situations. Smooth tests were used with the Legendre polynomial basis; the time transformation g was based on the distribution function.

1.6.2 Results on level

I examine the behaviour of the test procedures under the null hypothesis. I repeatedly generated two samples of unit exponential variables, censored them by independently generated uniform variables and performed the fixed-dimensional smooth test and both nested subsets

Table 1.1: Estimated sizes of fixed-dimensional and adaptive tests on the nominal level 5 % with asymptotic critical values. The distribution of survival times is unit exponential. Estimates based on 20 000 replications (standard deviation 0.0015).

	$d = 4$, $d_0 = 0$				$d = 7$, $d_0 = 4$	
(n_1, n_2)	T_d (χ^2_4)	T_S^{nested} (χ^2_1)	T_S^{nested} (two-term)	T_S^{all} ($\max \chi^2_1$)	T_S^{nested} (χ^2_4)	T_S^{all} (χ^2_4)
			Censoring U(0, 10) (10 %)			
(25, 25)	0.0664	0.1265	0.0695	0.0701	0.0945	0.1167
(50, 50)	0.0608	0.0960	0.0560	0.0600	0.0860	0.1084
(100, 100)	0.0600	0.0766	0.0554	0.0528	0.0772	0.0987
(200, 200)	0.0537	0.0656	0.0528	0.0516	0.0662	0.0848
(15, 35)	0.0769	0.1359	0.0770	0.0740	0.1158	0.1368
(30, 70)	0.0698	0.0960	0.0586	0.0586	0.0986	0.1215
(60, 140)	0.0636	0.0814	0.0604	0.0548	0.0832	0.1026
(120, 280)	0.0609	0.0695	0.0550	0.0519	0.0760	0.0944
			Censoring U(0, 2) (43 %)			
(25, 25)	0.0512	0.1132	0.0554	0.0620	0.0717	0.0898
(50, 50)	0.0548	0.0911	0.0536	0.0602	0.0710	0.0915
(100, 100)	0.0516	0.0701	0.0512	0.0542	0.0664	0.0854
(200, 200)	0.0522	0.0642	0.0490	0.0508	0.0632	0.0792
(15, 35)	0.0654	0.1238	0.0664	0.0734	0.0948	0.1129
(30, 70)	0.0572	0.0916	0.0542	0.0616	0.0785	0.0978
(60, 140)	0.0560	0.0762	0.0569	0.0566	0.0726	0.0899
(120, 280)	0.0534	0.0668	0.0535	0.0518	0.0654	0.0815

and all subsets adaptive tests with and without specifying high priority basis functions. Various sample sizes n_1, n_2 and two censoring distributions (U(0, 10) and U(0, 2)) were considered. The tests were performed on the nominal level 5 % using asymptotic critical values.

Table 1.1 provides empirical sizes. It is seen that the tests often exceed the nominal level. There are two sources of inaccuracy: bad performance of the asymptotic normal approximation for the score vector in some cases, and slow convergence of selection criteria to the smallest dimension.

First, we may observe that when the censoring is light the fixed-dimensional test T_d is anticonservative. A similar phenomenon could be observed for $G^{0,\gamma}$ tests (especially with $\gamma > 0$). Like these tests, our tests give some weight to late differences too.

A second, apparently more serious problem concerns data-driven tests. It is mainly seen for the nested subsets test with $d_0 = 0$ and for both variants with $d_0 > 0$ (here $d_0 = 4$) that the $\chi^2_{d^*}$ approximation is unacceptable even for the sample size 400. The reason of the inaccuracy is the slow convergence of the selection criterion to the smallest dimension. Table 1.2 reports estimated selection probabilities for sets of three smallest dimensions. It shows that the concentration of $|S|$ in d^* is insufficient for small samples. There is an exception: the criterion with all subsets with $d_0 = 0$ is more concentrated in smallest (one-dimensional) sets and the asymptotic distribution (i.e., the maximum of χ^2_1 variables) performs much better (the size is comparable to the size of the test with fixed dimension). This is not so surprising because in

Table 1.2: Estimated selection probabilities for subsets with dimension d^*, d^*+1, d^*+2 (three smallest dimensions) under the null hypothesis (unit exponential). Censoring U(0, 2), various sample sizes $n = n_1 + n_2$ (with $n_1 = n_2$). Estimates based on 20 000 replications (standard deviation at most 0.0035).

	$d = 4$, $d_0 = 0$			$d = 7$, $d_0 = 4$		
n	$\|S\| = 1$	$\|S\| = 2$	$\|S\| = 3$	$\|S\| = 4$	$\|S\| = 5$	$\|S\| = 6$
			Nested subsets			
50	0.9366	0.0518	0.0094	0.9482	0.0444	0.0063
100	0.9607	0.0325	0.0060	0.9662	0.0294	0.0038
200	0.9770	0.0198	0.0026	0.9758	0.0219	0.0021
400	0.9848	0.0134	0.0016	0.9844	0.0145	0.0008
			All subsets			
50	0.9804	0.0176	0.0014	0.8857	0.1101	0.0038
100	0.9891	0.0099	0.0009	0.9184	0.0796	0.0020
200	0.9938	0.0056	0.0003	0.9428	0.0559	0.0012
400	0.9967	0.0030	0.0003	0.9594	0.0400	0.0006

this case the selection rule is asymptotically concentrated in d one-dimensional sets whereas with the other classes of subsets the rule asymptotically selects one set (of dimension d^*). For the nested subsets criterion with $d_0 = 0$ we have the two-term approximation of Section 1.4. It successfully removes the problem of the slow convergence of S, the size is then similar to the size of the fixed-dimensional test (see Table 1.1).

To make the inference valid I use the permutation principle (Neuhaus, 1993). It assumes that the pairs $(N_{j,i}, Y_{j,i})$ (or $(T_{j,i}, \delta_{j,i})$ for survival data) are independent identically distributed under the null hypothesis and the distribution of the test statistic is exchangeable (permutation invariant). Hence, in the survival context, the censoring distributions in the two samples should be equal. The test is then exact. If the censoring distributions differ, the permutation procedure is valid asymptotically. Neuhaus (1993) and Heller and Venkatraman (1996) show that the permutation method remains reliable when the assumption of equal censoring distributions is not satisfied. I used the permutation test with 2000 random permutations which seemed enough as the rejection probability lied between 0.0470 and 0.0530 for all of the situations of Table 1.1. Note that alternatively instead of permutations (sampling without replacement) one may use the bootstrap (sampling with replacement); bootstrap results both under the null and under alternatives were very similar to permutation results.

I do not report detailed results on the null behaviour of other two-sample tests. Just note that they often do not have size close to the nominal level when the asymptotic distribution is used. The weighted logrank $G^{0,\gamma}$ tests with normal approximation exceed the level especially with light or without censoring. On the contrary, the Kolmogorov–Smirnov and other tests based on the logrank process are conservative. (For these tests one may alternatively use the simulation approximation of Lin, Wei and Ying (1993) which removes the conservatism to some extent.) Therefore, hereafter in simulations of power, all of the tests are performed using the permutation principle (with 2000 permutations).

1.6.3 Alternative configurations

Several configurations found in the literature were analysed. I report only situations previously studied by other authors not to be suspect of designing the study to favour the tests I propose. I investigated many other situations with similar conclusions. Configurations I–IV correspond to I–IV of Fleming, Harrington and O'Sullivan (1987), Configuration V corresponds to IV of Lee (1996). I admit that some of these alternatives may look somewhat peculiar when written in terms of hazard functions, they, however, do not look so when survival functions are plotted (see Figure 1.1). Here are their forms.
Configuration I (proportional hazards).

$$\alpha_1(t) = 1, \qquad \alpha_2(t) = 2.$$

Configuration II (late difference).

$$\alpha_1(t) = 2 \times 1_{[0,0.5)}(t) + 4 \times 1_{[0.5,\infty)}(t), \qquad \alpha_2(t) = 2 \times 1_{[0,0.5)}(t) + 0.4 \times 1_{[0.5,\infty)}(t).$$

Configuration III (middle/early difference).

$$\begin{aligned}\alpha_1(t) &= 2 \times 1_{[0,0.1)}(t) + 3 \times 1_{[0.1,0.4)}(t) + 0.75 \times 1_{[0.4,0.7)}(t) + 1_{[0.7,\infty)}(t),\\ \alpha_2(t) &= 2 \times 1_{[0,0.1)}(t) + 0.75 \times 1_{[0.1,0.4)}(t) + 3 \times 1_{[0.4,0.7)}(t) + 1_{[0.7,\infty)}(t).\end{aligned}$$

Configuration IV (early difference).

$$\begin{aligned}\alpha_1(t) &= 3 \times 1_{[0,0.2)}(t) + 0.75 \times 1_{[0.2,0.4)}(t) + 1_{[0.4,\infty)}(t),\\ \alpha_2(t) &= 0.75 \times 1_{[0,0.2)}(t) + 3 \times 1_{[0.2,0.4)}(t) + 1_{[0.4,\infty)}(t).\end{aligned}$$

Configuration V (middle difference).

$$\begin{aligned}\alpha_1(t) &= 2 \times 1_{[0,0.2)}(t) + 3 \times 1_{[0.2,0.6)}(t) + 0.75 \times 1_{[0.6,0.9)}(t) + 1_{[0.9,\infty)}(t),\\ \alpha_2(t) &= 2 \times 1_{[0,0.2)}(t) + 0.75 \times 1_{[0.2,0.6)}(t) + 5 \times 1_{[0.6,0.9)}(t) + 1_{[0.9,\infty)}(t).\end{aligned}$$

The sample size was always 100 (each group 50), the censoring distribution was uniform on $(0, 2)$ giving censoring rates from 28 % to 38 %.

1.6.4 Comparison of fixed-dimensional and various data-driven tests

In Table 1.3, several variants of smooth tests are compared in two of the considered situations. For Configuration I (proportional hazards) the best test is with $d = 1$, hence it is not surprising that increasing d decreases the power (other basis functions than $\varphi_1 \equiv 1$ are superfluous). In Configuration III we can see that the power decreases for $d > 6$ ($d = 6$ gives the best power because the hazard ratio is rather complicated and its description requires several functions). Now let us see how the data-driven tests with various classes of subsets behave.

First consider $d_0 = 0$ (no basis functions of primary interest). The all subsets version seems to suffer from the same problem as the test with a fixed dimension: when d is too high, the power decays. This is caused by the dependence of the null distribution of the test statistic on d. The nested subsets criterion gives stable power for various values of d in both configurations. In Configuration I, this test has higher power than the other tests because the selection rule mostly selects the smallest set containing only the intercept which describes

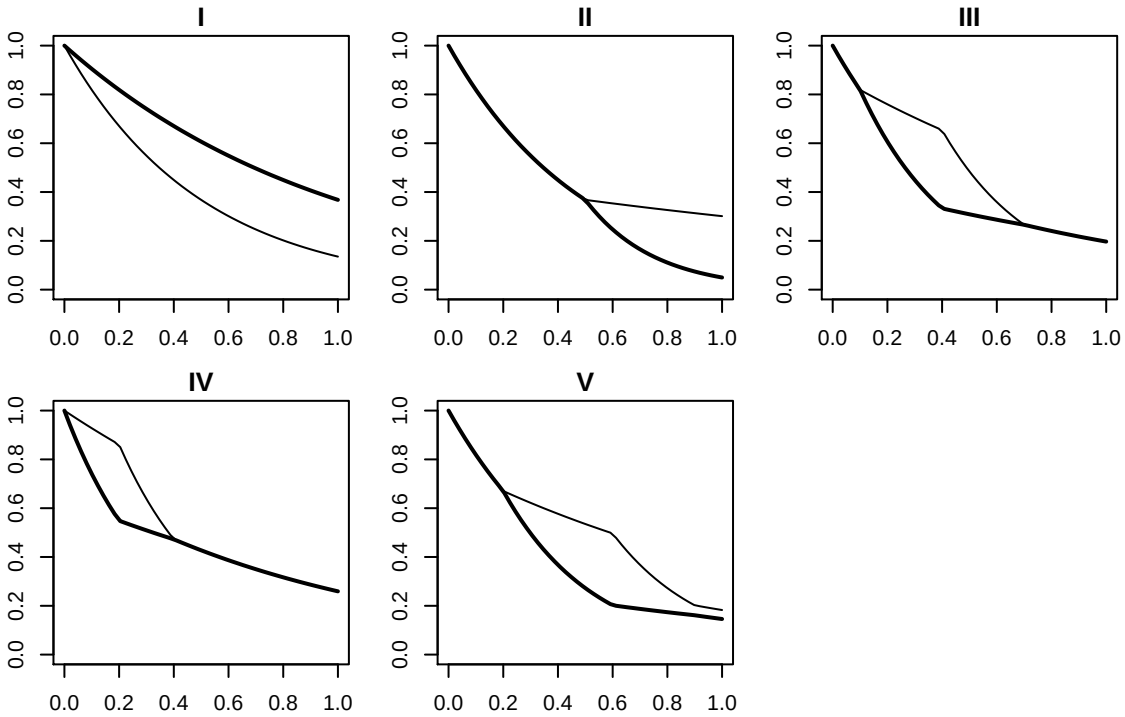


Figure 1.1: Survival functions S_1 (thick lines) and S_2 (thin) under the simulation scenarios I to V

Table 1.3: Comparison of power for fixed-dimensional and various data-driven tests with various values of d and d_0. Censoring $\mathrm{U}(0, 2)$, sample sizes $n_1 = n_2 = 50$, nominal level 5 % (permutation test). Estimates based on 5000 replications (standard deviation at most 0.007).

		$d_0 = 0$		$d_0 = 4$	
d	T_d	T_S^{nested}	T_S^{all}	T_S^{nested}	T_S^{all}
		Configuration I			
4	0.603	0.677	0.639	—	—
6	0.553	0.680	0.585	0.565	0.555
8	0.526	0.677	0.544	0.562	0.515
10	0.502	0.677	0.508	0.564	0.478
12	0.488	0.675	0.480	0.562	0.458
14	0.465	0.678	0.459	0.563	0.433
		Configuration III			
4	0.713	0.507	0.678	—	—
6	0.817	0.539	0.763	0.771	0.789
8	0.804	0.542	0.751	0.767	0.764
10	0.784	0.542	0.734	0.770	0.750
12	0.753	0.541	0.721	0.770	0.729
14	0.736	0.541	0.707	0.767	0.708

the data well. On the other hand, in Configuration III, this test has lower power than the test with fixed dimension because the selection rule penalises complicated functions which in this rather complicated play an important role.

Now let $d_0 = 4$. Again, the power of the test with nested subsets is stable. For Configuration I it is now lower than with $d_0 = 0$ (because now the BIC concentrates in the set $\{1, 2, 3, 4\}$ instead of $\{1\}$), for Configuration III the power is higher (four basis functions catch the hazard difference better than smaller sets often selected with $d_0 = 0$). The power of the all subsets test decreases with increasing d which is somewhat surprising since, unlike with $d_0 = 0$, the null distribution now does not depend on d. It perhaps may be explained by the slow convergence of the criterion to the four-dimensional set as already seen in Table 1.2 (with the all subsets criterion the limiting set $\{1, 2, 3, 4\}$ has much more competitors than with nested sets).

To summarise, mainly the nested subsets approach helps to avoid the use of too large d. Specifying several basis functions of high priority seems to be a good strategy when the hazard ratio is expected to be complicated.

1.6.5 Comparison with other tests

Let us compare Neyman's smooth tests with other two-sample methods. Firstly, I consider weighted logrank tests with $G^{0,0}$, $G^{2,0}$, $G^{0,2}$ and $G^{2,2}$ weights and tests combining these four statistics (the statistic T^{sum} equals the sum of absolute values of these four standardised statistic while $T^{\max}$ equals their maximum).

Secondly, functionals of the whole path of the logrank process leading to the Kolmogorov–Smirnov (KS), Cramér–von Mises (CM) and Anderson–Darling (AD) type tests are considered. They are of two kinds: those using the untransformed process (denoted KS-W, CM-W, AD-W) and those using the process transformed in a way similar to the construction of the Hall–Wellner confidence bands (these tests are denoted KS-B, CM-B, AD-B). The former test process $U_0(t)/\hat{\sigma}_0(\tau)^{1/2}$ is asymptotically distributed as a Brownian motion in transformed time, specifically $W(h_0(t))$, where W stands for the standard Brownian motion and $h_0(t) = \sigma_0(t)/\sigma_0(\tau)$ is an increasing continuous mapping of $[0, \tau]$ on $[0, 1]$. The latter kind of tests uses the process

$$\frac{U_0(t)/\hat{\sigma}_0(\tau)^{1/2}}{1 + \hat{\sigma}_0(t)/\hat{\sigma}_0(\tau)}$$

converging to $W(h_0(t))/[1 + h_0(t)]$, which is well-known to have the same distribution as $B(h_0(t)/[1 + h_0(t)])$, where B denotes the standard Brownian bridge. For details see Section V.4.1 of Andersen et al. (1993).

All of these tests are performed as two-sided since Neyman's tests are naturally two-sided. In all situations, the permutation approch is employed.

Table 1.4 presents estimated powers for the above situations I–V. Before looking at the results we should realize what we expect from versatile tests. Certainly, it is impossible to hope that they will outperform all other methods in all situations. Rather, one may wish to have tests whose behaviour is not bad under a broad range of situations, that is, one seeks tests with robust power.

To assess robustness of power I computed a quantity, presented in the last column of the table, as follows. For each situation (each column) the ratio of the power of each test and the power of the best test (in the column) is computed. Then for each test (each row) the minimum of these ratios is presented as a measure of robustness of power. In other words,

the last column contains row minima of standardised powers (where standardisation means division by column maxima).

Table 1.4: Comparison of power of various two-sample tests. Censoring $\mathrm{U}(0, 2)$, sample sizes $n_1 = n_2 = 50$, nominal level 5 % (permutation test). Estimates based on 5000 replications (standard deviation at most 0.007).

	Configuration					Robustness
	I	II	III	IV	V	of power
$G^{0,0}$	0.792	0.340	0.232	0.134	0.306	0.161
$G^{2,0}$	0.655	0.056	0.357	0.562	0.173	0.064
$G^{0,2}$	0.517	0.876	0.070	0.097	0.121	0.087
$G^{2,2}$	0.676	0.302	0.241	0.135	0.588	0.162
T^{sum}	0.782	0.500	0.216	0.147	0.344	0.177
$T^{\max}$	0.734	0.796	0.319	0.474	0.457	0.397
KS-W	0.772	0.274	0.556	0.558	0.468	0.313
KS-B	0.718	0.195	0.620	0.811	0.449	0.223
CM-W	0.705	0.058	0.478	0.511	0.319	0.066
CM-B	0.620	0.050	0.425	0.737	0.192	0.057
AD-W	0.646	0.052	0.433	0.696	0.228	0.059
AD-B	0.575	0.050	0.356	0.767	0.151	0.057
T_d $(d = 4)$	0.601	0.854	0.713	0.761	0.547	0.759
T_d $(d = 8)$	0.526	0.796	0.803	0.832	0.672	0.664
T_S^{nested} $(d = 8, d_0 = 0)$	0.677	0.803	0.541	0.733	0.419	0.624
T_S^{all} $(d = 8, d_0 = 0)$	0.547	0.687	0.751	0.788	0.609	0.691
T_S^{nested} $(d = 8, d_0 = 4)$	0.563	0.826	0.769	0.797	0.638	0.711
T_S^{all} $(d = 8, d_0 = 4)$	0.516	0.793	0.766	0.785	0.619	0.652

As expected, directional tests $G^{\rho,\gamma}$ have very low robustness scores because they perform excellently in situations they are designed for but often do very badly for other situations. Among versatile tests previously proposed in the literature the test $T^{\max}$ as well as the Kolmogorov–Smirnov type tests (mainly KS-W) appear to have more stable power. The behaviour of power of smooth tests proposed in this chapter seems much better (regarding stability over various alternatives) than of the other versatile tests. Smooth tests of course often lose against some of the other tests but not so much as the other tests sometimes do. These conclusions should be looked at with caution as they are based on the limited set of Configurations I–V. However, I studied various other situations but never found smooth tests completely failing.

A closer look at results for various configurations reveals several findings. Tests employing the untransformed logrank process detect late differences better than those with the transformed process and vice versa (because the Hall–Wellner transformation downweights late differences). Surprisingly, the integral type tests completely failed in Configuration II, thus they do not appear as versatile as expected. In the next section I attempt to explain this phenomenon by the Karhunen–Loève decomposition of the test process and principal components analysis of the integral statistic.

The behaviour of $T^{\max}$ in Configurations III and V is interesting. Situation III was termed

a middle difference in Fleming et al. (1987) but it rather seems to be something between a middle and early difference as is seen from powers of $G^{\rho,\gamma}$ tests. None of $G^{0,0}$, $G^{2,0}$, $G^{2,2}$ tests clearly dominates but $T^{\max}$ must choose one of them. Hence, the $T^{\max}$ test loses some power compared to smooth tests which can combine more than one direction (no matter that directions are given by different functions for the two approaches). In Configuration V studied by Lee (1996) the difference is more clearly in the middle ($G^{2,2}$ is much better than the other $G^{\rho,\gamma}$ tests), so $T^{\max}$ does better. In this regard, the behaviour of the adaptive test of Pecková and Fleming (2003) is expected to be similar as this test is also forced to select one of the weighted logrank statistics (simulations in their paper show that the power of this test in most cases lies above the power of $T^{\max}$ and below the best power in the cluster).

In view of the simulation results, it is difficult to give a general recommendation whether a fixed or some (and which) of data-driven procedures should be used. It seems impossible to say that one method should always be preferred to the other. But hopefully the study presented here gives an image of their behaviour which may be helpful in a particular application. In general, I think that the test with a relatively small (like $d = 3$) fixed dimension will do well quite often. The adaptive choice with nested subsets with $d_0 = 1$ (or perhaps $d_0 = 2$) will slightly prefer simpler alternatives, which reflects the natural idea that simple situations occur in reality more often than complicated ones. Anyway, differences between variants of the test do not appear large. All versions of smooth tests provide a procedure with power that seems to be more stable than power of other methods.

1.7 Principal components analysis of the integral test

This section gives some insight into the performance of the Cramér–von Mises test by means of the Karhunen–Loève decomposition and the principal components analysis. Principal components of integral-type L^2 statistics were studied in the traditional goodness-of-fit situation by, e.g., Anderson and Darling (1952), Durbin and Knott (1972), Durbin, Knott and Taylor (1975), or in a nonparametric regression setting for instance by Stute (1997).

The Cramér–von Mises statistic is the L^2 norm of the logrank process of the form

$$\int_0^\tau U_0(t)^2/\hat{\sigma}_0(\tau)d\hat{h}_0(t),$$

where $\hat{h}_0(t) = \hat{\sigma}_0(t)/\hat{\sigma}_0(\tau)$ is the empirical counterpart of $h_0(t) = \sigma_0(t)/\sigma_0(\tau)$. The standardised process $U_0(t)/\hat{\sigma}_0(\tau)^{1/2}$ is asymptotically distributed as the time-transformed Brownian motion $W(h_0(t))$. The limiting process admits the Karhunen–Loève expansion (e.g., Ash and Gardner, 1975, Section 1.4)

$$W_0(h_0(t)) = \sum_{j=1}^{\infty} \lambda_j^{1/2} b_j l_j(h_0(t)),$$

where the series converges in L^2, uniformly in $t \in [0, \tau]$. Here $l_j(u) = \sqrt{2}\sin((j - \frac{1}{2})\pi u), u \in [0, 1]$ are orthonormal eigenfunctions and $\lambda_j = 1/((j - \frac{1}{2})\pi)^2$ corresponding eigenvalues of the covariance kernel $k(u, v) = u \wedge v$ of the standard Brownian motion on $[0, 1]$, i.e., $\int_0^1 k(u, v)l_j(v)dv = \lambda_j l_j(u)$. The standardised Fourier coefficients

$$b_j = \lambda_j^{-1/2} \int_0^\tau W(h_0(t))l_j(h_0(t))dh_0(t), \quad j = 1, 2, \ldots$$

are independent standard normal variables. Their empirical counterparts

$$B_j = \lambda_j^{-1/2} \int_0^\tau U_0(t)/\hat{\sigma}_0(\tau)^{1/2} l_j(\hat{h}_0(t)) d\hat{h}_0(t)$$

may be inverted using integration by parts to obtain

$$B_j = \int_0^\tau f_j(\hat{h}_0(t)) dU_0(t)/\hat{\sigma}_0(\tau)^{1/2},$$

where $f_j(u) = \sqrt{2}\cos((j-\frac{1}{2})\pi u)$. Observe that the variable B_j is a standardised weighted logrank statistic with the weight $f_j(\hat{h}_0(t))$.

In the limit, the Cramér–von Mises statistic admits the representation

$$\int_0^\tau W(h_0(t))^2 dh_0(t) = \sum_{j=1}^\infty \lambda_j b_j^2,$$

where the principal components b_j^2 are independent χ_1^2-distributed. In finite samples, this corresponds to

$$\int_0^\tau U_0(t)^2/\hat{\sigma}_0(\tau) d\hat{h}_0(t) = \sum_{j=1}^\infty \lambda_j B_j^2.$$

This is an infinite weighted sum of squares of asymptotically independent weighted logrank statistics. Under local alternatives of Section 1.5 the limit components b_j^2 are independent χ_1^2-distributed with noncentrality parameter $[\int_0^\tau f_j(h_0(t))\eta(t)dh_0(t)]^2$. Thus each of the components B_j^2 reflects a specific departure from the hypothesis. The weights $f_j(h_0(t)) = \sqrt{2}\cos((j-\frac{1}{2})\pi h_0(t))$ are equal to 0 at τ and hence they downweight late differences. This corresponds to the bad performance of the Cramér–von Mises test for Configuration II (late difference) in Table 1.4.

From the above expansion a relation between the Cramér–von Mises test and Neyman's smooth test may be seen. The weights λ_j rapidly (quadratically) decrease, and hence components corresponding to higher frequencies of the hazard ratio are downweighted. When this series is truncated and summands are given equal weights, we arrive at $\sum_{j=1}^d B_j^2$, which asymptotically coincides with the statistic T_d of Neyman's smooth test with the basis functions $\varphi_j = f_j$ (by orthonormality of these functions). Such a test distributes its power evenly among the first d directions $f_1, \ldots, f_d$. Note, however, that different bases (e.g., $\varphi_j(u) = \sqrt{2}\cos(j\pi u)$ or Legendre polynomials) instead of $\varphi_j = f_j$ are preferred for Neyman's tests because they do not downweight late differences.

1.8 Illustration

Stablein and Koutrouvelis (1985) studied data from a trial comparing two types of treatment of gastric cancer: chemotherapy versus chemotherapy combined with radiotherapy. There were 45 patients in each group (2 and 6 were censored, respectively). This dataset is a popular example when methods for crossing curves are dealt with, see, for instance, Yang and Prentice (2005) who developed a two-sample model to accommodate crossing curves, and Bagdonavičius, Levuliene, Nikulin and Zdorova-Cheminade (2004) who proposed a test for non-location alternatives. Figure 1.2 displays crossing survival curves. It is not obvious

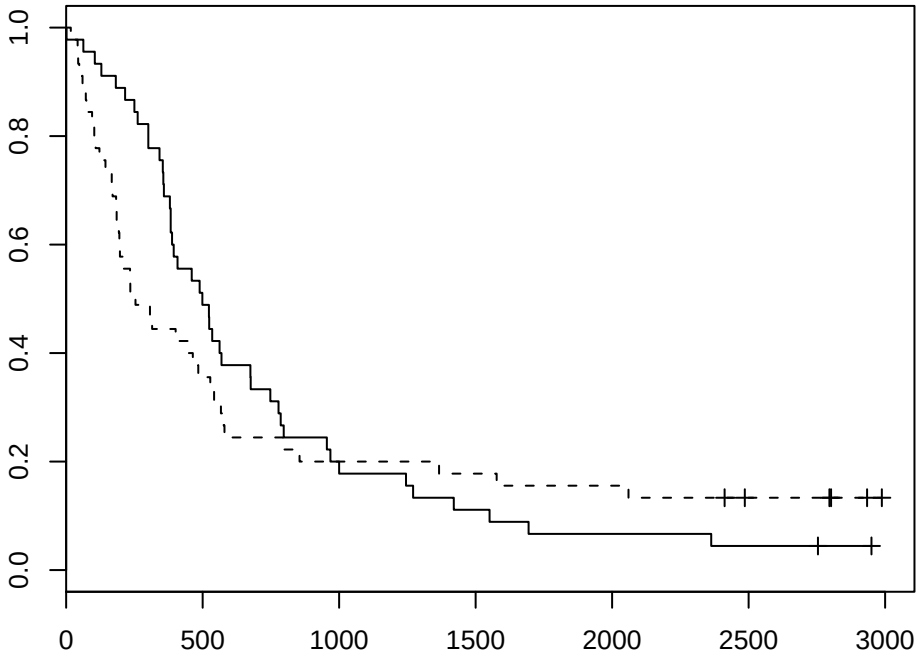


Figure 1.2: Kaplan–Meier estimates for chemotherapy (solid) and chemotherapy plus radiotherapy (dashed) for the gastric cancer data. Survival times in days.

against which alternative we should test, hence a versatile test is handy. On the conventional level 5 % all of them reject the hypothesis of no difference between the two treatments.

The test statistic of Neyman's smooth test with $d = 8$ Legendre polynomials (of order $0, \ldots, 7$) is 17.55 with p-value 0.023 (based on 5000 permutations). The selection rule with $d_0 = 4$ selects the smallest possible set $\{1, 2, 3, 4\}$ for both nested and all subsets search. The test statistic equals 13.59 with $p = 0.018$ for nested subsets and $p = 0.03$ for all subsets. If no functions of primary interest are specified ($d_0 = 0$) then the nested subsets criterion selects $\{1, 2\}$ with the statistic 13.45 and p-value 0.005 while the all subsets rule gives the set $\{2\}$, statistic 13.32 and p-value 0.01.

The tests $G^{0,0}$, $G^{2,0}$, $G^{0,2}$, $G^{2,2}$ have statistics 0.47 ($p = 0.637$), 2.59 (0.009), 1.99 (0.053), 0.41 (0.684), respectively. The p-value of the maximal statistic 2.59 is 0.021. The value of the KS-W statistic is 2.20 with $p = 0.047$, the KS-B statistic equals 1.58 with $p = 0.008$.

2 Comparison of two samples in the presence of competing risks

Summary

In this chapter a method is developed for comparison of two samples of survival data with competing risks. In competing risks data the probability that a failure from a particular cause in the presence of other risks of failure occurs by some time is summarised by the cumulative incidence function. A new test is proposed for the hypothesis that cumulative incidence functions for a particular type of failure are equal in two samples. The procedure is based on Neyman's idea of smooth tests. The new test has stable power against a wider spectrum of alternatives than tests previously proposed in the literature. In particular, the method exhibits much better power in situations with crossing curves. Asymptotic results, simulations and a real example are presented.

2.1 Introduction

In competing risks situations, individuals may fail from one of K causes. Observations consist of the failure time and the cause of failure. Formally, let $R \geq 0$ be the survival time and let $\varepsilon \in \{1, \ldots, K\}$ be the cause of death. There are two main quantities describing the occurrence of events of type k: the cause-specific hazard rate (crude transition intensity)

$$\alpha(t,k) = \lim_{\Delta \to 0} \frac{\Pr(t \leq R < t + \Delta, \varepsilon = k | R \geq t)}{\Delta},$$

and the cumulative incidence function

$$F(t,k) = \Pr(R \leq t, \varepsilon = k) = \int_0^t S(s)\alpha(s,k)ds,$$

where $S(t) = \Pr(R > t)$ is the overall survival function. Observations are allowed to be right-censored, that is, we actually observe (T, δ), $T = R \wedge C$, $\delta = \varepsilon 1[R \leq C]$, where the censoring time C is independent of R and ε.

The cause-specific hazard is the instantaneous rate of failure of the particular type. However, unlike in the classical survival analysis with one type of event, the integral of this rate does not translate to an interpretable survival probability. Specifically, the function $\exp\{-\int_0^t \alpha(s,k)ds\}$ cannot be interpreted as the survival function of the latent time of failure of type k (latent times are notional times, the actual failure time is the minimum of them) unless unverifiable assumptions like independence of latent times are adopted (see Section 8.2 of Kalbfleisch and Prentice, 2002). Without unverifiable assumptions the marginal distributions of the latent times are not estimable (not identifiable). Cause-specific hazards as well as cumulative incidence functions are estimable. Cumulative incidence functions are often preferred to cause-specific hazards because they have direct probability interpretations.

I consider two samples of competing risks data: $(T_{j,i}, \delta_{j,i})$, $j = 1, 2$, $i = 1, \ldots, n_j$. The goal is to compare the occurrence of failures from one particular cause, say 1. Without loss of generality I assume that the number of possible endpoints K is 2; all event types different from 1, which are not of interest, may be merged in type 2. The comparison of two samples can be done either in terms of the cause-specific hazards $\alpha_j(\cdot, 1)$ or in terms of the cumulative incidence functions $F_j(\cdot, 1)$. Here I focus on the cumulative incidences. Note that the hypotheses $\alpha_1(\cdot, 1) = \alpha_2(\cdot, 1)$ and $F_1(\cdot, 1) = F_2(\cdot, 1)$ are not equivalent, which is explained, e.g., by Gray (1988) or Lin (1997) and vividly illustrated by simulations of Bajorunaite and Klein (2007). For instance, if there is no difference between the cause-specific hazards for cause 1, say $\alpha_1(t, 1) = \alpha_2(t, 1) = 1$, but the cause-specific hazards for cause 2 differ, say $\alpha_1(t, 2) = 1$ and $\alpha_2(t, 2) = 0.5$, then also the cumulative incidence functions for cause 1 differ, $F_1(t, 1) < F_2(t, 1)$. Examples with the effect of a treatment on the cause-specific hazard opposite to its effect on the cumulative incidence function can be found (e.g., Example 8.1 of Kalbfleisch and Prentice, 2002).

Cause-specific hazard rates can be compared by standard methods (e.g., the logrank test) by working with failures from the other causes as with censored observations. On the other hand, the comparison of cumulative incidence curves requires special methods. In this chapter, I develop a method for testing the nonparametric null hypothesis $F_1(\cdot, 1) = F_2(\cdot, 1)$ against the alternative that these functions differ.

Several tests have been previously proposed for this task. Gray (1988) developed a class of tests based on weighted integrals with respect to the difference of estimated cumulative subdistribution hazard functions corresponding to the subdistributions $F_j(t, 1)$, defined as $\Gamma_j(t, k) = -\log(1 - F_j(t, k))$. The test statistic follows the form

$$\int_0^\tau L(t)(d\hat{\Gamma}_2(t, 1) - d\hat{\Gamma}_1(t, 1)),$$

where $L(t)$ is a weight function (a predictable process), $\hat{\Gamma}_j(t, 1)$ are consistent estimators of $\Gamma_j(t, 1)$, and $\tau < \infty$ is the end of the observation period $[0, \tau]$. These tests are good for detection of ordered subdistribution hazards $\gamma_j(t, 1) = d\Gamma_j(t, 1)/dt$ but may fail to detect crossing subdistribution hazards. Note that these statistics are similar to weighted logrank statistics for the traditional situation with one type of failure. For this similarity, I call this test the logrank-type test, which should not be confused with the ordinary logrank test applied to cause-specific hazards: the ordinary logrank test compares cause-specific hazards whereas Gray's logrank-type test compares subdistribution hazards.

Another test was proposed by Pepe (1991) who used the integrated difference of estimates of the cumulative incidence functions

$$\int_0^\tau (\hat{F}_2(t, 1) - \hat{F}_1(t, 1))dt.$$

This test often possesses good power against ordered cumulative incidence functions but may be less powerful when these curves cross. Note that this test is not a rank test.

Pepe's integral test and Gray's logrank-type test lose some power against alternatives with crossing curves (cumulative incidences or subdistribution hazards) because positive differences in some part of the observation period are negated by negative differences in another part. Lin (1997) suggested to use a Kolmogorov–Smirnov type test based on the supremum of the absolute value of the difference of estimated cumulative incidence functions $\sup_{t\in[0,\tau]} |\hat{F}_2(t, 1)-$

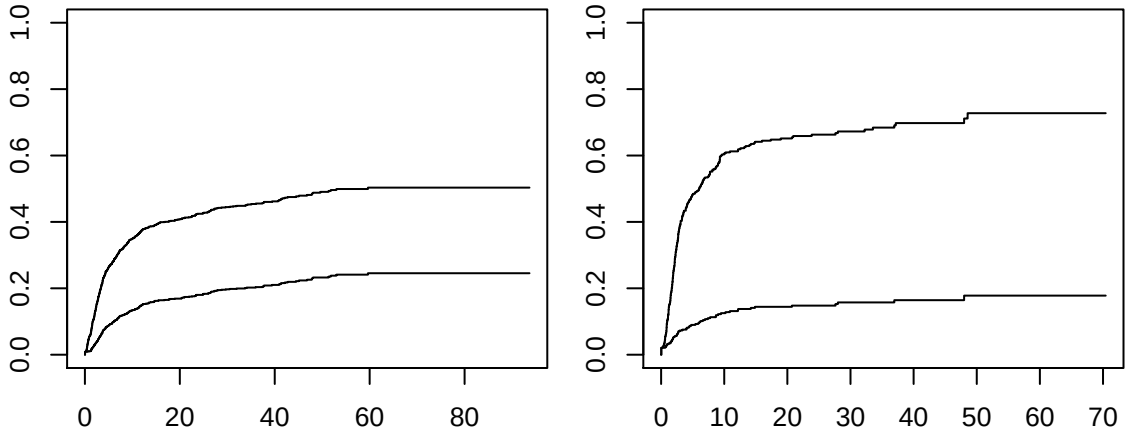


Figure 2.1: Summary plots for HLA-identical sibling donors (left panel) and HLA-matched unrelated donors (right panel). In each plot, the lower curve is the cumulative incidence of relapse, the upper curve is the sum of the relapse and death in remission cumulative incidences, the complement of the upper curve is the disease free survival probability. Time from the bone marrow transplantation is in months.

$\hat{F}_1(t,1)|$. While such a test is theoretically consistent against any alternative, its power is low quite often.

Therefore, it is desirable to develop a test which is good at detecting a spectrum of practically relevant alternatives (including crossing situations not covered by tests of Gray (1988) and Pepe (1991)) with better performance than the supremum test proposed by Lin (1997). This chapter deals with the class of Neyman's smooth tests.

Situations with complicated departures from the hypothesis (such as crossing functions) occur in real applications. As an example I consider data from a bone marrow transplant study discussed by Bajorunaite and Klein (2007). The treatment of leukaemia by the bone marrow transplantation may fail from one of two causes: recurrence of the disease (relapse), and death in remission (treatment-related death). There are two groups of patients to be compared: 1224 individuals with a human leukocyte antigen (HLA) identical sibling donor, and 383 with an HLA-matched unrelated donor. Summary plots for the two groups are displayed in Figure 2.1. Figure 2.2 compares estimates of cumulative incidence functions for both samples for each type of treatment failure. Numerical results of Section 2.4 show that (some of) tests sensitive against ordered alternatives do not detect the difference between relapse cumulative incidence curves, which contradicts the visual impression. The reason is that the relapse cumulative incidence functions for these two groups cross.

The structure of this chapter is as follows. In Section 2.2 the Neyman-type smooth test is constructed. Section 2.3 presents results of a simulation study. Results for the bone marrow transplant data are reported in Section 2.4. Asymptotic results are formulated and proved in Section 2.5.

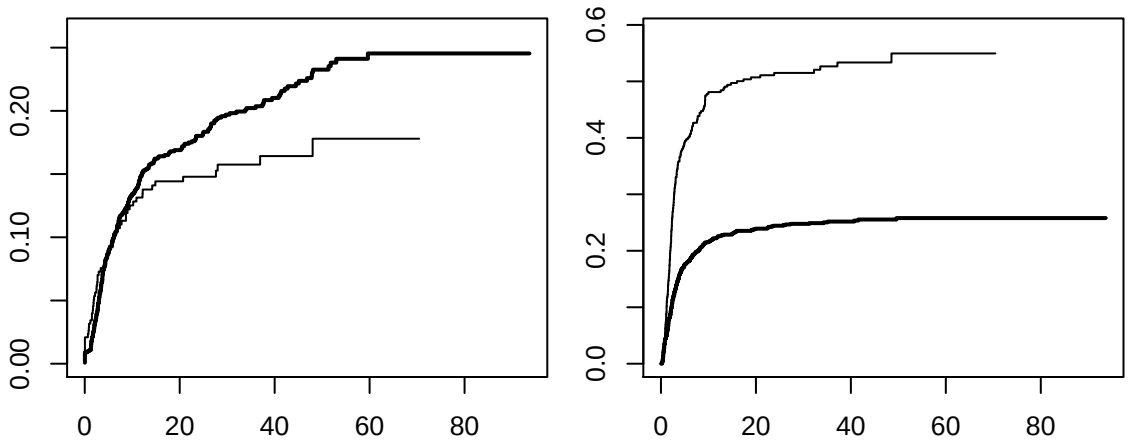


Figure 2.2: Cumulative incidence functions for relapse (left panel) and death in remission (right panel) for HLA-identical sibling donors (thick lines) and HLA-matched unrelated donors (thin lines). Time from the bone marrow transplantation is in months.

2.2 Neyman's embedding and development of the score test

In this section I show how Neyman's embedding idea can be applied in the two-sample competing risks situation. Neyman's smooth goodness-of-fit procedure is based on embedding the null hypothesis in a 'smooth' alternative model described by a finite number of parameters.

Traditionally (Rayner and Best, 1989), the embedding is formulated in terms of densities. In the goodness-of-fit problem of testing the simple hypothesis that the data come from a distribution with density $f = f_0$ the null hypothesis is embedded into the d-dimensional alternative

$$f(x;\theta) = f_0(x)\exp\{\theta^{\mathsf{T}}\varphi(F_0(x)) - c(\theta)\}, \quad x \in \mathbb{R}, \tag{2.1}$$

where $\varphi(u) = (\varphi_1(u), \ldots, \varphi_d(u))^{\mathsf{T}}, u \in [0,1]$ are some square integrable basis functions, $c(\theta) = \log \int_{\mathbb{R}} f_0(x)\exp\{\theta^{\mathsf{T}}\varphi(F_0(x))\}dx$ is a normalising constant and F_0 the distribution function corresponding to f_0. The general alternative $f \neq f_0$ is replaced by $\theta \neq 0$. Neyman's test is the score test of $\theta = 0$ in the above model. Note that the function $f(x;\theta)$ is properly normalised, i.e., any value θ gives rise to a possible alternative distribution.

In the standard (single endpoint) survival context the embedding is most conveniently achieved in terms of hazard functions. The alternative takes the form

$$\alpha(t;\theta) = \alpha_0(t)\exp\{\theta^{\mathsf{T}}\varphi(F_0(t))\}, \quad t \geq 0. \tag{2.2}$$

See Peña (1998a) for details and extensions to composite hypotheses. Notice that in this formulation there is no normalising constant. The integral of a hazard function may be arbitrary positive. Therefore, any value of θ yields a well-defined hazard rate.

Let us turn to competing risks problems. First, consider a one sample situation with a simple hypothesis of the full specification of the cumulative incidence function for failures of

type 1, that is $F(\cdot,1) = F_0(\cdot,1)$. Recall that $F(t,1) = \int_0^t f(s,1)ds$, where $f(t,1) = S(t)\alpha(t,1)$. The first idea is to formulate a smooth alternative in terms of $f(t,1)$. This strategy is, however, infeasible because $f(t,1)$ is a subdistribution density. That is, its integral $F(\infty,1)$ is neither fixed nor unbounded ($F(t,1)$ is a subdistribution function, hence $F(\infty,1)$ may be anything between 0 and 1). The smooth alternative cannot be expressed in the form (2.1) since there is no normalising constant. On the other hand, we cannot use an unnormalised form like (2.2) because of the upper bound 1 for the integral of a subdensity. The cause-specific hazard $\alpha(t,1)$ could be embedded similarly to (2.2) but hypotheses about cause-specific hazards are not equivalent to hypotheses about cumulative incidence functions. However, there is a characteristic suitable for embedding: the subdistribution hazard function $\gamma(t,1)$ defined as

$$\gamma(t,k) = \frac{d}{dt}\Gamma(t,k) = \frac{f(t,k)}{1-F(t,k)},$$

where $\Gamma(t,k) = -\log(1-F(t,k))$. The functions $\gamma(t,1)$ and $\Gamma(t,1)$ may be seen as the hazard rate and the cumulative hazard function of the subdistribution $F(t,k)$ of the improper random variable $\tilde{R}^{(k)}$ defined by $\tilde{R}^{(k)} = R$ if $\varepsilon = k$, $\tilde{R}^{(k)} = \infty$ otherwise. Due to the one-to-one correspondence between $F(t,k)$ and $\gamma(t,k)$ hypotheses about $F(t,1)$ and $\gamma(t,1)$ are equivalent.

Now consider the two-sample hypothesis $F_1(\cdot,1) = F_2(\cdot,1) = F_0(\cdot,1)$, equivalently $\gamma_1(\cdot,1) = \gamma_2(\cdot,1) = \gamma_0(\cdot,1)$. The null model is viewed as a submodel of

$$\gamma_2(t,1) = \gamma_1(t,1)\exp\{\theta^{\mathsf{T}}\psi(t)\}.$$

The logarithm of the subdistribution hazard ratio is expressed as a linear combination of some functions. Here $\psi_l(t)$, $t \in [0,\tau]$, $l = 1,\dots,d$ are of the form $\psi_l(t) = \varphi_l(F_0(t,1)/F_0(\tau,1))$, where $\varphi_l(u), u \in [0,1]$ are some linearly independent basis functions (for example, orthogonal Legendre polynomials of order $0,1,\dots,d-1$, or cosines $\sqrt{2}\cos((l-1)\pi u)$).

I shall develop a score test of the hypothesis $\theta = 0$ versus $\theta \neq 0$.

The observations $(T_{j,i}, \delta_{j,i})$ can be represented by marked point processes as follows. For $k \in \{1,2\}$ denote $N_{j,i}(t,k) = 1[T_{j,i} \le t, \delta_{j,i} = k]$, the counting process counting events of type k up to time t on the ith individual of the jth sample. Its intensity process is $\lambda_{j,i}(t,k) = Y_{j,i}(t)\alpha_j(t,k)$, where $Y_{j,i}(t) = 1[T_{j,i} \ge t]$ is the risk indicator process.

Estimators used in the following derivations are

$$\hat{F}_j(t,k) = \int_0^t \hat{S}_j(s-)\frac{d\bar{N}_j(s,k)}{\bar{Y}_j(s)}, \qquad \hat{\Gamma}_j(t,k) = \int_0^t \frac{d\hat{F}_j(s,k)}{1-\hat{F}_j(s-,k)} = \int_0^t \frac{d\bar{N}_j(s,k)}{\bar{R}_j(s,k)},$$

where $\hat{S}_j$ is the Kaplan–Meier estimator of S_j, $\bar{N}_j = \sum_{i=1}^{n_j} N_{j,i}$, $\bar{Y}_j = \sum_{i=1}^{n_j} Y_{j,i}$, and $\bar{R}_j(t,k) = \bar{Y}_j(t)(1-\hat{F}_j(t-,k))/\hat{S}_j(t-)$. Under the null hypothesis, there are consistent pooled sample estimators

$$\hat{F}_0(t,1) = \int_0^t \frac{d\bar{N}_1(s,1) + d\bar{N}_2(s,1)}{\bar{Y}_1(s)/\hat{S}_1(s-) + \bar{Y}_2(s)/\hat{S}_2(s-)}, \qquad \hat{\Gamma}_0(t,1) = \int_0^t \frac{d\bar{N}_1(s,1) + d\bar{N}_2(s,1)}{\bar{R}_1(s,1) + \bar{R}_2(s,1)}$$

introduced by Gray (1988, formulae (2.11) and (2.5)).

The logarithm of the likelihood takes the form

$$\sum_{j=1}^{2}\sum_{i=1}^{n_j}\int_0^\tau \sum_{k=1}^{2}\log(\lambda_{j,i}(t,k))dN_{j,i}(t,k) - \sum_{j=1}^{2}\sum_{i=1}^{n_j}\int_0^\tau \sum_{k=1}^{2}\lambda_{j,i}(t,k)dt$$
$$= \sum_{j=1}^{2}\int_0^\tau \sum_{k=1}^{2}\log(\alpha_j(t,k))d\bar{N}_j(t,k) - \sum_{j=1}^{2}\int_0^\tau \bar{Y}_j(t)\sum_{k=1}^{2}\alpha_j(t,k)dt.$$

Using the relation $\gamma_j(t,k) = S_j(t)\alpha_j(t,k)/(1-F_j(t,k))$ we get

$$\sum_{j=1}^{2}\int_0^\tau \sum_{k=1}^{2}\log(\gamma_j(t,k))d\bar{N}_j(t,k) + \sum_{j=1}^{2}\int_0^\tau \sum_{k=1}^{2}\log\Big(\frac{1-F_j(t,k)}{S_j(t)}\Big)d\bar{N}_j(t,k)$$
$$- \sum_{j=1}^{2}\int_0^\tau \bar{Y}_j(t)\gamma_j(t,1)\frac{1-F_j(t,1)}{S_j(t)}dt - \sum_{j=1}^{2}\int_0^\tau \bar{Y}_j(t)\gamma_j(t,2)\frac{1-F_j(t,2)}{S_j(t)}dt.$$

In the above expression not only $\gamma_j(t,1)$ but also $F_j(t,k)$ and $S_j(t)$ depend on the parameter θ. However, things simplify when we replace $F_j(t,k)$ and $S_j(t)$ by their consistent estimators $\hat{F}_j(t,k)$ and $\hat{S}_j(t)$. Then only the first and third term contain θ. Taking derivatives with respect to θ we arrive at

$$\int_0^\tau \psi(t)\Big[d\bar{N}_2(t,1) - \bar{Y}_2(t)\frac{1-\hat{F}_j(t,2)}{\hat{S}_j(t)}\exp\{\theta^\mathsf{T}\psi(t)\}\gamma_0(t,1)dt\Big].$$

Since $\gamma_0(t,1)$ is unknown, we use its null Breslow-type estimator $\hat{\Gamma}_0(t,1)$. Finally, we obtain the score vector

$$U(\tau) = \int_0^\tau \psi(t)\Big[d\bar{N}_2(t,1) - \bar{R}_2(t,1)\exp\{\theta^\mathsf{T}\psi(t)\}\frac{d\bar{N}_1(t,1)+d\bar{N}_2(t,1)}{\bar{R}_1(t,1)+\bar{R}_2(t,1)}\Big]$$
$$= \int_0^\tau L(t)(d\hat{\Gamma}_2(t,1) - d\hat{\Gamma}_1(t,1)),$$

where

$$L(t) = \psi(t)\frac{\bar{R}_1(t,1)\bar{R}_2(t,1)}{\bar{R}_1(t,1)+\bar{R}_2(t,1)}.$$

This vector resembles a partial likelihood score vector but here the risksets are reweighted. It is a vector of weighted logrank-type statistics for comparing subdistribution hazard functions. When $U(\tau)$ is one-dimensional ($d=1$) and $\varphi_1(t)=1$, it agrees with the statistic of Gray (1988).

In practice, the time transformation in $\psi(t) = \varphi(F_0(t,1)/F_0(\tau,1))$ must be estimated, i.e., $\hat{F}_0(\cdot,1)$ replaces $F_0(\cdot,1)$.

In Theorem 2.1 in Section 2.5, I show that under the null hypothesis the score vector $n^{-1/2}U(\tau)$ (where $n=n_1+n_2$) is asymptotically normal with mean zero and variance matrix which is consistently estimated by $n^{-1}\hat{\sigma}(\tau,\tau)$ given in that theorem. Consequently, the quadratic score statistic $T = U(\tau)^\mathsf{T}\hat{\sigma}(\tau,\tau)^{-1}U(\tau)$ is asymptotically χ^2 distributed with d degrees of freedom. Significantly large values of T contradict the hypothesis.

Theorem 2.2 provides a condition for consistency of the test. Unlike the Kolmogorov–Smirnov test of Lin (1997), this test is not consistent against an arbitrary alternative. Alternatives that will be rejected with probability converging to 1 are given by the choice of the

basis functions. The consistency condition essentially says that the test is consistent unless the basis functions are orthogonal to the true distribution in certain sense. If we take three or four basis functions, the true difference between subdistributions would have to be quite unusually complicated for the test to be inconsistent. For instance, Legendre polynomials of order $0, 1, 2$ will be able to detect proportional subdistribution hazards as well as monotone and nonmonotone (convex or concave) subdistribution hazard log-ratios.

The number of the basis functions can be chosen with the help of Schwarz's selection rule as discussed in Chapter 1, though here I work with a fixed number of basis function.

2.3 Simulations

I conducted a simulation study to investigate properties of the proposed test and compare them with other existing tests both under the null hypothesis and under alternatives. Datasets of size 100 (50 in each sample) are generated. The number of Monte Carlo runs for each model is 20 000 under the hypothesis and 5000 under alternatives. The data generating procedure in the jth sample is as follows: first, the failure type is set to k with probability $p_{jk} = F_j(\infty, k)$, $k \in \{1, 2\}$, then the failure time is drawn from the conditional distribution $F_j(t, k)/p_{jk}$, and, finally, the observation is possibly censored. Censoring times are generated from the uniform distribution on $[0, c]$ (values of c are reported below).

Neyman-type tests proposed in this chapter are performed with $d = 3$ Legendre polynomials (of order $0, 1, 2$). Lin's Kolmogorov–Smirnov-type test uses 1000 resampled test processes (see Lin (1997) for the description of the simulation procedure), with pooled sample null estimates of $F_0(t, 1)$ which was found by Bajorunaite and Klein (2007) to give a more accurate approximation than with individual samples estimators. In Pepe's integral test I use the asymptotic normal approximation with the martingale-based variance estimator derived by Bajorunaite and Klein (2007). Lebesgue integrals involved in this statistic are computed from 0 to $\tau = c$.

In the first set of simulations I investigate the behaviour of tests under H_0. Cumulative incidence functions take the form $F_0(t, 1) = p_1(1-e^{-t})$, $F_j(t, 2) = (1-p_1)(1-e^{-t})$. Probability p_1 of failure type 1 is 0.25, 0.5 and 0.75. The paremeter of the censoring distribution is $c = 7$ (about 15 % censored in all of the situations) and $c = 2.5$ (about 37 %). Rejection probabilities on the nominal level 5 % are reported in Table 2.1. The accuracy of the level of the smooth test and Gray's logrank-type test appears acceptable. The Kolmogorov–Smirnov-type test and integral tests tend to be slightly conservative when the censoring rate is low (see Bajorunaite and Klein (2007) for a detailed analysis).

Next I consider five alternative configurations. Figure 2.3 shows subdistribution characteristics for type 1 events.

Configuration A.

$$F_1(t, 1) = 0.5(1 - e^{-t}), \qquad F_2(t, 1) = 1 - (1 - F_1(t, 1))^2,$$
$$F_1(t, 2) = 0.5(1 - e^{-t}), \qquad F_2(t, 2) = 0.25(1 - e^{-t}).$$

This situation was considered by Gray (1988). Subdistribution hazard rates for events of type 1 are proportional. With $c = 4$ there is 23 % censored observations.

Configuration B.

$$F_1(t, 1) = \frac{\pi(1 - e^{-t})}{1 - \pi + \pi(1 - e^{-t})}, \qquad F_2(t, 1) = \frac{\pi\theta(1 - e^{-t})}{1 - \pi + \pi\theta(1 - e^{-t})},$$

Table 2.1: Empirical levels on the nominal level of 5 %. Figures based on 20 000 Monte Carlo repetitions (standard deviation 0.0015).

	$c = 7$ (15 % censored)			$c = 2.5$ (37 % censored)		
	p_1			p_1		
	0.25	0.5	0.75	0.25	0.5	0.75
Neyman	0.0608	0.0579	0.0450	0.0562	0.0682	0.0602
KS	0.0300	0.0160	0.0196	0.0491	0.0495	0.0557
Pepe	0.0333	0.0205	0.0172	0.0468	0.0464	0.0476
Gray	0.0582	0.0590	0.0623	0.0550	0.0562	0.0572

$$F_1(t,2) = (1-\pi)(1-e^{-t}), \qquad F_2(t,2) = (1-\pi)(1-e^{-t})/(1-\pi+\pi\theta).$$

With $\pi = 0.5$, $\theta = e^{0.75}$, this is Model 1 of Bajorunaite and Klein (2007). Set $c = 3$ (24 % censoring).

Configuration C.

$$F_j(t,1) = p_{j1}(1-e^{-t/p_{j1}}), \qquad F_j(t,2) = (1-p_{j1})(1-e^{-t/p_{j1}}).$$

In this situation considered by Bajorunaite and Klein (2007, Model 3) cause-specific intensities of events of type 1 are the same (equal to 1) in both samples. I take $(p_{11}, p_{21}) = (0.3, 0.7)$ and $c = 2$ (giving 28 % censoring).

Configuration D.

$$F_1(t,1) = \tfrac{2}{3}(1-e^{-t}), \qquad F_2(t,1) = \tfrac{2}{3}(1-e^{-t^{0.5}}),$$
$$F_1(t,2) = \tfrac{1}{3}(1-e^{-0.8t}), \qquad F_2(t,2) = \tfrac{1}{3}(1-e^{-1.2t}).$$

Peng and Fine (2007) used this situation in which both cumulative incidence curves and subdistribution hazards cross. For $c = 4$, 26 % is censored.

Configuration E.

$$F_1(t,1) = 0.3(1-e^{-1.3t^{2.4}}), \qquad F_2(t,1) = 0.4(1-e^{-t}),$$
$$F_1(t,2) = 0.7(1-e^{-t}), \qquad F_2(t,2) = 0.6(1-e^{-t}).$$

In this model subdistribution hazards cross but cumulative incidence functions are ordered. The censoring proportion is 25 % with $c = 4$.

Empirical powers are summarised in Table 2.2. In Configurations A and B, Gray's logrank-type test performs best which is not surprising as the subdistribution hazards differ largely (at least at the beginning) and do not cross. The Neyman-type smooth test and Pepe's integral test do not lose much. In Configuration C, the difference of the subdistributions appears later in time which makes the smooth test slightly more powerful but the difference is not dramatic. In Configuration D, the logrank-type test and the integral test fail because functions they are based on cross. Similarly, in the last Configuration E, Neyman's test outperforms the other methods (here the bad performance of Pepe's test is rather unexpected as the subdistribution functions are ordered). Interestingly, Lin's Kolmogorov–Smirnov-type test has also low power for D and E despite its 'omnibus' property (consistency against any alternative).

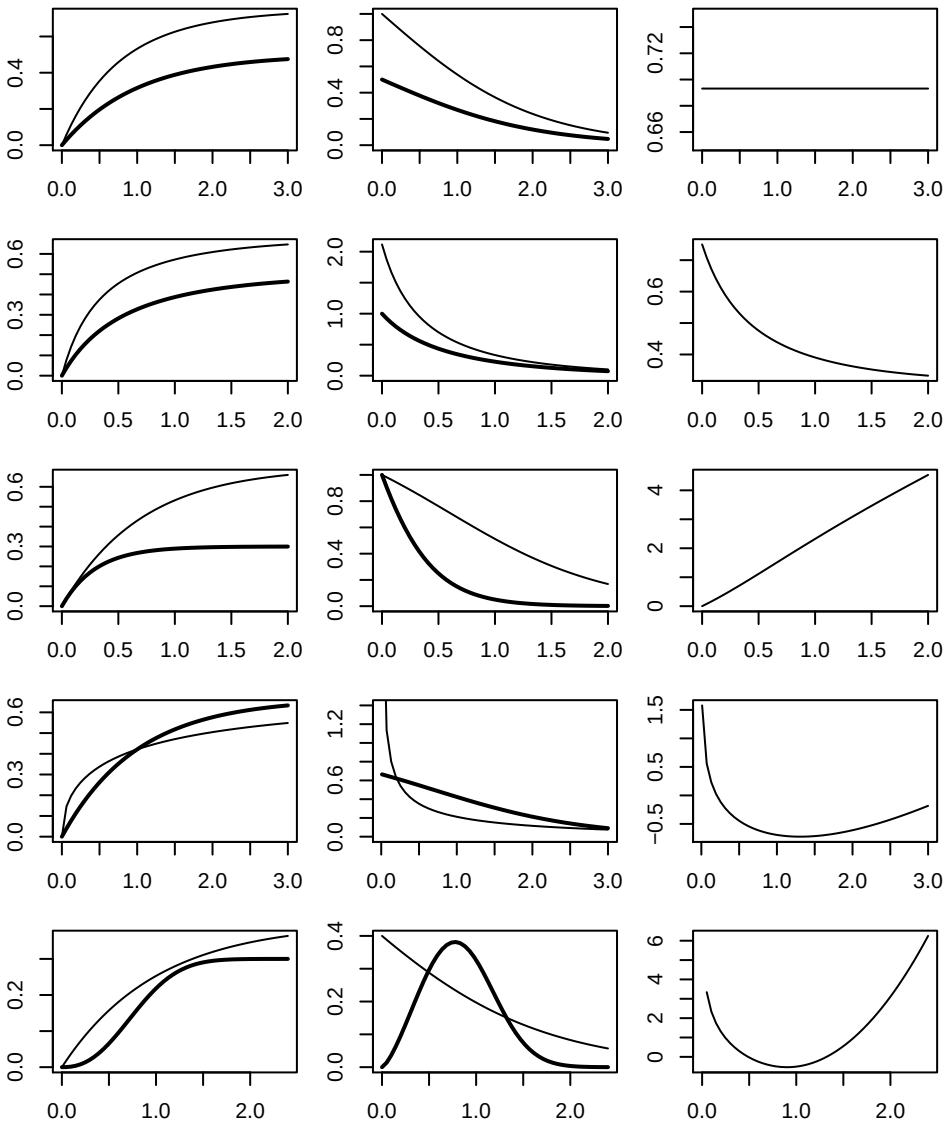


Figure 2.3: Alternative configurations A–E (from top to bottom). On each row: Left plot: cumulative incidence functions $F_1(t, 1)$ (thick line) and $F_2(t, 1)$ (thin line). Middle plot: corresponding subdistribution hazards. Right plot: logarithm of subdistribution hazard ratios.

Table 2.2: Estimated powers on the nominal level 5 %. Based on 5000 Monte Carlo repetitions (standard deviation 0.007).

	A	B	C	D	E
Neyman	0.507	0.334	0.672	0.531	0.521
KS	0.344	0.338	0.477	0.130	0.199
Pepe	0.593	0.417	0.581	0.053	0.161
Gray	0.690	0.505	0.550	0.058	0.235

The proposed smooth tests appear to have stable power over a wide range of realistic alternatives and thus seem to be virtually 'omnibus'. They did not 'completely fail' in any of practically relevant situations considered in this simulation. Smooth test procedures can be recommended as an alternative to the supremum-type procedure for their better performance in complicated situations. Moreover, in simple situations, they do not lose much compared to other existing methods.

2.4 Real example

In the bone marrow transplant study introduced in Section 2.1 there were two competing risks (relapse and death in remission) and two groups of patients (with an HLA-identical sibling donor and with an HLA-matched unrelated donor).

First consider the risk of relapse. Gray's logrank-type test does not lead to rejection of the hypothesis of equal relapse cumulative incidence functions for the two kinds of donors: the test statistic is -1.66 with p-value 0.098. Pepe's test statistic is -2.09 with $p = 0.036$, Lin's Kolmogorov–Smirnov test statistic equals 0.0672 with p-value 0.027 (based on 5000 simulated processes). In contrast to these marginally significant results, the Neyman-type smooth test with $d = 3$ basis functions strongly rejects the hypothesis with the test statistic 14.1, $p = 0.0028$. This conclusion agrees with the visual impression drawn from Figure 2.2.

Cumulative incidence functions for the risk of death in remission are ordered and well separated. All of the tests discussed here reject with $p < 0.0001$.

2.5 Asymptotic results and proofs

The following results are derived under the assumption that $n^{-1}\bar{Y}_1, n^{-1}\bar{Y}_2$ converge in probability uniformly on $[0, \tau]$ to some functions $\bar{y}_1, \bar{y}_2$, respectively, which are bounded away from zero. By the Glivenko–Cantelli theorem, this basically holds with $\bar{y}_j(t) = a_j S_j(t)(1 - G_j(t))$, where $G_j(t)$ is the distribution function of censoring times in the jth group and $a_j = \lim_{n\to\infty} n_j/n$, provided $S_j(\tau) > 0$, $1 - G_j(\tau) > 0$ and $a_j \in (0, 1)$. Denote by $\bar{r}_j(t, 1) = \bar{y}_j(t)(1 - F_j(t, 1))/S_j(t)$ uniform limits in probability of $n^{-1}\bar{R}_j(t, 1)$ (the limits exist by uniform consistency of $\hat{S}_j(t)$ and $\hat{F}_j(t, 1)$).

Theorem 2.1 (Asymptotic distribution)**.** *The score vector $n^{-1/2}U(\tau)$ is under the null hypothesis $F_1(\cdot, 1) = F_2(\cdot, 1)$ asymptotically (as $n \to \infty$) distributed as a zero-mean Gaussian vector with covariance matrix whose consistent estimator is $n^{-1}\hat{\sigma}(\tau, \tau) = n^{-1}(\hat{\sigma}_1(\tau, \tau) + \hat{\sigma}_2(\tau, \tau))$ with $\hat{\sigma}_j(s, t)$ given by* (2.3) *below.*

Proof. Under the null hypothesis $\Gamma_1(t,1) = \Gamma_2(t,1)$ the score process may be expressed as $U(t) = U_2(t) - U_1(t)$ with

$$U_j(t) = \int_0^t L(s)(d\hat{\Gamma}_j(s,1) - d\Gamma_j(s,1)) = \int_0^t L(s)\Big(\frac{d\hat{F}_j(s,1)}{1-\hat{F}_j(s-,1)} - \frac{dF_j(s,1)}{1-F_j(s-,1)}\Big).$$

Hence the process U_j is a function of $\hat{F}_j(\cdot,1)$, and thus the asymptotic distribution of $n^{-1/2}U_j$ can be inferred from that of $n^{1/2}(\hat{F}_j(\cdot,1) - F_j(\cdot,1))$ by a use of the functional delta method (see Section II.8 of Andersen et al. (1993), or Chapter 3.9 of van der Vaart and Wellner (1996)).

Asymptotic results for $n^{1/2}(\hat{F}_j(\cdot,1) - F_j(\cdot,1))$ were derived by Lin (1997). He found a martingale representation in the form $n^{1/2}(\hat{F}_j(\cdot,1) - F_j(\cdot,1)) = n^{1/2}V_j + o_P(1)$, where

$$n^{1/2}V_j(t) = n^{1/2}\int_0^t \frac{1-F_j(s,2)}{\bar{Y}_j(s)}d\bar{M}_j(s,1) + n^{1/2}\int_0^t \frac{F_j(s,1)}{\bar{Y}_j(s)}d\bar{M}_j(s,2) \\ - n^{1/2}F_j(t,1)\int_0^t \frac{d\bar{M}_j(s,1) + d\bar{M}_j(s,2)}{\bar{Y}_j(s)}$$

(beware of misprints in eq. (2) in Lin's paper) with counting process martingales $\bar{M}_j(t,k) = \bar{N}_j(t,k) - \int_0^t \bar{Y}_j(s)\alpha_j(s,k)ds$. By the martingale central limit theorem (Andersen et al., 1993, Theorem II.5.1) this process converges weakly to a zero-mean continuous Gaussian process with covariance function consistently estimated by $n\hat{\rho}(s,t)$ with

$$\hat{\rho}(s,t) = \int_0^{s\wedge t} \frac{(1-\hat{F}_j(u,2))^2}{\bar{Y}_j(u)^2}d\bar{N}_j(u,1) + \int_0^{s\wedge t} \frac{\hat{F}_j(u,1)^2}{\bar{Y}_j(u)^2}d\bar{N}_j(u,2) \\ + \hat{F}_j(s,1)\hat{F}_j(t,1)\int_0^{s\wedge t} \frac{d\bar{N}_j(u,1) + d\bar{N}_j(u,2)}{\bar{Y}_j(u)^2} \\ - (\hat{F}_j(s,1) + \hat{F}_j(t,1))\Big(\int_0^{s\wedge t} \frac{1-\hat{F}_j(u,2)}{\bar{Y}_j(u)^2}d\bar{N}_j(u,1) + \int_0^{s\wedge t} \frac{\hat{F}_j(u,1)}{\bar{Y}_j(u)^2}d\bar{N}_j(u,2)\Big).$$

The delta method (together with the chain rule, and Proposition II.8.6 of Andersen et al. (1993) or Lemma 3.9.17 of van der Vaart and Wellner (1996)) yields that $n^{-1/2}U_j(\cdot)$ is asymptotically equivalent to

$$\int_0^\cdot n^{-1}L(s)\frac{1}{1-F_j(s,1)}n^{1/2}dV_j(s) + \int_0^\cdot n^{-1}L(s)\frac{n^{1/2}V_j(s)}{(1-F_j(s,1))^2}dF_j(s,1).$$

By integration by parts this equals

$$\int_0^\cdot n^{-1}Q_j(s)n^{1/2}dV_j(s) + n^{-1}H_j^L(\cdot)n^{1/2}V_j(\cdot),$$

where

$$Q_j(t) = \frac{L(t)}{1-F_j(t,1)} - H_j^L(t),$$

$$H_j^L(t) = \int_0^t L(s)H_j(s,1), \qquad H_j(t,1) = \int_0^t \frac{dF_j(s,1)}{(1-F_j(s,1))^2}$$

($H_j(t,1)$ is equal to $F_j(t,1)/(1-F_j(t,1))$ and can be viewed as a subdistribution odds function). Using the fact that $n^{-1}L$ uniformly converges in probability (by the assumption of the theorem and by consistency of estimators involved in L) we obtain that $n^{-1/2}U_j$ is asymptotically a zero-mean continuous Gaussian process. Its limiting covariance matrix function can be consistently estimated by

$$n^{-1}\hat\sigma_j(s,t) = n^{-1}\int_0^s\int_0^t \hat Q_j(u)\hat Q_j(v)\hat\rho_j(du,dv) + n^{-1}\int_0^s \hat Q_j(u)\hat\rho_j(du,t)\hat H_j^L(t)^\mathsf{T} + n^{-1}\hat H_j^L(s)\int_0^t \hat Q_j(v)^\mathsf{T}\hat\rho_j(s,dv) + n^{-1}\hat H_j^L(s)\hat\rho_j(s,t)\hat H_j^L(t)^\mathsf{T} \quad (2.3)$$

($\hat Q_j$ and $\hat H_j^L$ are defined like Q_j and H_j^L with $\hat F_j(\cdot,1)$ in place of $F_j(\cdot,1)$).

The weak convergence result achieved above holds jointly for $n^{-1/2}(U_1,U_2)$ with U_1,U_2 asymptotically independent (by independence of the two samples). Finally, the score vector $n^{-1/2}U(\tau) = n^{-1/2}U_2(\tau) - n^{-1/2}U_1(\tau)$ converges in distribution to a zero-mean normal vector with variance matrix which is consistently estimated by $n^{-1}(\hat\sigma_1(\tau,\tau)+\hat\sigma_2(\tau,\tau))$. □

Theorem 2.2 (Consistency)**.** *Assume that $\gamma_1(t,1)\neq\gamma_2(t,1)$ on a non-null set. Denote by $F_0^*(\cdot,1)$ the limit in probability of $\hat F_0(\cdot,1)$ and set $\psi^*(t)=\varphi(F_0^*(t,1)/F_0^*(\tau,1))$. Then the test is consistent provided the condition*

$$\int_0^\tau \psi^*(t)\frac{\bar r_1(t,1)\bar r_2(t,1)}{\bar r_1(t,1)+\bar r_2(t,1)}(\gamma_2(t,1)-\gamma_1(t,1))dt \neq 0 \quad (2.4)$$

(at least one component is nonzero) is satisfied.

Proof. We have

$$n^{-1}U(\tau) = n^{-1}\int_0^\tau L(t)(d\hat\Gamma_2(t,1)-d\Gamma_2(t,1)) - n^{-1}\int_0^\tau L(t)(d\hat\Gamma_1(t,1)-d\Gamma_1(t,1)) + n^{-1}\int_0^\tau L(t)(d\Gamma_2(t,1)-d\Gamma_1(t,1)).$$

The first two terms on the right-hand side converge in probability to zero by the previous proof, and the last term converges in probability to the left-hand side of (2.4). Therefore, $n^{-1}T$ converges in probability to a nonzero quantity, and the test rejecting for large T is consistent. □

3 Testing fit of two-sample proportional rate transformation models

Summary

Transformation models for two samples of censored data are considered. Main examples are the proportional hazards and proportional odds model. The key assumption of these models is that the ratio of transformation rates (e.g., hazard rates or odds rates) is constant in time. A method of verification of this proportionality assumption is developed. The proposed procedure is based on the idea of Neyman's smooth test and its data-driven version. The method is suitable for detecting monotonic as well as nonmonotonic ratios of rates.

3.1 Introduction

This chapter deals with simple models for two samples (e.g., the control and treatment group) of survival data under random censorship. Various models have been proposed in the literature to describe the situation when the survival distributions in two samples differ. The aim of this chapter is to develop new methods of assessment of fit for one class of these models, proportional rate models.

The most frequent model is the proportional hazards model which assumes that the ratio of the hazard rates $\alpha_1(t), \alpha_2(t)$ is constant over time, that is there exists a real constant β such that $\alpha_2(t)/\alpha_1(t) = e^\beta$ for all t. The effect of treatment on the failure rate remains the same in the course of time. In some situations the effect of treatment decays for large times and hazard rates converge to each other. A popular model for this situation is the proportional odds model. Let $S_k(t) = e^{-A_k(t)}$ be the survival function in the kth sample, $k = 1, 2$. Denote $\Gamma_k(t) = (1 - S_k(t))/S_k(t)$ the odds function giving the odds of dying before time t versus surviving up to t. The proportional odds model assumes $\Gamma_2(t)/\Gamma_1(t) = e^\beta$ for all times. A common feature of these two main examples is that they assume constancy of the ratio of some functions. It is important to check this assumption.

These two models are considered within a wider class of semiparametric linear transformation models as follows (for more details and references see, for instance, Bagdonavičius and Nikulin (2001) or Martinussen and Scheike (2006)). Let S_ω be a known survival function (of a continuous nonnegative variable ω), let $A_\omega = -\log S_\omega$ be the corresponding cumulative hazard. Assume that there exists a continuous increasing function G_k defined on the positive half-line with $G_k(0) = 0$ such that in the kth sample the survival function is $S_k(t) = S_\omega(G_k(t))$, and the cumulative hazard is $A_k(t) = A_\omega(G_k(t))$. The functions G_k are called cumulative rates. Denote the (noncumulative) rate $g_k(t) = dG_k(t)/dt$ and the hazard function $\alpha_k(t) = dA_k(t)/dt$. These noncumulative functions are in the one-to-one relationship $\alpha_k(t) = \alpha_\omega(G_k(t))g_k(t) = q_\omega(A_k(t))g_k(t)$, where $q_\omega(t) = \alpha_\omega(A_\omega^{-1}(t))$.

It is assumed that the functions G_1, G_2 are proportional, i.e., there exists real β such that $G_1(t) = G_0(t)$, $G_2(t) = e^\beta G_0(t)$ for all $t \in [0, \tau]$. The baseline cumulative rate G_0 is unknown

and not specified parametrically. Denote by R the survival time (distributed according to S_k in the kth sample). It is easily verified that in the kth sample the transformed survival time $G_k(R)$ follows the distribution S_ω. This implies the multiplicative model $G_0(R) = e^{-\beta z_k}\omega$, equivalently the linear model $\log G_0(R) = -\beta z_k + \log\omega$, where $z_k = 1[k = 2]$. That is, after the unknown transformation $\log G_0$ the survival times follow a location-shift model in the known error distribution of $\log\omega$.

Both main models, proportional hazards and proportional odds, fit in this framework.

The proportional hazards model is obtained for ω following the unit exponential distribution, that is $S_\omega(t) = e^{-t}$, $A_\omega(t) = t$, $\alpha_\omega(t) = q_\omega(t) = 1$. Then the cumulative rate G_k is the cumulative hazard A_k, g_k is the hazard rate α_k, and the model for $\log G_0(R)$ is a location model in the extreme value distribution.

When ω comes from the log-logistic distribution with $S_\omega(t) = 1/(1+t)$, $A_\omega(t) = \log(1+t)$, $\alpha_\omega(t) = 1/(1+t)$ and $q_\omega(t) = e^{-t}$, we get the proportional odds model since the cumulative rate G_k has the meaning of the odds function (because $G_k(t) = S_\omega^{-1}(S_k(t)) = (1 - S_k(t))/S_k(t)$). The rate $g_k(t) = dG_k(t)/dt$ may be called the odds-rate. The transformed time $\log G_0(R)$ has a shifted logistic distribution. In this model the hazard rates are $\alpha_k(t) = e^{\beta z_k} g_0(t)/(1 + e^{\beta z_k} G_0(t))$. Thus the hazard ratio $e^\beta(1 + G_0(t))/(1 + e^\beta G_0(t))$ converges to 1 as $t \to \infty$, and this convergence is monotonic (from above when $e^\beta > 1$, from below when $e^\beta < 1$). Therefore, this model is a popular alternative to proportional hazards when the hazards appear to approach each other for large times.

Reliability theory provides a different view of transformation models. The function S_ω is called resource, and G_k is the rate of resource usage. So in the proportional hazards model there are proportional rates of the exponential resource usage, in proportional odds the resource is log-logistic. Another example is a lognormal resource.

Linear transformation models are related to frailty models. Let U be frailty variables, i.e., unobservable positive random variables with a known distribution with expectation 1 which act multiplicatively on the hazard rate. That is, the conditional hazard of observations in the kth sample is $\alpha_k(t|U = u) = e^{\beta z_k} g_0(t)u$. Then the marginal survival function is $S_k(t) = \mathsf{E}\, S_k(t|U) = \mathsf{E}\exp\{-e^{\beta z_k} G_0(t)U\} = L_U(e^{\beta z_k} G_0(t))$, where L_U denotes the Laplace transform of the distribution of U. When R comes from the kth sample, the survival function of $e^{\beta z_k} G_0(R)$ is L_U (because $S_k(R)$ is uniformly distributed). Hence the Laplace transform L_U of the frailty distribution equals the survival function S_ω of the error variable ω in the transformation model. Also, as the conditional (on $U = u$) proportional hazards model $e^{\beta z_k} g_0(t)u$ is the transformation model $\log G_0(R) = -\beta z_k - \log u + \log\omega_0$ with $\log\omega_0$ being extreme-value distributed, we see that $\log G_0(R)$ unconditionally follows a tranformation model with errors $\log\omega = -\log U + \log\omega_0$ (thus the error distribution is the distribution of the difference of an extreme-value variable and $\log U$, which are independent).

A model without frailties ($U = 1$ a.s.) has $L_U(t) = S_\omega(t) = e^{-t}$, thus it is a proportional hazards model. When frailties are unit exponential, $L_U(t) = S_\omega(t) = 1/(1 + t)$, so the model is a proportional odds model. This agrees with the fact that the difference of two independent extreme-value variables ($\log\omega_0$ and $\log U$) is logistic. More generally, if frailties are gamma distributed with parameters $(1/v, 1/v)$ (expectation 1, variance v), it follows that $L_U(t) = S_\omega(t) = (1 + vt)^{-1/v}$, $\alpha_\omega(t) = (1 + vt)^{-1}$. This model is the proportional generalised odds model of Dabrowska and Doksum (1988) (in this model G_k are the generalised odds functions $v^{-1}(1 - S_k(t)^v)/S_k(t)^v$, they are proportional, while the hazard rates $\alpha_k(t) = e^{\beta z_k} g_0(t)/(1 + ve^{\beta z_k} G_0(t))$ converge to each other).

Section 3.2 explains the simplified partial likelihood estimation procedure needed in sub-

sequent considerations. In Section 3.3, I develop Neyman's smooth test of the proportional rates assumption, the main contribution of the chapter. Section 3.4 reviews and extends some other testing methods. Smooth tests and other procedures are compared via simulations reported in Section 3.5. A real data illustration can be found in Section 3.6. Technical material (theorems and proofs) is deferred to Section 3.7, which closes the chapter.

3.2 Estimation procedure

Let the data consist of pairs $(T_{j,i}, \delta_{j,i})$, $j = 1, 2$, $i = 1, \ldots, n_j$, where $T_{j,i} = \min(R_{j,i}, C_{j,i})$ are possibly censored survival times $R_{j,i}$ ($R_{j,i}$ are independent, with hazard function α_j), $\delta_{j,i} = 1[T_{j,i} = R_{j,i}]$ are failure indicators, and censoring times $C_{j,i}$ are mutually independent and independent of $R_{j,i}$. The standard counting process notation is used. Set $N_{j,i}(t) = 1[T_{j,i} \le t, \delta_{j,i} = 1]$, $\bar{N}_j(t) = \sum_{i=1}^{n_j} N_{j,i}(t)$, $\bar{N}(t) = \bar{N}_1(t) + \bar{N}_2(t)$, $Y_{j,i}(t) = 1[T_{j,i} \ge t]$, $\bar{Y}_j(t) = \sum_{i=1}^{n_j} Y_{j,i}(t)$. Let these processes be observed on a finite interval $[0, \tau]$.

For the estimation of β, I use the procedure of Bagdonavičius and Nikulin (2000) based on a simplification of the partial likelihood as follows. The partial likelihood takes the form

$$C(\tau; \beta, A_1, A_2) = \sum_{j=1}^{2} \sum_{i=1}^{n_j} \int_0^\tau \log\left(\frac{\lambda_{j,i}(t)}{\sum_{k=1}^{2} \sum_{l=1}^{n_k} \lambda_{k,l}(t)} \right) dN_{j,i}(t) = \int_0^\tau \log[q_\omega(A_1(t))] d\bar{N}_1(t)$$
$$+ \int_0^\tau \log[q_\omega(A_2(t)) e^\beta] d\bar{N}_2(t) - \int_0^\tau \log[\bar{Y}_1(t) q_\omega(A_1(t)) + \bar{Y}_2(t) q_\omega(A_2(t)) e^\beta] d\bar{N}(t).$$

Here A_1, A_2 depend on β which complicates differentiation when we want to derive a score equation. However, A_1, A_2 can be estimated directly without knowing β by Nelson–Aalen estimators $\hat{A}_j(t) = \int_0^t \bar{Y}_j(s)^{-1} d\bar{N}_j(s)$ computed separately in each sample. Therefore, we work with $C(\tau; \beta, \hat{A}_1, \hat{A}_2)$ instead of $C(\tau; \beta, A_1, A_2)$. Here it considerably simplifies calculations, especially when taking derivatives with respect to β. Then the score vector $\frac{\partial}{\partial \beta} C(\tau; \beta, \hat{A}_1, \hat{A}_2)$ is $U_1(\tau; \beta, \hat{A}_1, \hat{A}_2)$, where the score process equals

$$U_1(t; \beta, A_1, A_2) = \bar{N}_2(t) - \int_0^t \frac{\bar{Y}_2(s) q_\omega(A_2(s)) e^\beta}{\bar{Y}_1(s) q_\omega(A_1(s)) + \bar{Y}_2(s) q_\omega(A_2(s)) e^\beta} d\bar{N}(s)$$
$$= \int_0^\tau \frac{\bar{Y}_1(s) q_\omega(A_1(s)) \bar{Y}_2(s) q_\omega(A_2(s))}{\bar{Y}_1(s) q_\omega(A_1(s)) + \bar{Y}_2(s) q_\omega(A_2(s)) e^\beta} \left(\frac{d\bar{N}_2(s)}{\bar{Y}_2(s) q_\omega(A_2(s))} - e^\beta \frac{d\bar{N}_1(s)}{\bar{Y}_1(s) q_\omega(A_1(s))} \right). \quad (3.1)$$

The estimator $\hat{\beta}$ of the parameter β is defined as the maximiser of $C(\tau; \beta, \hat{A}_1, \hat{A}_2)$, that is, by concavity of C as a function of β, the solution to $U_1(\tau; \beta, \hat{A}_1, \hat{A}_2) = 0$. (Here and further in the chapter, the left-continuous version of the Nelson–Aalen estimator $\hat{A}_j(t-)$ is used in the integrand in C and U_1 to preserve predictability.)

Note that for the proportional hazards model ($q_\omega \equiv 1$) this estimation procedure agrees with the usual partial likelihood method.

Having computed β, one can obtain a Breslow-type model-based estimator of G_0 ($= G_1$) in the form

$$\hat{G}_0(t) = \int_0^t \frac{d\bar{N}(s)}{\bar{Y}_1(s) q_\omega(\hat{A}_1(s)) + \bar{Y}_2(s) q_\omega(\hat{A}_2(s)) e^{\hat{\beta}}}.$$

Before proceeding to main results we need to know that the simplified partial likelihood estimation procedure yields a consistent estimator of β. This is verified by Lemma 3.1 in Section 3.7.

Other estimation procedures have been developed for regression models with general covariates. Bagdonavičius and Nikulin (1999) use the modified partial likelihood. Variants of this approach are reviewed in Section 8.2 of Martinussen and Scheike (2006). Murphy, Rossini and van der Vaart (1997) propose the full nonparametric maximum likelihood estimation. Chen, Jin and Ying (2002) develop a method based on the iterative solution of some martingale estimating equations.

3.3 Neyman's smooth test

The general idea of Neyman's smooth test (see Rayner and Best, 1989) is based on embedding the null model into a model where the departure from the null is expressed by a d-dimensional parameter. That is, the general alternative that the hypothesis does not hold is replaced by a d-dimensional alternative. In the present context the null model assumes that the ratio of rates is constant over time, i.e., $g_2(t) = e^\beta g_1(t)$. Thus the Neyman embedding is most conveniently and most naturally formulated in terms of these transformation rates $g_k(t) = dG_k(t)/dt$. Under the alternative model the logarithm of the time-varying rate ratio is expressed as a linear combination of some bounded basis functions $\psi_1(t), \dots, \psi_d(t)$, that is

$$g_2(t) = \exp\{\beta + \theta^\mathsf{T}\psi(t)\}g_1(t). \tag{3.2}$$

These functions must be linearly independent and independent of 1 (then the model is identifiable). The Neyman-type smooth test of goodness of fit of the proportional rate model is the score test of $\theta = 0$ versus $\theta \neq 0$ in (3.2).

In the proportional hazards model the formulation of the embedding in terms of the noncumulative rates g_k is the obvious choice because g_k is the hazard rate and the name of the model actually speaks about hazards. On the other hand, in the proportional odds model one can be tempted to work directly with the odds functions $G_k = \Gamma_k$. This is not a good idea because G_k is an increasing (cumulative) function, thus in a model like $G_2(t) = \exp\{\beta+\theta^\mathsf{T}\psi(t)\}G_1(t)$ one would have to work with some monotonicity constraints. On the contrary, noncumulative rates may be arbitrary positive which poses no restrictions on β and ψ_j in (3.2).

The functions $\psi_j(t)$ are typically some standard basis functions on $[0,1]$ in transformed time, i.e., of the form $\psi_j(t) = \varphi_j(P(t)/P(\tau))$, where $\varphi_j(u), u \in [0,1]$, are, for instance, Legendre polynomials of order $1, \dots, d$, or cosines $\sqrt{2}\cos(j\pi u)$. The time-transformation $P(t)$ is a nondecreasing nonnegative continuous function with $P(0) = 0$, thus $P(t)/P(\tau)$ maps $[0,\tau]$ on $[0,1]$. Its purpose is to make the course of time in some sense uniform and hence better exploit the flexibility of the shape of φ_j. In practice $P(t)$ must be replaced by an estimator $\hat{P}(t)$. Here I use $\hat{P}^*(t) = 1 - \exp\{-\hat{A}^*(t)\}$, where $\hat{A}^*(t)$ is the Nelson–Aalen estimator computed from the pooled sample. The quantity $\hat{A}^*(t)$ consistently estimates $A^*(t) = \int_0^t \bar{y}_1(s)/\bar{y}(s)dA_1(s) + \int_0^t \bar{y}_2(s)/\bar{y}(s)dA_2(s)$, where $\bar{y}_j(t) = a_j S_j(t)(1 - C_j(t))$ denotes the uniform limit in probability of $n^{-1}\bar{Y}_j(t)$, $C_j(t)$ is the distribution function of censoring times and $a_j \in (0,1)$ is the limit of n_j/n (see Section 3.7 for details). If the censoring distribution is the same in both samples ($C_1(t) = C_2(t)$), the limit of $\hat{P}^*(t)$ is the distribution function corresponding to the mixture of survival distributions S_1, S_2 with weights a_1, a_2, i.e., $P^*(t) = a_1(1-S_1(t)) + a_2(1-S_2(t))$. Thus $P^*(t)$ is the distribution of a 'typical' observation.

Now let us finally derive the score test of significance of θ. If $C(\tau; \beta, \theta, \hat{A}_1, \hat{A}_2)$ denotes the simplified partial likelihood in the extended model (3.2), the score vector for inference about θ is $U_2(\tau; \beta, \theta, \hat{A}_1, \hat{A}_2) = \frac{\partial}{\partial\theta}C(\tau; \beta, \theta, \hat{A}_1, \hat{A}_2)$. The score test of significance of θ employs the

score vector $U_2(\tau;\hat\beta,0,\hat A_1,\hat A_2)$, denoted $U_2(\tau;\hat\beta,\hat A_1,\hat A_2)$ for short. Notice that

$$U_2(\tau;\beta,A_1,A_2)=\int_0^\tau \psi(t)U_1(dt;\beta,A_1,A_2).$$

In Section 3.7, I show that the score $n^{-1/2}U_2(\tau;\hat\beta,\hat A_1,\hat A_2)$ is asymptotically (with $n\to\infty$) normal with mean zero and variance matrix consistently estimated by $n^{-1}\hat\Xi$ given by (3.8). Consequently, the distribution of the quadratic test statistic

$$T_d=U_2(\tau;\hat\beta,\hat A_1,\hat A_2)^\mathsf{T}\hat\Xi^{-1}U_2(\tau;\hat\beta,\hat A_1,\hat A_2)$$

is approximately χ^2 with d degrees of freedom. Large values of T_d lead to rejection of the hypothesis.

I consider a data-driven version of Neyman's smooth test. The problem of choosing the suitable number of basis functions is addressed by the approach based on a modification of Schwarz's selection rule as described in Chapter 1. The number of basis functions is the maximiser of penalised score statistics, i.e., $S=\arg\max_{k=1,\dots,d}\{T_k-k\log n\}$. The data-driven test statistic is T_S. Under the null hypothesis, the selector S converges in probability to 1, and thus T_S is asymptotically χ^2-distributed with one degree of freedom. As this approximation is inaccurate (anticonservative), the more accurate two-term approximation provided in Section 1.4 (eq. (1.6)) is used.

3.4 Other tests

3.4.1 Komogorov–Smirnov-type test

A simple test can be based on the simplified partial likelihood score process $U_1(t;\hat\beta,\hat A_1,\hat A_2), t\in[0,\tau]$. When the fit of the proportional rate model is good, this process fluctuates around zero. When the model is not valid, the score process is expected to be far from zero. This may be measured by the Kolmogorov–Smirnov-type statistic $\sup_{t\in[0,\tau]}|U_1(t;\hat\beta,\hat A_1,\hat A_2)|$. Wei (1984) used this test for the two-sample proportional hazards model ($A_\omega(t)=t$), Bagdonavičius and Nikulin (2000) extended it to general two-sample transformation models.

Bagdonavičius and Nikulin (2000) proved that under the proportional rate model the score process is asymptotically Gaussian with mean zero. Here this convergence is proved (Lemma 3.3 in Section 3.7) as an intermediate result for the proof of the asymptotic distribution of the Neyman test statistic. The process is of the bridge type (equal to zero at times 0 and τ). In the special case of the proportional hazards model the score process converges to the Brownian bridge. In general, however, its limiting covariance structure is complicated and one has to resort to simulations. The standard resampling technique of Lin et al. (1993) (see also Martinussen and Scheike, 2006) can be used as the martingale representation is available, see eqs. (3.6) and (3.7). We obtain simulated paths of the test process by replacing unobservable martingale increments $dM_{k,i}(t)$ at failure times by randomly generated independent standard normal variables.

3.4.2 Gill–Schumacher-type test

Gill and Schumacher (1987) proposed a simple procedure for verifying proportionality of hazard functions in two samples. The idea is to compare two weighted estimators of the

hazard ratio. Here I use their idea and extend this approach to the general transformation setting.

Consider a weight function $K(t), t \in [0, \tau]$, which is a nonnegative predictable process. Assume that $n^{-1}K(t)$ converges in probability to some deterministic function $k(t)$, uniformly in $t \in [0, \tau]$. Then the proportionality parameter $\eta = e^{\beta} = g_2(t)/g_1(t)$ may be estimated by

$$\hat{\eta} = \frac{\int_0^\tau K(t)d\hat{G}_2(t)}{\int_0^\tau K(t)d\hat{G}_1(t)}.$$

The variable $\hat{\eta}$ converges to $\{\int_0^\tau k(t)g_2(t)dt\}/\{\int_0^\tau k(t)g_1(t)dt\} = \eta$. Now consider weights K_1, K_2 with the same properties as K. Denote $\hat{\eta}_j = \hat{\rho}_{j2}/\hat{\rho}_{j1}$, $\hat{\rho}_{jk} = \int_0^\tau n^{-1}K_j(t)d\hat{G}_k(t)$, $j = 1, 2$, $k = 1, 2$. Under the null hypothesis both $\hat{\eta}_1$ and $\hat{\eta}_2$ consistently estimate η, hence their difference $\hat{\eta}_2 - \hat{\eta}_1$ will fluctuate around zero. On the other hand, when the rate ratio $g_2(t)/g_1(t)$ is nonconstant and K_1 and K_2 emphasize time periods with different values of $g_2(t)/g_1(t)$, the difference $\hat{\eta}_2 - \hat{\eta}_1$ will be far from zero. Following Gill and Schumacher (1987), rewrite

$$\hat{\eta}_2 - \hat{\eta}_1 = \frac{\hat{\rho}_{22}\hat{\rho}_{11} - \hat{\rho}_{21}\hat{\rho}_{12}}{\hat{\rho}_{21}\hat{\rho}_{11}},$$

and use $\hat{\rho}_{22}\hat{\rho}_{11} - \hat{\rho}_{21}\hat{\rho}_{12}$ as the test statistic. In Section 3.7, this statistic is shown to be asymptotically zero-mean normal, a variance estimator is provided, and a consistency result is given (the test is consistent against monotonic rate ratios provided the limit of $K_2(t)/K_1(t)$ is monotonic).

For testing proportional hazards Gill and Schumacher (1987) discussed several choices of the weight functions and recommended the logrank weight $\bar{Y}_1(t)\bar{Y}_2(t)/(\bar{Y}_1(t) + \bar{Y}_2(t))$ and the Prentice–Wilcoxon weight $\hat{S}^*(t-)\bar{Y}_1(t)\bar{Y}_2(t)/(\bar{Y}_1(t) + \bar{Y}_2(t))$, where $\hat{S}^*(t-)$ is the left-continuous Kaplan–Meier estimator computed from the combined sample. In transformation models analogs of these weights are

$$\frac{\bar{Y}_1(t)q_\omega(\hat{A}_1(t-))\bar{Y}_2(t)q_\omega(\hat{A}_2(t-))}{\bar{Y}_1(t)q_\omega(\hat{A}_1(t-)) + \bar{Y}_2(t)q_\omega(\hat{A}_2(t-))}, \qquad \hat{S}^*(t-)\frac{\bar{Y}_1(t)q_\omega(\hat{A}_1(t-))\bar{Y}_2(t)q_\omega(\hat{A}_2(t-))}{\bar{Y}_1(t)q_\omega(\hat{A}_1(t-)) + \bar{Y}_2(t)q_\omega(\hat{A}_2(t-))}. \quad (3.3)$$

Note that a test related to that of Gill and Schumacher (1987) was proposed by Sengupta, Bhattacharjee and Rajeev (1998). They focused on alternatives where the cumulative hazard ratio $A_2(t)/A_1(t)$ is monotonic, which is a slightly broader class of alternatives than alternatives with monotonic $\alpha_2(t)/\alpha_1(t)$. Dauxois and Kirmani (2003) applied their idea to testing proportional odds against monotonic $\Gamma_2(t)/\Gamma_1(t)$. These tests are based on statistics of the same form $\hat{\rho}_{22}\hat{\rho}_{11} - \hat{\rho}_{21}\hat{\rho}_{12}$ but with $\hat{\rho}_{jk}$ defined as $\int_0^\tau K_j(t)\hat{G}_k(t)dt$ instead of $\int_0^\tau K_j(t)d\hat{G}_k(t)$. Unlike the Gill–Schumacher-type tests, these tests are not rank tests as they depend on the actual spaces between event times due to the Lebesgue integration.

3.5 Simulation study

I carried out a simulation study of the behaviour of three tests of proportionality: the data-driven smooth test (T_S), the Kolmogorov–Smirnov (KS) test and the Gill–Schumacher (GS) test. I repeatedly (10 000 times) generated two samples (each of size 50) of survival times under four scenarios. These include proportional hazards, proportional odds, and monotonic and

Table 3.1: Scenarios for the simulation study

		$\alpha_1(t)$	$\alpha_2(t)$	25 % cens. (a, b)	45 % cens. (a, b)
A	(Prop. hazards)	0.5	2	$(0, 5)$	$(0, 2)$
B	(Prop. odds)	$e^{-1}(1 + e^{-1}t)^{-1}$	$e^{1.5}(1 + e^{1.5}t)^{-1}$	$(2, 5)$	$(0, 3)$
C	(Monot. ratios)	1	$\frac{5}{3}t^{2/3}$	$(0, 3.8)$	$(0, 2)$
D	(Nonmon. ratios)	0.75	$\frac{3}{2}(t-1)^2$	$(0, 5.5)$	$(0, 3)$

Table 3.2: Estimated rejection probabilities on the nominal level 5 % for configurations A to D with 25 % and 45 % censoring proportions. Sample sizes $n_1 = n_2 = 50$. Figures based on 10 000 simulation repetitions (standard deviation 0.005).

	A		B		C		D	
	25 %	45 %	25 %	45 %	25 %	45 %	25 %	45 %
Hypothesis: proportional hazards								
T_S	0.056	0.048	0.322	0.169	0.577	0.417	0.926	0.692
KS	0.052	0.050	0.528	0.322	0.595	0.487	0.584	0.493
GS	0.042	0.034	0.249	0.076	0.707	0.548	0.122	0.172
Hypothesis: proportional odds								
T_S	0.414	0.220	0.052	0.049	0.507	0.314	0.926	0.683
KS	0.218	0.114	0.043	0.044	0.490	0.358	0.304	0.505
GS	0.221	0.153	0.021	0.024	0.525	0.388	0.058	0.293

nonmonotonic ratios of hazard rates and odds rates. Survival times were censored by independent variables distributed uniformly on intervals (a, b) adjusted to produce approximately 25 % and 45 % censored observations. Parameters of the simulation design are summarised in Table 3.1. On the level of 5 %, tests of both proportional hazards and proportional odds were performed. The Kolmogorov–Smirnov test was performed with 1000 resampled processes. The data-driven smooth test was used with $d = 5$, with the Legendre polynomial basis, with the two-term approximation of the distribution of the test statistic. The GS test used the weights (3.3), the statistic was compared to asymptotic critical points.

Table 3.2 reports estimates of rejection probabilities. It is seen that Neyman's test and the Kolmogorov–Smirnov test preserve the level very well (see scenario A for proportional hazards and B for proportional odds). The Gill–Schumacher test tends to be slightly conservative, mainly under proportional odds. Under alternatives with monotonic ratios of hazard and/or odds rates (A–C), the overall performance seems to be comparable for all three tests. In the nonmonotonic situation D, it is no surprise that the Gill–Schumacher-type test does not do well as it is designed to be sensitive against monotonic alternatives. The main message of the results concerning the power of the proposed smooth test is that this test maintains stable power for a variety of departures from proportionality. I performed simulations for other combinations of distributions, and never met a situation where the smooth test dramatically lost compared to the other methods.

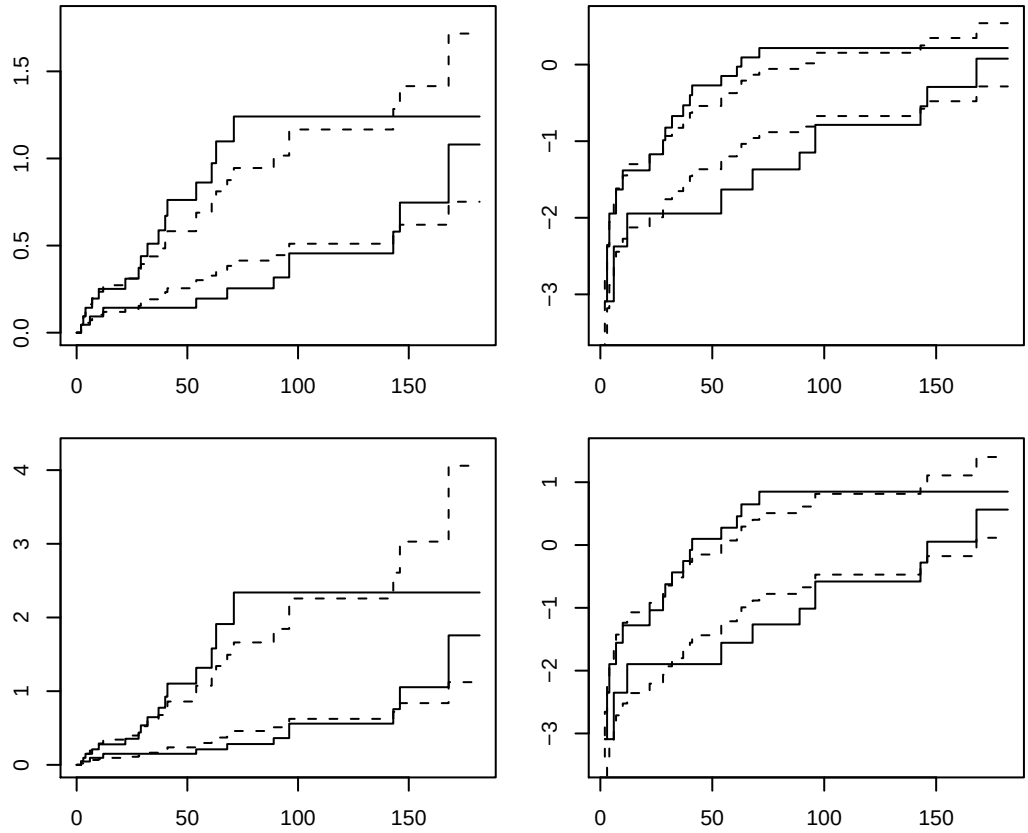


Figure 3.1: Estimated cumulative hazards and odds functions for the chronic active hepatitis data. Upper row: cumulative hazards (left panel) and log-cumulative hazards (right). Lower row: odds (left) and log-odds functions (right). In each plot: solid curves are estimates computed separately for the treatment group (lower curves) and control group (upper curves), dashed lines show corresponding model-based estimates. Time from the beginning of the trial is in months.

3.6 Illustration

A real example is taken from Collett (2003, Appendix D.1). The data concern survival times of patients with chronic active hepatitis. There were 44 patients, 22 of them (randomly selected) received a drug (11 died, 11 were censored), the remaining 22 were in the control group (16 deaths, 6 survivors). Figure 3.1 displays estimates of cumulative hazards, odds functions, and their logarithms (i.e., complementary log-log and logit transform of the Kaplan–Meier estimate). These estimates obtained separately from each sample are plotted by solid lines. If the proportional assumption holds, the vertical distance between log-curves should be approximately constant. Estimates based on proportional rate models are plotted by dashed lines. Results of goodness-of-fit tests are summarised in Table 3.3.

The partial likelihood estimate in the proportional hazards model is $\hat{\beta} = 0.826$ ($e^{\hat{\beta}} = 2.28$). The data-driven smooth test (with maximum dimension $d = 5$) rejects the hypothesis of proportional hazards. Schwarz's selection rule selects two basis functions (the linear and

Table 3.3: Results of tests of fit for the chronic active hepatitis data

	Proportional hazards		Proportional odds	
	Statistic	p-value	Statistic	p-value
T_1	2.09	0.148	0.45	0.502
T_2	7.71	0.021	5.36	0.069
T_3	7.85	0.049	5.67	0.129
T_4	8.30	0.081	6.09	0.192
T_5	9.36	0.096	7.29	0.200
T_S	7.71	0.005	5.36	0.061
KS	3.00	0.052	2.53	0.169
GS	1.23	0.220	0.46	0.648

quadratic Legendre polynomial), which corresponds to the fact that the hazard ratio appears to be nonmonotonic. The quadratic function contributes to the description of the hazard ratio most; using more than two basis functions does not increase the statistic much. The Gill–Schumacher test does not reject the hypothesis of proportionality (the logrank and Prentice–Wilcoxon weighted estimates of $\eta = e^\beta$ are 2.37 and 2.73, respectively) as this test is focused against alternatives with monotonic ratios.

If we are interested in the proportional odds model, the simplified partial likelihood procedure gives the estimate $\hat\beta = 1.29$ ($e^{\hat\beta} = 3.63$), and two weighted estimates used in the Gill–Schumacher test are $\hat\eta_1 = 3.74$, $\hat\eta_2 = 3.93$ (with weights (3.3)). Plots of (log-)odds functions indicate a similar type of departure from proportionality as plots of cumulative hazards; results, however, do not lead to rejection on the 5 % level.

3.7 Asymptotic results

3.7.1 Assumptions

It is assumed that $n^{-1}\bar Y_j(t)$, $j = 1, 2$, converge in probability uniformly in $t \in [0, \tau]$ to some functions $\bar y_j(t)$ bounded away from zero. This is satisfied if $n_j/n \to a_j \in (0, 1)$, $S_j(\tau) > 0$ and $1 - C_j(\tau) > 0$ (C_j is the distribution function of censoring variables) because then by the Glivenko–Cantelli theorem $\bar y_j(t) = a_j S_j(t)(1 - C_j(t))$.

Further assume that $q_\omega(t) = \alpha_\omega(A_\omega^{-1}(t))$ is continuously differentiable on $[0, \tau]$. For both main examples this is satisfied ($q_\omega(t) = 1$ for proportional hazards, $q_\omega(t) = e^{-t}$ for proportional odds). Denote the derivative $\dot q_\omega(t)$.

3.7.2 Consistency of the estimation procedure

Lemma 3.1 (Convergence of $\hat\beta$)**.** *Assume that the rate ratio $g_2(t)/g_1(t)$ is $e^{\beta_0(t)}$, i.e., it may or may not be constant. Then the estimator $\hat\beta$ defined as the solution to $U_1(\tau; \beta, \hat A_1, \hat A_2) = 0$ converges in probability to the solution $\bar\beta$ to the limiting estimating equation*

$$\int_0^\tau \frac{\bar y_1(t) q_\omega(A_1(t)) \bar y_2(t) q_\omega(A_2(t))}{\bar y_1(t) q_\omega(A_1(t)) + \bar y_2(t) q_\omega(A_2(t)) e^\beta} (e^{\beta_0(t)} - e^\beta) g_0(t) dt = 0. \tag{3.4}$$

Specifically, if the proportional rate model holds (β_0 is constant), $\hat{\beta}$ consistently estimates the true value β_0.

Proof. The proof is analogous to that for the Cox model (see Theorem 8.3.1 of Fleming and Harrington (1991) or Theorem VII.2.1 of Andersen et al. (1993), and also Struthers and Kalbfleisch (1986)). The maximiser of $C(\tau;\beta,\hat{A}_1,\hat{A}_2)$ is the same as the maximiser of

$$\begin{aligned}
&n^{-1}(C(\tau;\beta,\hat{A}_1,\hat{A}_2)-C(\tau;\bar{\beta},\hat{A}_1,\hat{A}_2))\\
&\quad= n^{-1}(\beta-\bar{\beta})\bar{N}_2(\tau)-n^{-1}\int_0^\tau \log\left(\frac{\bar{Y}_1(t)q_\omega(\hat{A}_1(t-))+\bar{Y}_2(t)q_\omega(\hat{A}_2(t-))e^{\beta}}{\bar{Y}_1(t)q_\omega(\hat{A}_1(t-))+\bar{Y}_2(t)q_\omega(\hat{A}_2(t-))e^{\bar{\beta}}}\right)d\bar{N}(t)\\
&\quad= n^{-1}(\beta-\bar{\beta})\bar{\Lambda}_2(\tau)-n^{-1}\int_0^\tau \log\left(\frac{\bar{Y}_1(t)q_\omega(\hat{A}_1(t-))+\bar{Y}_2(t)q_\omega(\hat{A}_2(t-))e^{\beta}}{\bar{Y}_1(t)q_\omega(\hat{A}_1(t-))+\bar{Y}_2(t)q_\omega(\hat{A}_2(t-))e^{\bar{\beta}}}\right)d\bar{\Lambda}(t)\\
&\qquad+ n^{-1}(\beta-\bar{\beta})\bar{M}_2(\tau)-n^{-1}\int_0^\tau \log\left(\frac{\bar{Y}_1(t)q_\omega(\hat{A}_1(t-))+\bar{Y}_2(t)q_\omega(\hat{A}_2(t-))e^{\beta}}{\bar{Y}_1(t)q_\omega(\hat{A}_1(t-))+\bar{Y}_2(t)q_\omega(\hat{A}_2(t-))e^{\bar{\beta}}}\right)d\bar{M}(t).
\end{aligned}$$

Here the last two terms converge to zero by Lenglart's inequality. Hence, by the uniform consistency of Nelson–Aalen estimators, $n^{-1}(C(\tau;\beta,\hat{A}_1,\hat{A}_2)-C(\tau;\bar{\beta},\hat{A}_1,\hat{A}_2))$ converges in probability to

$$\begin{aligned}
(\beta-\bar{\beta})\int_0^\tau \bar{y}_2(t)q_\omega(A_2(t))e^{\beta_0(t)}g_0(t)dt-\int_0^\tau \log\left(\frac{\bar{y}_1(t)q_\omega(A_1(t))+\bar{y}_2(t)q_\omega(A_2(t))e^{\beta}}{\bar{y}_1(t)q_\omega(A_1(t))+\bar{y}_2(t)q_\omega(A_2(t))e^{\bar{\beta}}}\right)\\
\times[\bar{y}_1(t)q_\omega(A_1(t))+\bar{y}_2(t)q_\omega(A_2(t))e^{\beta_0(t)}]g_0(t)dt. \quad (3.5)
\end{aligned}$$

Then, in the light of concavity of $n^{-1}(C(\tau;\beta,\hat{A}_1,\hat{A}_2)-C(\tau;\bar{\beta},\hat{A}_1,\hat{A}_2))$, Lemma 8.3.1 of Fleming and Harrington (1991) (see also Appendix II of Andersen and Gill, 1982) yields that the maximiser of $n^{-1}(C(\tau;\beta,\hat{A}_1,\hat{A}_2)-C(\tau;\bar{\beta},\hat{A}_1,\hat{A}_2))$ converges in probability to the maximiser of (3.5), which is by concavity the solution to (3.4). □

3.7.3 Asymptotics for Neyman's test

Lemma 3.2. *The process $n^{-1/2}U_1(\cdot;\beta_0,\hat{A}_1,\hat{A}_2)$ is asymptotically distributed as the process*

$$V_1(t)=\int_0^t l_{12}(s)dV_{12}(s)-\int_0^t l_{11}(s)dV_{11}(s)-\int_0^t V_{12}(s)dh_{12}(s)+\int_0^t V_{11}(s)dh_{11}(s),$$

where V_{1j} are independent zero-mean continuous Gaussian martingales with variance functions $\int_0^t \bar{y}_j(s)^{-1}dA_j(s)$, and the functions h_{1j} and l_{1j} are uniform limits in probability of $n^{-1}H_{1j}$ and $n^{-1}L_{1j}$ defined below in the proof.

Proof. By the martingale central limit theorem, the process $n^{1/2}(\hat{A}_1-A_1,\hat{A}_2-A_2)$ converges in distribution to (V_{11},V_{12}), which is a standard result on the Nelson–Aalen estimator. Rewrite $U_1(t;\beta_0,\hat{A}_1,\hat{A}_2)$ completely in terms of $\hat{A}_j$ as follows

$$\begin{aligned}
U_1(t;\beta_0,\hat{A}_1,\hat{A}_2)=\int_0^\tau \frac{\bar{Y}_1(s)q_\omega(\hat{A}_1(s))}{\bar{Y}_1(s)q_\omega(\hat{A}_1(s))+\bar{Y}_2(s)q_\omega(\hat{A}_2(s))e^{\beta_0}}\bar{Y}_2(s)d\hat{A}_2(s)\\
-\int_0^\tau \frac{\bar{Y}_2(s)q_\omega(\hat{A}_2(s))e^{\beta_0}}{\bar{Y}_1(s)q_\omega(\hat{A}_1(s))+\bar{Y}_2(s)q_\omega(\hat{A}_2(s))e^{\beta_0}}\bar{Y}_1(s)d\hat{A}_1(s).
\end{aligned}$$

When in this expression $\hat{A}_j$ are replaced by A_j, the result is zero. Thus the asymptotic distribution of $U_1(t;\beta_0,\hat{A}_1,\hat{A}_2)$ (minus zero) can be inferred from that of $\hat{A}_j - A_j$ with the help of the functional delta method. Using the chain rule and a lemma on differentiation of integration (Proposition II.8.6 in Andersen et al. (1993) or Lemma 3.9.17 in van der Vaart and Wellner (1996)), we obtain that $n^{-1/2}U_1(t;\beta_0,\hat{A}_1,\hat{A}_2)$ is asymptotically equivalent to

$$\int_0^t n^{-1}L_{12}(s)n^{1/2}(d\hat{A}_2(s) - dA_2(s)) - \int_0^t n^{-1}L_{11}(s)n^{1/2}(d\hat{A}_1(s) - dA_1(s)) \\ - \int_0^t n^{1/2}(\hat{A}_2(s) - A_2(s))n^{-1}dH_{12}(s) + \int_0^t n^{1/2}(\hat{A}_1(s) - A_1(s))n^{-1}dH_{11}(s), \quad (3.6)$$

where

$$L_{11}(t) = \frac{\bar{Y}_1(t)\bar{Y}_2(t)q_\omega(A_2(t))e^{\beta_0}}{\bar{Y}_1(t)q_\omega(A_1(t)) + \bar{Y}_2(t)q_\omega(A_2(t))e^{\beta_0}},$$
$$L_{12}(t) = \frac{\bar{Y}_1(t)q_\omega(A_1(t))\bar{Y}_2(t)}{\bar{Y}_1(t)q_\omega(A_1(t)) + \bar{Y}_2(t)q_\omega(A_2(t))e^{\beta_0}},$$
$$H_{11}(t) = \int_0^t \frac{\bar{Y}_1(s)\dot{q}_\omega(A_1(s))\bar{Y}_2(s)q_\omega(A_2(s))e^{\beta^0}}{[\bar{Y}_1(s)q_\omega(A_1(s)) + \bar{Y}_2(s)q_\omega(A_2(s))e^{\beta^0}]^2}[\bar{Y}_1(s)dA_1(s) + \bar{Y}_2(s)dA_2(s)],$$
$$H_{12}(t) = \int_0^t \frac{\bar{Y}_1(s)q_\omega(A_1(s))\bar{Y}_2(s)\dot{q}_\omega(A_2(s))e^{\beta_0}}{[\bar{Y}_1(s)q_\omega(A_1(s)) + \bar{Y}_2(s)q_\omega(A_2(s))e^{\beta_0}]^2}[\bar{Y}_1(s)dA_1(s) + \bar{Y}_2(s)dA_2(s)].$$

□

Lemma 3.3. *The process $n^{-1/2}U_1(\cdot;\hat{\beta},\hat{A}_1,\hat{A}_2)$ is asymptotically distributed as the process $V_1(t) - d_1(t;\beta_0,A_1,A_2)d_1(\tau;\beta_0,A_1,A_2)^{-1}V_1(\tau)$, where V_1 is the process of Lemma 3.2 and the function $d_1(t;\beta,A_1,A_2)$ is the uniform limit in probability of $n^{-1}D_1(t;\beta,\hat{A}_1,\hat{A}_2)$ defined below.*

Proof. The proof follows by Lemma 3.2 after a straightforward use of Taylor's expansion which gives

$$n^{-1/2}U_1(t;\hat{\beta},\hat{A}_1,\hat{A}_2) = n^{-1/2}U_1(t;\beta_0,\hat{A}_1,\hat{A}_2) - n^{-1}D_1(t;\beta_t^*,\hat{A}_1,\hat{A}_2)n^{1/2}(\hat{\beta}-\beta_0) \\ = n^{-1/2}U_1(t;\beta_0,\hat{A}_1,\hat{A}_2) - \{n^{-1}D_1(t;\beta_t^*,\hat{A}_1,\hat{A}_2)\}\{n^{-1}D_1(\tau;\beta_\tau^*,\hat{A}_1,\hat{A}_2)\}^{-1} \\ \times n^{-1/2}U_1(t;\beta_0,\hat{A}_1,\hat{A}_2), \quad (3.7)$$

where β_t^* lies on the line segment between β_0 and $\hat{\beta}$, and

$$D_1(t;\beta,\hat{A}_1,\hat{A}_2) = -\frac{\partial}{\partial\beta}U_1(t;\beta,\hat{A}_1,\hat{A}_2) = \int_0^t \frac{\bar{Y}_1(s)q_\omega(\hat{A}_1(s))\bar{Y}_2(s)q_\omega(\hat{A}_2(s))e^{\beta}}{[\bar{Y}_1(s)q_\omega(\hat{A}_1(s)) + \bar{Y}_2(s)q_\omega(\hat{A}_2(s))e^{\beta}]^2}d\bar{N}(s).$$

□

Theorem 3.4 (Asymptotic distribution of the score)**.** *The score vector $n^{-1/2}U_2(\tau;\hat{\beta},\hat{A}_1,\hat{A}_2)$ converges in distribution to a mean zero Gaussian vector with variance matrix that is consistently estimated by $n^{-1}\hat{\Xi}$ given below in* (3.8)*.*

Proof. By Taylor's expansion about β_0,

$$n^{-1/2}U_2(\tau;\hat\beta,\hat A_1,\hat A_2) = n^{-1/2}U_2(\tau;\beta_0,\hat A_1,\hat A_2) \\ - n^{-1}D_2(\tau;\beta^{**},\hat A_1,\hat A_2)\{n^{-1}D_1(\tau;\beta^*,\hat A_1,\hat A_2)\}^{-1}n^{-1/2}U_1(\tau;\beta_0,\hat A_1,\hat A_2),$$

where $D_2(\tau;\beta,\hat A_1,\hat A_2) = -\frac{\partial}{\partial\beta}U_2(\tau;\beta,\hat A_1,\hat A_2)$, and β^* and β^{**} are on the line segment between β_0 and $\hat\beta$ (to be technically precise, note that each component of D_2 has its own β^{**}, all between β_0 and $\hat\beta$). The variables $n^{-1}D_1(\tau;\beta^*,\hat A_1,\hat A_2), n^{-1}D_2(\tau;\beta^{**},\hat A_1,\hat A_2)$ converge in probability to $d_1(\tau;\beta_0,A_1,A_2), d_2(\tau;\beta_0,A_1,A_2)$, respectively, where d_1 is explained in Lemma 3.3 and $d_2(\tau;\beta,A_1,A_2) = \int_0^\tau \psi(t)d_{11}(dt;\beta,A_1,A_2)$. The $(1+d)$-dimensional vector $n^{-1/2}(U_1(\tau;\beta_0,\hat A_1,\hat A_2), U_2(\tau;\beta_0,\hat A_1,\hat A_2)^\mathsf{T})^\mathsf{T}$ jointly converges weakly to the zero-mean Gaussian vector $(V_1(\tau), V_2(\tau)^\mathsf{T})^\mathsf{T}$ with V_1 given in Lemma 3.3 and $V_2(\tau) = \int_0^\tau \psi(t)dV_1(t)$.

Let us derive suitable variance estimators. Integrating by parts (or using Fubini's theorem) in (3.6) yields that $n^{-1/2}U_1(\tau;\beta_0,\hat A_1,\hat A_2)$ has the same asymptotic distribution as

$$\int_0^\tau n^{-1}[L_{12}(t)+H_{12}(t))]n^{1/2}(d\hat A_2(s)-dA_2(s)) - n^{-1}H_{12}(\tau)n^{1/2}(\hat A_2(\tau)-A_2(\tau)) \\ - \int_0^\tau n^{-1}[L_{11}(t)+H_{11}(t)]n^{1/2}(d\hat A_1(s)-dA_1(s)) + n^{-1}H_{11}(\tau)n^{1/2}(\hat A_1(\tau)-A_1(\tau)).$$

For $n^{-1/2}U_2(\tau;\beta_0,\hat A_1,\hat A_2)$ we get an analogous expression with $L_{2j}(t) = \psi(t)L_{1j}(t)$ instead of $L_{1j}(t)$ and $H_{2j}(t) = \int_0^t \psi(s)dH_{1j}(s)$ instead of $H_{1j}(t)$. Thus the asymptotic variance Σ_{11} of $n^{-1/2}U_1(\tau;\beta_0,\hat A_1,\hat A_2)$, covariance vector Σ_{21} of $n^{-1/2}U_2(\tau;\beta_0,\hat A_1,\hat A_2)$, $n^{-1/2}U_1(\tau;\beta_0,\hat A_1,\hat A_2)$, and variance matrix Σ_{22} of $n^{-1/2}U_2(\tau;\beta_0,\hat A_1,\hat A_2)$ can be consistently estimated by

$$n^{-1}\hat\Sigma_{kk'} = \sum_{j=1}^2 \int_0^\tau n^{-1}[\hat L_{kj}(t)+\hat H_{kj}(t)-\hat H_{kj}(\tau)][\hat L_{k'j}(t)+\hat H_{k'j}(t)-\hat H_{k'j}(\tau)]^\mathsf{T}\frac{d\hat A_j(t)}{\bar Y_j(t)},$$

$k,k' = 1,2$. Here $\hat L_{kj}$ and $\hat H_{kj}$ are defined like L_{kj} and H_{kj} with β_0, A_1, A_2 replaced by $\hat\beta, \hat A_1, \hat A_2$ (with left-continuous Nelson–Aalen estimators in integrands because of predictability). The very final conclusion is that the asymptotic variance of $n^{-1/2}U_2(\tau;\hat\beta,\hat A_1,\hat A_2)$ is estimated by

$$n^{-1}\hat\Xi = n^{-1}\hat\Sigma_{22} - n^{-1}\hat D_2\hat D_1^{-1}\hat\Sigma_{21}^\mathsf{T} - n^{-1}\hat\Sigma_{21}\hat D_1^{-1}\hat D_2^\mathsf{T} + n^{-1}\hat D_2\hat D_1^{-1}\hat\Sigma_{11}\hat D_1^{-1}\hat D_2^\mathsf{T} \tag{3.8}$$

(having set $\hat D_k = D_k(\tau;\hat\beta,\hat A_1,\hat A_2)$). □

Theorem 3.5 (Consistency of Neyman's test)**.** *Assume that the true rate ratio is time-varying of the form $e^{\beta_0(t)}$. Let $\bar\beta$ be as in Lemma 3.1. Suppose that the basis functions satisfy the condition*

$$\int_0^\tau \psi(t)\frac{\bar y_1(t)q_\omega(A_1(t))\bar y_2(t)q_\omega(A_2(t))}{\bar y_1(t)q_\omega(A_1(t)) + \bar y_2(t)q_\omega(A_2(t))e^{\bar\beta}}(e^{\beta_0(t)}-e^{\bar\beta})g_0(t)dt \neq 0 \tag{3.9}$$

(at least one component differs from zero). Then the rejection probability of Neyman's test approaches 1 *as $n\to\infty$.*

Proof. In view of the definition $U_2(\tau;\hat\beta,\hat A_1,\hat A_2)=\int_0^\tau \psi(t)U_1(dt;\hat\beta,\hat A_1,\hat A_2)$ and Lemma 3.1, $n^{-1}U_2(\tau;\hat\beta,\hat A_1,\hat A_2)$ converges in probability to the left-hand side of (3.9). The variance matrix estimator $n^{-1}\hat\Xi$ converges to some finite matrix. Thus n^{-1} times the score test statistic converges in probability to a nonzero number. □

The consistency condition means, loosely speaking, that the choice of the basis functions is not 'completely wrong'. More precisely, the left-hand side of (3.9) is the limiting estimating equation for the parameters $\theta=(\theta_1,\dots,\theta_d)^\mathsf{T}$ in the smooth model (3.2) evaluated at $\theta=0$. The inequality (3.9) means that $\theta=0$ does not solve the estimating equation, that is, the basis function contribute to the description of the true time-varying rate ratio. In other words, the test is consistent against alternatives whose projection on the smooth model (3.2) does not fall to the null model.

3.7.4 Asymptotics for the Gill–Schumacher test

Assume that $n^{-1}K_j(t)$, $j=1,2$, converge in probability uniformly in $t\in[0,\tau]$ to some functions $k_j(t)$ bounded away from zero. For instance, the logrank-type and Prentice–Wilcoxon-type weights (3.3) satisfy this condition by the convergence of $n^{-1}\bar Y_j(t)$ and the Kaplan–Meier estimator.

Theorem 3.6 (Asymptotic distribution of the GS statistic)**.** *Under the null hypothesis of proportionality of g_1,g_2, the test statistic $n^{1/2}(\hat\rho_{22}\hat\rho_{11}-\hat\rho_{21}\hat\rho_{12})$ is asymptotically normal with mean zero and variance given by* (3.11) *below, which is consistently estimated by* (3.12)*.*

Proof. Denote $\rho_{jk}=\int_0^\tau k_j(t)dG_k(t)$ and rewrite

$$\begin{aligned}&\hat\rho_{22}\hat\rho_{11}-\hat\rho_{21}\hat\rho_{12}\\&\quad=(\hat\rho_{22}-\rho_{22})\hat\rho_{11}+(\hat\rho_{11}-\rho_{11})\rho_{22}-(\hat\rho_{21}-\rho_{21})\hat\rho_{12}-(\hat\rho_{12}-\rho_{12})\rho_{21}+\rho_{22}\rho_{11}-\rho_{21}\rho_{12}.\end{aligned}\tag{3.10}$$

Under the hypothesis it is $\rho_{j2}=\eta\rho_{j1}$, hence the last two terms together are zero. Further, $\hat\rho_{jk}$ converges in probability to ρ_{jk}. It remains to explore the weak convergence of $n^{1/2}(\hat\rho_{jk}-\rho_{jk})$ jointly for $j=1,2$, $k=1,2$.

Recall that $G_k(t)=A_\omega^{-1}(A_k(t))$ and $\hat G_k(t)=A_\omega^{-1}(\hat A_k(t))$. By the functional delta method $n^{1/2}(\hat G_k(\cdot)-G_k(\cdot))$ is asymptotically equivalent to

$$\frac{1}{q_\omega(A_k(\cdot))}n^{1/2}(\hat A_k(\cdot)-A_k(\cdot)).$$

Thus, the asymptotic distribution of $n^{1/2}(\hat\rho_{jk}-\rho_{jk})$ is the same as the asymptotic distribution of

$$\begin{aligned}&\int_0^\tau n^{-1}K_j(t)d\biggl(\frac{1}{q_\omega(A_k(t))}n^{1/2}(\hat A_k(t)-A_k(t))\biggr)\\&\qquad=\int_0^\tau n^{1/2}(\hat A_k(t)-A_k(t))n^{-1}dB_{jk}(t)+\int_0^\tau\frac{n^{-1}K_j(t)}{q_\omega(A_k(t))}n^{1/2}(d\hat A_k(t)-dA_k(t)),\end{aligned}$$

where $dB_{jk}(t)=K_j(t)dB_k(t)$, $dB_k(t)=d(1/q_\omega(A_k(t)))=-\dot q_\omega(A_k(t))/q_\omega(A_k(t))dG_k(t)$. This asymptotic distributional equivalence holds jointly for $j=1,2$, $k=1,2$. Integrating by parts

we arrive at

$$\int_0^\tau n^{-1}R_{jk}(t)n^{1/2}(d\hat{A}_k(t) - dA_k(t)) + n^{-1}B_{jk}(\tau)n^{1/2}(\hat{A}_k(\tau) - A_k(\tau)),$$

where

$$R_{jk}(t) = \frac{K_j(t)}{q_\omega(A_k(t))} - B_{jk}(t).$$

Denote by b_{jk} and r_{jk} the limits of $n^{-1}B_{jk}$ and $n^{-1}R_{jk}$, respectively. Then by the martingale central limit theorem $n^{1/2}(\hat{\rho}_{jk} - \rho_{jk})$, $j = 1, 2$, $k = 1, 2$, converge to zero mean jointly normal variables. The asymptotic covariance of $n^{1/2}(\hat{\rho}_{jk} - \rho_{jk})$ and $n^{1/2}(\hat{\rho}_{j'k} - \rho_{j'k})$ is

$$\int_0^\tau (r_{jk}(t) + b_{jk}(\tau))(r_{j'k}(t) + b_{j'k}(\tau))\frac{dA_k(t)}{\bar{y}_k(t)},$$

while the asymptotic covariance of $n^{1/2}(\hat{\rho}_{jk} - \rho_{jk})$ and $n^{1/2}(\hat{\rho}_{j'k'} - \rho_{j'k'})$ is zero for $k \neq k'$.

Therefore, using the fact $\rho_{j2} = \eta\rho_{j1}$, it follows that the asymptotic variance of the statistic $n^{1/2}(\hat{\rho}_{22}\hat{\rho}_{11} - \hat{\rho}_{21}\hat{\rho}_{12})$ is

$$\int_0^\tau [(r_{11}(t) + b_{11}(\tau))\rho_{21} - (r_{21}(t) + b_{21}(\tau))\rho_{11}]^2\eta^2\frac{dA_1(t)}{\bar{y}_1(t)} + \int_0^\tau [(r_{12}(t) + b_{12}(\tau))\rho_{11} - (r_{22}(t) + b_{22}(\tau))\rho_{21}]^2\frac{dA_2(t)}{\bar{y}_2(t)}.$$

Finally, as $dG_2(t) = \eta dG_1(t)$ and $dG_k(t) = dA_k(t)/q_\omega(A_k(t))$, we arrive at

$$\int_0^\tau [(r_{11}(t) + b_{11}(\tau))\rho_{21} - (r_{21}(t) + b_{21}(\tau))\rho_{11}]^2\eta\frac{q_\omega(A_1(t))dA_2(t)}{q_\omega(A_2(t))\bar{y}_1(t)} + \int_0^\tau [(r_{12}(t) + b_{12}(\tau))\rho_{11} - (r_{22}(t) + b_{22}(\tau))\rho_{21}]^2\eta\frac{q_\omega(A_2(t))dA_1(t)}{q_\omega(A_1(t))\bar{y}_2(t)}, \quad (3.11)$$

which may be consistently estimated by

$$\int_0^\tau [n^{-1}(\hat{R}_{11}(t) + \hat{B}_{11}(\tau))\hat{\rho}_{21} - n^{-1}(\hat{R}_{21}(t) + \hat{B}_{21}(\tau))\hat{\rho}_{11}]^2\hat{\eta}_0\frac{q_\omega(\hat{A}_1(t-))d\hat{A}_2(t)}{q_\omega(\hat{A}_2(t-))\bar{Y}_1(t)/n} + \int_0^\tau [n^{-1}(\hat{R}_{12}(t) + \hat{B}_{12}(\tau))\hat{\rho}_{11} - n^{-1}(\hat{R}_{22}(t) + \hat{B}_{22}(\tau))\hat{\rho}_{21}]^2\hat{\eta}_0\frac{q_\omega(\hat{A}_2(t-))d\hat{A}_1(t)}{q_\omega(\hat{A}_1(t-))\bar{Y}_2(t)/n}. \quad (3.12)$$

Here $\hat{R}_{jk}(t)$ and $\hat{B}_{jk}(t)$ are defined like $R_{jk}(t)$ and $B_{jk}(t)$ but with unknown quantities replaced by their estimators (the Nelson–Aalen estimator is used in the left-continuous version where necessary to preserve predictability). The estimator $\hat{\eta}_0$ is defined as $(\hat{\eta}_1 + \hat{\eta}_2)/2$ to preserve some sort of symmetry (but any other consistent estimator of η may be used as well). □

Note that the variance estimator (3.12) is always positive. A different estimator of a form similar to that of Gill and Schumacher (1987, eq. (4)) can be derived. However, such an estimator may be (and my experience is that sometimes actually is) negative.

Theorem 3.7 (Consistency of the Gill–Schumacher test)**.** *The Gill–Schumacher-type test is consistent against alternatives satisfying* $\rho_{22}\rho_{11} - \rho_{21}\rho_{12} \neq 0$*. This particularly holds for alternatives with monotonic* $g_2(t)/g_1(t)$ *whenever* $k_2(t)/k_1(t)$ *is monotonic.*

Proof. The variance estimator (3.12) converges to some finite nonzero quantity even under the alternative, and, thus, the first assertion follows from (3.10). The proof of the rest is the same as the proof in Gill and Schumacher (1987, p. 293) with hazard rates replaced by transformation rates g_k. □

Part II

Tests for the proportional hazards regression

4 Testing the proportional hazards assumption for one covariate

Summary

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.

4.1 Introduction

We consider the Cox proportional hazards regression model (Cox, 1972) in the counting process formulation of Andersen and Gill (1982) (see also Andersen et al., 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 thesis, we assume that the conditions of Andersen and Gill (1982) guaranteeing certain asymptotic 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 the 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 this chapter, my 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.

One of the most important inference tools 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 et al. (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 $g(t) = \log t$. Another choice of $g(t)$ is the left-continuous Kaplan–Meier estimate (computed from the data ignoring covariates) which is the default method in the function `cox.zph` in the package 'survival' in R/S-PLUS. These methods and related plotting techniques are described in detail in Chapter 6 of Therneau and Grambsch (2000) (see also Grambsch and Therneau, 1994).

A compromise between the two classical tests (global and directional) is represented by Neyman's smooth tests, which I develop here. 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 et al. (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 chapter is as follows. In Section 4.2 we develop the smooth test of proportionality and establish asymptotic properties of the test statistic. Section 4.4 deals with the data-driven version of the test based on Schwarz's selection rule. In Section 4.5, our tests are compared through simulations with the other methods in various situations.

4.2 Smooth tests

As mentioned in the previous section, the null model

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{\beta^{\mathsf{T}} Z_i(t)\} \tag{4.1}$$

is embedded in

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{\beta^{\mathsf{T}} Z_i(t) + \theta^{\mathsf{T}}\xi_i(t)\}, \tag{4.2}$$

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

$$\psi_j(t) = \varphi_j(\Lambda_0(t)/\Lambda_0(\tau)), \qquad j = 1, \ldots, k \tag{4.3}$$

or

$$\psi_j(t) = \varphi_j(F_0(t)/F_0(\tau)), \qquad j = 1, \ldots, k. \tag{4.4}$$

Here $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ is the cumulative baseline hazard 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 (4.2). Then $C(\tau;\beta) := C(\tau;\beta,0)$ is the log partial likelihood for the Cox model (4.1). 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 (4.1) (or the restricted maximum partial likelihood estimate in (4.2) under $\theta = 0$). The score process for θ in the model (4.2) 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

$$U_2(\tau;\hat{\beta}) = U_2(\tau;\beta_0) - D(\tau;\beta^*)(\hat{\beta} - \beta_0), \tag{4.5}$$

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 (4.5) we obtain

$$n^{-1/2} U_2(\tau;\hat{\beta}) = n^{-1/2} U_2(\tau;\beta_0) - \{n^{-1} D(\tau;\beta^*)\}\{n J(\tau;\tilde{\beta})^-\}\{n^{-1/2} U_1(\tau;\beta_0)\}. \tag{4.6}$$

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 (4.6). 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

$$T_k = U_2(\tau;\hat\beta)^\mathsf{T} V(\tau;\hat\beta)^- U_2(\tau;\hat\beta), \tag{4.7}$$

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,

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

Let us make a note on conditions for the test to be consistent. Assume now that the proportional hazards assumption about the p-th covariate is violated, that is, the true model is

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\left\{\sum_{j=1}^{p-1}\beta_j Z_{ij}(t) + \beta_p(t)Z_{ip}(t)\right\}, \tag{4.8}$$

where the function $\beta_p(t)$ is nonconstant. Struthers and Kalbfleisch (1986, Theorem 2.1) (cf. Lin and Wei, 1989) showed that the partial likelihood estimator $\hat\beta$ (the solution to $U_1(\tau;\beta) = 0$) in the misspecified Cox model (4.1) converges in probability to some well-defined constant vector $\bar\beta$. This $\bar\beta$ is the solution to the limiting estimating equation

$$u_1(\tau;\bar\beta) := \int_0^\tau \left(s^{(1)}(t) - \frac{s^{(1)}(t;\bar\beta)}{s^{(0)}(t;\bar\beta)}s^{(0)}(t)\right)\lambda_0(t)dt = 0. \tag{4.9}$$

Here $s^{(k)}(t;\beta)$ and $s^{(k)}(t)$ are limits in probability of $n^{-1}S^{(k)}(t;\beta)$ and $n^{-1}S^{(k)}(t)$, respectively, with

$$S^{(k)}(t) = \sum_{i=1}^n Y_i(t)Z_i(t)^{\otimes k}\exp\left\{\sum_{j=1}^{p-1}\beta_j Z_{ij}(t) + \beta_p(t)Z_{ip}(t)\right\}.$$

The vector $u_1(\tau;\bar\beta)$ is the limit in probability of $n^{-1}U_1(\tau;\hat\beta)$ (under the true model (4.8)). Then the Neyman score $n^{-1}U_2(\tau;\hat\beta) = n^{-1}\int_0^\tau \psi(t)U_{1p}(dt;\hat\beta)$ converges in probability to $u_2(\tau;\bar\beta) = \int_0^\tau \psi(t)u_{1p}(dt;\bar\beta)$. Thus the variable $n^{-1}T_k$ converges in probability to the value

$u_2(\tau;\bar\beta)^\mathsf{T} v(\tau;\bar\beta)^{-1} u_2(\tau;\bar\beta)$. Therefore, Neyman's test is consistent against (4.8) if the basis functions are such that the condition

$$u_2(\tau;\bar\beta) = \int_0^\tau \psi(t)\left(s_p^{(1)}(t) - \frac{s_p^{(1)}(t;\bar\beta)}{s^{(0)}(t;\bar\beta)} s^{(0)}(t)\right)\lambda_0(t)dt \neq 0 \tag{4.10}$$

holds (at least one component is nonzero).

This condition may be interpreted as follows. The limiting estimating equations for the parameter $(\beta^\mathsf{T},\theta^\mathsf{T})^\mathsf{T}$ in the extended model (4.2) are $u_1(\tau;\beta,\theta)=0$, $u_2(\tau;\beta,\theta)=0$, where $u_j(\tau;\beta,\theta)$ are defined obviously (like $u_j(\tau;\beta)$, with $s^{(j)}(t;\beta)$ replaced by $s^{(j)}(t;\beta,\theta)$, the limit of $n^{-1}S^{(j)}(t;\beta,\theta)$). The left hand of the consistency condition (4.10) equals $u_2(\tau;\bar\beta,0)$. Thus, the condition means that if one estimates the extended model (4.2), he does not end up with the same result as in the null model (4.1), hence the basis functions at least partly explain the time-varying coefficient $\beta_p(t)$.

Finally some practical remarks.

The time transformation in the smooth functions $\psi_j(t)$ (in (4.3) or (4.4)) 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 ((4.3) or (4.4)) should we use? For survival data (i.e., for counting processes with at most one jump) I prefer the transformation (4.4) based on the baseline distribution function F_0. The reason is as follows. If we use the transformation (4.3), 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 (4.3), and it is better to use (4.4). 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 (4.3) 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.

4.3 Relation to principal components of integral tests

Let us investigate the relationship between Neyman's smooth tests and tests based on L^2 integrals of the score process. The analysis is elaborated in the special situation of models with

one covariate only (for multiple covariates the problem becomes more complicated because the asymptotic distribution of the test process $U_1(\cdot;\hat\beta)$ is complex). The comparison is based on the principal components analysis of the Cramér–von Mises and Anderson–Darling statistics. It is shown that the smooth test with the cosine basis is related to the Cramér–von Mises test and the smooth test with the Legendre basis to the Anderson–Darling test.

Consider test statistics of the form of an L^2 norm on $[0,\tau]$

$$\|U_1(\cdot;\hat\beta)/\Sigma_{11}(\tau;\hat\beta)^{1/2}\|^2_{2,w,R}=\int_0^\tau U_1(t;\hat\beta)^2/\Sigma_{11}(\tau;\hat\beta)w(R(t;\hat\beta))R(dt;\hat\beta),$$

where $R(t;\beta)=\Sigma_{11}(t;\beta)/\Sigma_{11}(\tau;\beta)$. The choice of the weight function $w\equiv 1$ yields the Cramér–von Mises statistic, $w(u)=(u(1-u))^{-1}, u\in(0,1)$ leads to the Anderson–Darling test.

Let us investigate the Cramér–von Mises statistic

$$T_{\mathrm{CM}}=\|U_1(\cdot;\hat\beta)/\Sigma_{11}(\tau;\hat\beta)^{1/2}\|^2_{2,1,R}=\int_0^\tau U_1(t;\hat\beta)^2/\Sigma_{11}(\tau;\hat\beta)R(dt;\hat\beta).$$

It is a well-known fact (following from a Taylor expansion like (4.6), see, e.g., Lemma 4.5.1 of Fleming and Harrington (1991) for details) that the standardised process $U_1(\cdot;\hat\beta)/\Sigma_{11}(\tau;\hat\beta)^{1/2}$ converges in distribution to a zero-mean continuous Gaussian process which is in the one covariate model distributed as the time-transformed Brownian bridge $B(r(\cdot;\beta_0))$. Here B is the standard Brownian bridge on $[0,1]$ and $r(t;\beta)=\sigma_{11}(t;\beta)/\sigma_{11}(\tau;\beta)$ continuously maps $[0,\tau]$ on $[0,1]$. The Karhunen–Loève decomposition of the limiting process $B(r(\cdot;\beta_0))$ is

$$B(r(\cdot;\beta_0))=\sum_{j=1}^\infty \lambda_j^{1/2}c_j l_j(r(\cdot;\beta_0)),$$

where the series converges in L^2, uniformly on $[0,\tau]$, specifically

$$\sup_{t\in[0,\tau]}\mathsf{E}\left\{\left(B(r(t;\beta_0))-\sum_{j=1}^k\lambda_j^{1/2}c_jl_j(r(t;\beta_0))\right)^2\right\}\xrightarrow{k\to\infty}0.$$

Here $l_j(u)=\sqrt{2}\sin(j\pi u), u\in[0,1]$ are orthonormal eigenfunctions and $\lambda_j=1/(j\pi)^2$ corresponding eigenvalues of the covariance kernel $k(u,v)=\mathrm{cov}(B(u),B(v))=u\wedge v-uv$ of the standard Brownian bridge, i.e., $\int_0^1 k(u,v)l_j(v)dv=\lambda_j l_j(u)$, $u\in[0,1]$. The variables

$$c_j=\lambda_j^{-1/2}\int_0^\tau B(r(t;\beta_0))l_j(r(t;\beta_0))r(dt;\beta_0),\quad j=1,2,\ldots$$

are independent standard normal. See, for instance, Section 1.4 of Ash and Gardner (1975) for details. Then the limiting Cramér–von Mises statistic

$$T^\infty_{\mathrm{CM}}=\|B(r(\cdot;\beta_0))\|^2_{2,1,r}=\int_0^\tau B(r(\cdot;\beta_0))^2r(dt;\beta_0)$$

is represented as

$$T^\infty_{\mathrm{CM}}=\sum_{j=1}^\infty\lambda_jc_j^2$$

(the series converges in L^2, hence in distribution). The principal components c_j^2 are independent χ_1^2-distributed.

Integrating by parts, we may rewrite their empirical counterparts

$$C_j = \lambda_j^{-1/2} \int_0^\tau U_1(t;\hat\beta)/\Sigma_{11}(\tau;\hat\beta)^{1/2} l_j(R(t;\hat\beta))R(dt;\hat\beta) \\ = \int_0^\tau \varphi_j(R(t;\hat\beta))U_1(dt;\hat\beta)/\Sigma_{11}(\tau;\hat\beta)^{1/2},$$

where $\varphi_j(u) = \sqrt{2}\cos(j\pi u)$. Then we get

$$T_{\mathrm{CM}} = \sum_{j=1}^{\infty} \lambda_j C_j^2. \tag{4.11}$$

If the series with the decreasing weights $\lambda_j = (j\pi)^{-2}$ is replaced by the finite sum $\sum_{j=1}^k C_j^2$ with equal weights, the resulting statistic asymptotically coincides with the statistic (4.7) of Neyman's smooth test with $\psi_j(t) = \varphi_j(R(t;\hat\beta))$. This is seen by observing that

$$C_j = \int_0^\tau \varphi_j(R(t;\hat\beta))U_1(dt;\hat\beta)/\Sigma_{11}(\tau;\hat\beta)^{1/2} = U_{2j}(\tau;\hat\beta)/\Sigma_{11}(\tau;\hat\beta)^{1/2} \\ = U_{2j}(\tau;\hat\beta)/V_{jj}(\tau;\hat\beta)^{1/2} + o_P(1)$$

because by orthogonality of the functions φ_j associated to the eigenfunctions l_j the limiting variance matrix $v(\tau;\beta_0)$ is diagonal with $\sigma_{11}(\tau;\beta_0)$ on the diagonal (because $\sigma_{22}(\tau;\beta_0) = \mathrm{diag}[\sigma_{11}(\tau;\beta_0)]$ and by orthogonality of φ_j and 1 it is $\sigma_{21}(\tau;\beta_0) = 0$).

As for the Anderson–Darling test (that is, $w(u) = (u(1-u))^{-1}$), the eigenfunctions l_j and eigenvalues λ_j satisfying the integral equation $\int_0^1 k(u,v)l_j(v)w(v)dv = \lambda_j l_j(u)$ are orthonormal associated Legendre polynomials on $[0,1]$ multiplied by $w(u)^{-1/2}$ and $\lambda_j = (j(j+1))^{-1}$ (again rapidly decreasing). The functions φ_j are then the orthonormal Legendre polynomials on $[0,1]$, and we can see the relation between the Anderson–Darling test and the smooth test with the Legendre basis.

Thus, in addition to the embedding idea we have seen another motivation for Neyman's smooth tests as tests based on the truncated series of L^2 integral tests. (Note that the L^2 tests are related to Neyman's tests with the basis functions with the time transformation $R(t;\beta)$, not the previously described transformation based on the distribution function.)

4.4 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 (4.2), where k was fixed (chosen prior to testing). Simulations (reported in Section 4.5) 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 the 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 (4.7) for the k-dimensional alternative. Then the selection rule is

$$S = \operatorname*{arg\,max}_{k\in\{1,\ldots,d\}} \{T_k - k\log n\}. \tag{4.12}$$

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 the asymptotic distribution of the statistic with dimension selected by Schwarz's rule. Under the null the selection rule is asymptotically concentrated in 1, the smallest possible dimension, i.e., $\Pr[S = 1] \xrightarrow[n\to\infty]{} 1$. This is apparent from the fact that for $k = 2, \ldots, d$

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

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

It will be seen in the Monte Carlo simulations in Section 4.5 that the χ^2_1 approximation is not accurate in small samples and the two-term approximation (1.6) derived in Section 1.4 gives better results.

4.5 Simulation study

We investigate the performance of the proposed tests through simulations. Our tests are compared with (1) tests based on various functionals of the score process (of the Kolmogorov–Smirnov, hereafter KS, Cramér–von Mises, CM, and Anderson–Darling, AD, type), and (2) the test of Grambsch and Therneau (1994) (GT) mentioned in the introduction (this is the method provided in R as `cox.zph`, testing submodels in models extended by artificial covariates equal to the original covariates times the Kaplan–Meier estimate). The standard tests were thoroughly examined by Kvaløy and Neef (2004) 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 used in all simulations. The time transformation for the basis functions is in the form (4.4).

For the null distribution of the data-driven test statistic T_S we consider both the χ^2_1 approximation and the improved approximation (1.6).

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 number of simulated paths which have the same asymptotic distribution as the score process; the number of the paths is 1000 everywhere). Another possibility is based on Khmaladze's transformation (Khmaladze, 1981; Martinussen and Scheike, 2006, Appendix A) which consists of transforming the test process to a process which is asymptotically a Gaussian martingale. My experience in a different regression context (Kraus, 2004) is that the behaviour of both methods (simulation and transformation) is similar.

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 4.1 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 Kvaløy and Neef (2004). Results are reported in Table 4.1. 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 (1.6) 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 Kvaløy and Neef (2004) 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 Kvaløy and Neef, 2004) as well as with $\mathrm{U}(0, 1)$ censoring. The model with nonmonotonic hazard ratios

Table 4.1: 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
	GT	0.038	0.039	0.040	0.042

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 Kvaløy and Neef, 2004). Estimated rejection probabilities are given in Table 4.2.

In Table 4.2 we can see the effect 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.

In comparison to the score process based tests, in this situation our test is less powerful for detecting monotonic deviations from proportionality but more powerful for detecting non-monotonic hazard ratios. The performance of the GT test is very good against monotonic hazard ratios but the test does not perform well under the nonmonotonic alternative. This is not surprising as this test is based on the monotonic modelling of the coefficient (by the Kaplan–Meier curve).

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 $\lambda(t) = \exp(0.5tZ_1 + Z_2 - 8)$.

Table 4.2: 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
	GT	0.409	0.236	0.108	0.070

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 Kvaløy and Neef (2004). The second model is $\lambda(t) = \exp(\beta(t)Z_1 + Z_2 - 8)$, 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 4.3 and 4.4.

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 4.3 and 4.4 should be looked at with caution especially for high values of ρ.

As for Z_1, a similar behaviour as in the one covariate situation of Table 4.2 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 mainly when there is high association between the covariates, 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 seem to be more

Table 4.3: 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
	GT	0.395	0.034	0.370	0.035	0.283	0.028

Table 4.4: 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
	GT	0.224	0.057	0.220	0.052	0.180	0.051

powerful for detecting monotonically time-varying effects (see Table 4.3). In the nonmonotonic situation of Table 4.4, both kinds of tests behave similarly.

To summarise, the simulation study showed that the proposed smooth test and its data-driven version could be a reasonable alternative to the tests of the proportional hazards assumption based on functionals of the score process. The GT test is good at detecting monotonic alternatives but may fail in nonmonotonic cases. Although the new procedure does not universally dominate the standard methods, I believe that the proposed approach is worth studying.

In simulations some doubt has been cast upon the accuracy of the procedures for situations with multiple covariates. It is a challenge to improve the tests to make them more capable to distinguish which covariates are proportional and which not. In Tables 4.3 and 4.4 we have seen that for highly correlated covariates the smooth tests (as well as the score process tests) are not reliable for such a discrimination. In this regard, the behaviour of the GT test is better. The next chapter addresses these issues more thoroughly.

5 Identifying nonproportional covariates in the Cox model

Summary

The problem of testing whether an individual covariate in the Cox model has a proportional (i.e., time-constant) effect on the hazard is dealt with. Two existing methods are considered: one is based on the component of the score process and the other is Neyman's smooth test. Simulations show that when the model contains both proportional and nonproportional covariates, these methods are not reliable tools for discrimination. A simple, yet effective solution is proposed based on smooth modelling of the effects of the covariates not in focus.

5.1 Introduction

Consider the Cox proportional hazards regression model (Cox, 1972) for right-censored survival data in the form

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{\beta^\mathsf{T} Z_i(t)\} \tag{5.1}$$

(Andersen and Gill, 1982). 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-vector of covariates, $\lambda_0(t)$ stands for an unknown baseline hazard function, and β is a vector of unknown regression coefficients.

In the Cox model the key assumption is proportionality of the effects of the covariates which means that the hazard ratio for two individuals does not depend on time. The assumption is not satisfied, for instance, when some of the coefficients $\beta_1, \ldots, \beta_p$ varies with time.

The aim of this chapter is to study methods of assessment of the proportional hazards assumption for a single covariate, say the p-th one. More specifically, we wish to test the hypothesis that the coefficient β_p is constant against the alternative of time-varying $\beta_p(t)$.

The problem of existing methods (namely score process based tests and Neyman-type smooth tests) is that they cannot distinguish reliably which covariates are proportional and which not. In models with both proportional and nonproportional covariates, the hypothesis of proportionality is often rejected even for the proportional covariate, that is, the size of the test dramatically exceeds the nominal level. Therefore, these tests serve as a tool for the overall assessment of proportionality rather than for individual covariate checks. I propose an improvement that consists of modelling the effects of the other covariates, which are not of interest, as linear combinations of some smooth functions. This makes the test more precise in identifying nonproportional covariates.

Simulation results of Section 5.2 warn against the use of methods derived under the valid proportional hazards model for testing proportionality of individual covariates. In Section 5.3 I present a solution whose performance is investigated through simulations in Section 5.4.

5.2 Warning against individual covariate tests

Two methods already introduced in the previous chapter are considered. Recall that tests based on the score process

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

use the Kolmogorov–Smirnov, Cramér–von Mises or Anderson–Darling statistic computed from the component $U_{1p}(t;\hat\beta)$, which reflects departures from the proportionality of the p-th covariate. For Neyman's smooth tests the original model (5.1) is embedded in the k-dimensional model

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{\beta^\mathsf{T} Z_i(t) + \theta^\mathsf{T}\varphi(F_0(t)/F_0(\tau))Z_{ip}(t)\}, \tag{5.2}$$

where F_0 is the distribution function associated with the baseline hazard λ_0 (in practice, F_0 is replaced by an estimator $\hat F_0$), and $\varphi = (\varphi_1,\dots,\varphi_d)^\mathsf{T}$ are some bounded functions in $L_2[0,1]$ such that $\{1,\varphi_1,\dots,\varphi_d\}$ is a set of linearly independent functions, as discussed previously. The smooth test is the score test of $\theta = 0$ against $\theta \neq 0$. Here I work only with a fixed choice of d (but the results of this chapter may be used for the data-driven test as well).

It may be misleading and dangerous to draw conclusions about proportionality of individual covariates from these tests in models with multiple covariates. To illustrate the extent of this problem, I performed a simulation study.

Let us consider three models, all of them have two covariates. In the first model of the form

$$\lambda(t) = \exp\{0.7Z_1 + 0.3Z_2\} \tag{5.3}$$

both covariates have proportional effects. The other two models have one covariate (Z_1) with a nonproportional effect and one (Z_2) proportional. The models follow the form

$$\lambda(t) = \exp\{0.5tZ_1 + Z_2 - 8\} \tag{5.4}$$

and

$$\lambda(t) = \exp\{\beta(t)Z_1 + Z_2 - 8\}, \tag{5.5}$$

where $\beta(t) = 0.4 + 0.7\times 1_{[1.2,2]}(t)$. The coefficient of Z_1 in (5.4) is monotonic, whereas in (5.5) it is not. In both models the variables Z_1, Z_2 are jointly normal with expectation 4, variance 1 and various values of correlation ρ. Independent censoring times were $\mathrm{U}(0,5)$ distributed in the models (5.3) and (5.4) (giving for all of the values of correlation the censoring rate about 24 % and 45 % in (5.3) and (5.4), respectively) and constant equal to 5 in the model (5.5) (about 31 % censoring).

I repeatedly generated samples of $n = 200$ observations and estimated rejection probabilities. The number of Monte Carlo runs in each situation was 5000 giving estimates of rejection probabilities with standard deviation at most $\sqrt{0.5\times(1-0.5)/5000} = 0.007$. The smooth test (denoted in the tables as T_d) was used with the dimension $d = 3,4,5,6$. The Kolmogorov–Smirnov (KS), Cramér–von Mises (CM), and Anderson–Darling test were performed using the simulation approximation of Lin et al. (1993) (with 1000 simulated paths of the score process).

Results for the model (5.3) are reported in Table 5.1 (results for Z_1 are given only, for Z_2 they are similar). The null hypothesis of proportionality is satisfied and the estimates of

Table 5.1: Estimated rejection probabilities on the nominal level of 5 % for the covariate Z_1 in the model $\lambda(t) = \exp\{0.7Z_1 + 0.3Z_2\}$ with $\text{cor}(Z_1, Z_2) = \rho$.

	$\rho = 0$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.9$
T_3	0.054	0.057	0.055	0.050	0.055
T_4	0.052	0.057	0.060	0.050	0.059
T_5	0.053	0.055	0.055	0.054	0.059
T_6	0.060	0.055	0.055	0.056	0.059
KS	0.054	0.051	0.053	0.050	0.052
CM	0.052	0.044	0.050	0.049	0.052
AD	0.050	0.043	0.046	0.047	0.050

Table 5.2: Estimated rejection probabilities on the nominal level of 5 % for the proportional covariate Z_2 in the model $\lambda(t) = \exp\{0.5tZ_1 + Z_2 - 8\}$ with $\text{cor}(Z_1, Z_2) = \rho$.

	$\rho = 0$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.9$
T_3	0.136	0.070	0.063	0.118	0.265
T_4	0.125	0.068	0.067	0.116	0.231
T_5	0.116	0.066	0.064	0.105	0.212
T_6	0.112	0.067	0.065	0.098	0.190
KS	0.149	0.061	0.085	0.181	0.382
CM	0.157	0.057	0.080	0.192	0.430
AD	0.149	0.053	0.076	0.186	0.425

rejection probabilitites are close to the nominal level of 5 %, so everything seems to be all right.

Results for the models (5.4) and (5.5) are shown in Tables 5.2 and 5.3 (results for Z_2 which satisfies the null hypothesis are reported only). As the hypothesis of proportionality of Z_2 is valid, the figures in Tables 5.2 and 5.3 should be close to the nominal level of 5 %. However, it turns out that in some cases the level is dramatically exceeded. Some of the figures are really alarming, especially (but not only) in the case of highly associated covariates.

The reason is that the score process method and the smooth method are valid only under the assumption of time-constancy of the effects of all the other covariates. When the proportionality is violated for some (nuisance) covariate that is not of interest, the techniques become unreliable. The procedures can indicate that the proportionality is not valid but are not capable to distinguish which covariate is 'guilty' and which not. This phenomenon has been previously pointed out, e.g., by Scheike and Martinussen (2004, pp. 58–59). My simulations show that the problem is serious even if the covariates are independent which was not seen so markedly in their simulation study. Also, Kvaløy and Neef (2004, p. 147) conjectured: "[. . .] intuitively, this should mainly be a problem in cases with strongly correlated covariates [. . .] Thus as long as the variation of other covariates is fairly nonsystematic compared to the covariate under examination a possible nonproportionality in some of the other covariates should not play any important role." My simulation illustrates that it is not advisable to rely on this intuition.

Let us investigate these problems formally. In Section 4.2 we saw that the partial likelihood

Table 5.3: Estimated rejection probabilities on the nominal level of 5 % for the proportional covariate Z_2 in the model $\lambda(t) = \exp\{\beta(t)Z_1 + Z_2 - 8\}$ ($\beta(t) = 0.4 + 0.7 \times 1_{[1.2,2]}(t)$) with correlation ρ between Z_1 and Z_2.

	$\rho = 0$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.9$
T_3	0.091	0.068	0.086	0.152	0.297
T_4	0.085	0.060	0.078	0.142	0.276
T_5	0.086	0.060	0.072	0.136	0.256
T_6	0.078	0.058	0.075	0.128	0.244
KS	0.055	0.069	0.099	0.210	0.392
CM	0.052	0.068	0.101	0.200	0.373
AD	0.053	0.066	0.095	0.188	0.341

estimator in the misspecified Cox model (4.8) (the fixed alternative with $\beta_1, \ldots, \beta_{p-1}$ constant and $\beta_p(t)$ time-varying) converges in probability to the solution $\bar{\beta}$ to the limiting estimating equation $u_1(\tau; \beta) = 0$, where

$$u_1(\tau; \bar{\beta}) = \int_0^\tau \left(s^{(1)}(t) - \frac{s^{(1)}(t; \bar{\beta})}{s^{(0)}(t; \bar{\beta})} s^{(0)}(t) \right) \lambda_0(t) dt = 0$$

(see eq. (4.9)). Now let the proportionality be violated for some of the other covariate(s) instead of (or in addition to) the p-th one, that is let the true model generally be

$$\lambda_i(t) = Y_i(t)\lambda_0(t) \exp\left\{ \sum_{j=1}^{p} \beta_j(t) Z_{ij}(t) \right\}.$$

Then $\hat{\beta}$ converges to some $\bar{\beta}$ satisfying $u_1(\tau; \bar{\beta}) = 0$, where $u_1(t; \beta)$ is of the same form as above but now with $s^{(k)}(t), k = 0, 1$, being defined as the uniform limits in probability of

$$n^{-1} S^{(k)}(t) = n^{-1} \sum_{i=1}^{n} Y_i(t) Z_i(t)^{\otimes k} \exp\left\{ \sum_{j=1}^{p} \beta_j(t) Z_{ij}(t) \right\}.$$

The p-th component of the score process, $n^{-1} U_{1p}(t; \hat{\beta})$, uniformly converges in probability to $u_{1p}(t; \bar{\beta})$ which in general is a nonzero function even if β_p is constant. Thus the tests based on the score process $n^{-1/2} U_{1p}(t; \hat{\beta})$ will reject with probability going to 1. Similarly, the Neyman score $n^{-1} U_2(\tau; \hat{\beta})$ converges to $u_2(\tau; \bar{\beta}) = \int_0^\tau \psi(t) u_{1p}(dt; \bar{\beta})$, which is generally nonzero, and hence also the smooth test asymptotically rejects in spite of the proportionality of the p-th covariate.

We see that generally the rejection probability of the tests supposedly checking the proportionality of the p-th covariate actually converges to 1 whenever any of the covariates is nonproportional, not necessarily the p-th one.

5.3 Improvement

The tests for individual covariates work correctly provided the model is correct for all the other covariates. Therefore, a simple idea seems to be reasonable: make the model for all

the other covariates correct enough to remove or diminish the influence of their potential nonproportionality on the test. This means to model the time-varying effects of the other covariates. To this end, I use smooth functions similarly as in Neyman's smooth tests.

Recall that the aim is to test proportionality of the p-th covariate.

Instead of the original model

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\left\{\sum_{j=1}^{p} Z_{ij}(t)\beta_j\right\}, \tag{5.6}$$

the null model now follows the form

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\left\{\sum_{j=1}^{p-1} Z_{ij}(t)\left(\beta_j + \sum_{k=1}^{d_j}\theta_{jk}\varphi_k(F_0(t)/F_0(\tau))\right) + Z_{ip}(t)\beta_p\right\}, \tag{5.7}$$

which allows for smoothly time-varying coefficients of all the covariates but the p-th one. This large model is an ordinary Cox model with artificial time-dependent covariates and with parameters $\beta_1, \ldots, \beta_p$ and $\theta_{jk},\ j = 1, \ldots, p-1,\ k = 1, \ldots, d_j$. The unknown time-transformation F_0 is estimated from the original model (5.6) and viewed as given when working with the large model (5.7). That is we in fact work with

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\left\{\sum_{j=1}^{p-1} Z_{ij}(t)\left(\beta_j + \sum_{k=1}^{d_j}\theta_{jk}\varphi_k(\hat{F}_0(t)/\hat{F}_0(\tau))\right) + Z_{ip}(t)\beta_p\right\}. \tag{5.8}$$

The proposed test procedure is as follows. First, one estimates the coefficients in the large model (5.8) by the standard partial likelihood method and then performs some of the tests. That is either the test based on the component of the score process corresponding to Z_{ip} in (5.8), or the smooth test which is the score test of significance of $\theta_{p1}, \ldots, \theta_{p,d_p}$ in

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\left\{\sum_{j=1}^{p-1} Z_{ij}(t)\left(\beta_j + \sum_{k=1}^{d_j}\theta_{jk}\varphi_k(\hat{F}_0(t)/\hat{F}_0(\tau))\right)\right.$$
$$\left. + Z_{ip}(t)\left(\beta_p + \sum_{k=1}^{d_p}\theta_{pk}\varphi_k(\hat{F}_0(t)/\hat{F}_0(\tau))\right)\right\}.$$

Here in the Z_{ip}-part $\hat{F}_0$ may be replaced by $\tilde{F}_0$ computed from (5.8) (which is closer to the true baseline distribution).

The tests are carried out in the same way as in the original versions of the tests: for the score process based test the simulation approximation can be used, and the distribution of the statistic of the smooth test is approximated by the $\chi^2_{d_p}$ distribution. A data-driven choice of d_p is possible.

The proposed approach practically works as is observed in the simulation study in the next section. To justify it theoretically one would have to let $d_1, \ldots, d_{p-1}$ tend to infinity at a suitable rate as n grows. The convergence must be fast enough to control the approximation error but not too fast to guarantee stability of estimation. It calls for further research to give conditions under which the Z_{ip}-component of the score process computed in (5.7) converges to a zero-mean Gaussian process. A similar problem was previously considered by Murphy and Sen (1991) who dealt with a histogram sieve estimator in the Cox model with time-varying coefficients.

Table 5.4: Estimated rejection probabilities on the nominal level 5 % in the model $\lambda(t) = \exp\{0.7Z_1 + 0.3Z_2\}$ with $\text{cor}(Z_1, Z_2) = \rho$. Various numbers of smooth functions for the other covariate.

		Z_1				Z_2			
		$d_2 = 0$	$d_2 = 2$	$d_2 = 3$	$d_2 = 4$	$d_1 = 0$	$d_1 = 2$	$d_1 = 3$	$d_1 = 4$
$\rho = 0$	T_3	0.054	0.057	0.058	0.058	0.056	0.060	0.062	0.063
	T_4	0.052	0.061	0.064	0.063	0.057	0.059	0.059	0.060
	T_5	0.053	0.061	0.065	0.064	0.054	0.058	0.060	0.062
	T_6	0.060	0.062	0.064	0.065	0.057	0.060	0.060	0.062
	KS	0.054	0.057	0.055	0.053	0.051	0.051	0.050	0.050
	CM	0.052	0.053	0.052	0.050	0.049	0.047	0.049	0.046
	AD	0.050	0.050	0.049	0.048	0.048	0.047	0.048	0.046
$\rho = 0.5$	T_3	0.055	0.063	0.064	0.066	0.054	0.059	0.063	0.064
	T_4	0.060	0.065	0.068	0.066	0.052	0.056	0.057	0.060
	T_5	0.055	0.062	0.064	0.066	0.056	0.059	0.059	0.060
	T_6	0.055	0.061	0.061	0.062	0.054	0.058	0.058	0.056
	KS	0.053	0.055	0.056	0.056	0.057	0.060	0.059	0.058
	CM	0.050	0.055	0.057	0.054	0.053	0.052	0.049	0.051
	AD	0.046	0.052	0.054	0.055	0.050	0.046	0.047	0.050
$\rho = 0.9$	T_3	0.055	0.058	0.055	0.059	0.048	0.051	0.057	0.058
	T_4	0.059	0.058	0.060	0.058	0.054	0.054	0.058	0.061
	T_5	0.059	0.060	0.061	0.061	0.051	0.051	0.056	0.057
	T_6	0.059	0.063	0.064	0.064	0.056	0.054	0.054	0.055
	KS	0.052	0.053	0.054	0.051	0.056	0.052	0.053	0.049
	CM	0.052	0.050	0.051	0.052	0.050	0.045	0.051	0.049
	AD	0.050	0.047	0.048	0.050	0.048	0.041	0.043	0.046

5.4 Simulations

The aim of the Monte Carlo study is to explore whether the proposed improvement works (i.e., whether the level is preserved) and how it influences the power.

The design of the study is the same as in Section 5.2. The tests were used with various numbers of smooth functions describing the effect of the covariate that is not tested (various values of d_2 for tests of Z_1 and d_1 for Z_2). Results for the models (5.3), (5.4) and (5.5) are displayed in Tables 5.4, 5.5 and 5.6.

Table 5.4 shows that the proposed modification works correctly in the situation where even the original test (without smooth modelling of the other covariates) was valid. In Tables 5.5 and 5.6, results for Z_2 show that the prescribed level of 5 % is preserved when the effect of the covariate Z_1 is modelled smoothly. In the models of the study, it was enough to use two smooth functions because the time-varying coefficient of Z_1 has a relatively simple form in both models. Generally, it may be necessary to use more basis functions.

The power results for Z_1 show that if one uses more smooth functions than necessary, the decrease of power is not dramatic. The factor that prevents us from including very large numbers of basis functions (i.e., new artificial covariates) is the sample size, hence the numerical stability.

Table 5.5: Estimated rejection probabilities on the nominal level 5 % in the model $\lambda(t) = \exp\{0.5tZ_1 + Z_2 - 8\}$ with $\text{cor}(Z_1, Z_2) = \rho$. Various numbers of smooth functions for the other covariate.

		Z_1				Z_2			
		$d_2 = 0$	$d_2 = 2$	$d_2 = 3$	$d_2 = 4$	$d_1 = 0$	$d_1 = 2$	$d_1 = 3$	$d_1 = 4$
$\rho = 0$	T_3	0.771	0.690	0.685	0.683	0.136	0.061	0.060	0.062
	T_4	0.726	0.644	0.638	0.640	0.125	0.062	0.059	0.061
	T_5	0.698	0.611	0.612	0.609	0.116	0.064	0.059	0.061
	T_6	0.663	0.577	0.574	0.576	0.112	0.059	0.055	0.059
	KS	0.797	0.738	0.735	0.736	0.149	0.048	0.050	0.050
	CM	0.855	0.808	0.809	0.810	0.157	0.047	0.048	0.047
	AD	0.861	0.813	0.817	0.814	0.149	0.045	0.047	0.046
$\rho = 0.5$	T_3	0.657	0.626	0.625	0.615	0.063	0.063	0.062	0.059
	T_4	0.605	0.582	0.577	0.572	0.067	0.060	0.060	0.060
	T_5	0.566	0.536	0.534	0.533	0.064	0.059	0.061	0.060
	T_6	0.529	0.494	0.498	0.494	0.065	0.056	0.059	0.061
	KS	0.698	0.675	0.672	0.674	0.085	0.060	0.053	0.052
	CM	0.774	0.754	0.748	0.746	0.080	0.055	0.051	0.049
	AD	0.777	0.758	0.755	0.754	0.076	0.053	0.048	0.047
$\rho = 0.9$	T_3	0.467	0.238	0.235	0.225	0.265	0.071	0.063	0.061
	T_4	0.414	0.210	0.200	0.200	0.231	0.068	0.066	0.062
	T_5	0.374	0.181	0.182	0.179	0.212	0.064	0.063	0.061
	T_6	0.341	0.164	0.174	0.165	0.190	0.061	0.059	0.059
	KS	0.554	0.266	0.254	0.263	0.382	0.075	0.059	0.057
	CM	0.633	0.340	0.325	0.334	0.430	0.070	0.054	0.052
	AD	0.641	0.336	0.323	0.335	0.425	0.061	0.049	0.048

Table 5.6: Estimated rejection probabilities on the nominal level 5 % in the model $\lambda(t) = \exp\{\beta(t)Z_1 + Z_2 - 8\}$ ($\beta(t) = 0.4 + 0.7 \times 1_{[1.2,2]}(t)$) with $\text{cor}(Z_1, Z_2) = \rho$. Various numbers of smooth functions for the other covariate.

		Z_1				Z_2			
		$d_2 = 0$	$d_2 = 2$	$d_2 = 3$	$d_2 = 4$	$d_1 = 0$	$d_1 = 2$	$d_1 = 3$	$d_1 = 4$
$\rho = 0$	T_3	0.662	0.569	0.547	0.569	0.091	0.051	0.053	0.052
	T_4	0.676	0.630	0.608	0.612	0.085	0.052	0.052	0.048
	T_5	0.651	0.612	0.595	0.594	0.086	0.051	0.056	0.049
	T_6	0.634	0.599	0.587	0.579	0.078	0.053	0.059	0.052
	KS	0.596	0.605	0.612	0.617	0.055	0.053	0.052	0.049
	CM	0.525	0.541	0.538	0.546	0.052	0.049	0.047	0.048
	AD	0.559	0.567	0.563	0.567	0.053	0.047	0.048	0.046
$\rho = 0.5$	T_3	0.570	0.539	0.499	0.483	0.086	0.050	0.052	0.052
	T_4	0.569	0.529	0.497	0.485	0.078	0.052	0.048	0.052
	T_5	0.534	0.511	0.481	0.467	0.072	0.051	0.050	0.050
	T_6	0.516	0.495	0.464	0.449	0.075	0.054	0.053	0.052
	KS	0.603	0.577	0.559	0.559	0.099	0.053	0.052	0.051
	CM	0.568	0.541	0.527	0.525	0.101	0.053	0.049	0.051
	AD	0.553	0.521	0.510	0.505	0.095	0.050	0.049	0.048
$\rho = 0.9$	T_3	0.452	0.200	0.172	0.149	0.297	0.057	0.058	0.060
	T_4	0.436	0.200	0.179	0.145	0.276	0.060	0.068	0.064
	T_5	0.412	0.200	0.177	0.152	0.256	0.065	0.068	0.065
	T_6	0.394	0.195	0.175	0.150	0.244	0.063	0.067	0.062
	KS	0.541	0.261	0.242	0.238	0.392	0.075	0.063	0.057
	CM	0.522	0.261	0.239	0.233	0.373	0.060	0.055	0.053
	AD	0.481	0.209	0.196	0.191	0.341	0.051	0.052	0.053

6 Global assessment of proportional hazards

Summary

The idea of smooth tests is extended to the overall verification of the proportional hazards assumption. Strategies for the data-driven selection of the smooth alternative are discussed.

6.1 Introduction

The goal is to verify the proportional hazards assumption globally, without a focus on a particular covariate. That is, the hypothesis of proportional hazards is to be tested against the alternative that there is a covariate with time-varying effect.

Lin et al. (1993) proposed Kolmogorov–Smirnov supremum tests based on combinations of components of the score process. For instance, the test process may be

$$\sum_{j=1}^{p} \{\Sigma_{11}(\tau;\hat{\beta})_{jj}\}^{-1/2} |U_{1j}(t;\hat{\beta})|.$$

The martingale simulation technique of Lin et al. (1993) is standardly used to approximate the p-value.

Another approach is the global version of the test of Grambsch and Therneau (1994) (see also Therneau and Grambsch, 2000, Chapter 6). Each possibly time-varying effect is written as a constant plus a function of time, i.e., $\beta_j(t) = \beta_j + \theta_j g_j(t)$. The significance of the p new covariates is tested by a chi-square test on p degrees of freedom. The functions $g_j(t)$ are typically some monotonic functions, such as $\log t$ or the Kaplan–Meier estimator. This method is offered in the standard R procedure `cox.zph`.

6.2 Global smooth test and selection procedures

Let us see how the ideas of the previous chapters extend to the present situation. It is straightforward. All p potentially time-varying coefficients are expressed as linear combinations of the basis functions $\psi_k(t)$, $k = 1, \ldots, d_j$ (which are some standard basis functions in transformed time, i.e., $\psi_k(t) = \varphi_k(F_0(t)/F_0(\tau))$, as before). The null proportional hazards model

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

is embedded in the model

$$\lambda_i(t) = Y_i(t)\lambda_0(t) \exp\left\{\beta^{\mathsf{T}} Z_i(t) + \sum_{j=1}^{p} \sum_{k=1}^{d_j} \theta_{j,k} \xi_{i,j,k}(t)\right\},$$

where the artificial covariates

$$\xi_{i,j,k}(t) = \psi_k(t) Z_{i,j}(t), \qquad i = 1, \dots n, \quad j = 1, \dots, p, \quad k = 1, \dots, d_j$$

are tested for significance by the partial likelihood score test with $\bar{d} = \sum_{j=1}^{p} d_j$ degrees of freedom.

Now the question is whether we can make a data-driven choice of the basis functions so as to avoid using too many redundant functions. Obviously, the artificial covariates

$$\xi_{i,1,1}(t), \dots, \xi_{i,1,d_1}(t), \dots, \xi_{i,p,1}(t), \dots, \xi_{i,p,d_p}(t)$$

cannot be naturally ordered according to the increasing complexity of the model they induce. Therefore, Schwarz's selection rule cannot pick out of nested models.

Instead, we may apply the idea of Claeskens and Hjort (2004) and let the selection rule search among all nonempty subsets of the set of the $\bar{d}$ basis functions. However, in the two-sample context, we saw in Table 1.3 that the all subsets selection did not lead to one of the main goals of the data-driven approach which is to avoid testing against an overcomplicated alternative. The behaviour of the test was akin to that of the test with a fixed set of basis functions in that the power decreased when the number of the functions increased above a number sufficiently describing the alternative. Some simulations not reported here indicated that in the present regression situation the behaviour was similar.

Therefore, I try a different procedure which is more similar to the nested subsets (order selection) procedure. Rather than all nonempty subsets of the set of $\bar{d}$ artificial time-dependent covariates, I consider nested subsets for each of the covariates. In other words, for each covariate the order of the approximation for the respective coefficient is selected. For each $j = 1, \dots, p$, the possibly time-varying coefficient of the j-th covariate may be approximated by a linear combination of the functions $\psi_1(t), \dots, \psi_{k_j}(t)$ for $k_j = 1, \dots, d_j$ or may be left constant ($k_j = 0$). Each of the possible alternatives is described by the vector of dimensions $(k_1, \dots, k_p)$, where $0 \le k_j \le d_j$, with at least one k_j positive, i.e., $\bar{k} = \sum_{j=1}^{p} k_j > 0$ (there must be at least one artificial covariate in the alternative, the empty set must be excluded because otherwise the BIC would consistently estimate the true empty model). This gives rise to the total number of $\prod_{j=1}^{p}(1 + d_j) - 1$ alternative models. Note that this is much less than $2^{\bar{d}} - 1$ nonempty sets when the all subsets search is used which considerably reduces the computational burden (e.g., for $p = 3$ and $d_j = 4$ this is 124 versus 4095 alternatives).

The selector $S = (S_1, \dots, S_p)$ is defined as the configuration that maximises the penalised score statistic over possible configurations $(k_1, \dots, k_p)$, i.e.,

$$S = \underset{\substack{(k_1,\dots,k_p):\\ 0 \le k_j \le d_j, \bar{k} > 0}}{\arg\max} \{T_{(k_1,\dots,k_p)} - \bar{k} \log n\},$$

where $T_{(k_1,\dots,k_p)}$ obviously denotes the partial likelihood score statistic of significance of the $(k_1, \dots, k_p)$ artificial covariates. The test is then based on T_S.

Under the null hypothesis, for all $(k_1, \dots, k_p)$ with $\bar{k} > 0$ the statistic $T_{(k_1,\dots,k_p)}$ is asymptotically χ^2 distributed with $\bar{k}$ degrees of freedom. Therefore, the selection rule asymptotically concentrates in one-dimensional models, i.e., $\Pr[\bar{S} = 1] \to 1$ as $n \to \infty$ (where $\bar{S} = \sum_{j=1}^{p} S_j$) because

$$\Pr[\bar{S} = 1] \ge \Pr\left[\max_{(k_1,\dots,k_p):\bar{k}=1} \{T_{(k_1,\dots,k_p)} - \log n\} > \max_{(k_1,\dots,k_p):\bar{k}=2} \{T_{(k_1,\dots,k_p)} - 2 \log n\} \right] \to 1.$$

Hence, asymptotically only singletons survive the selection, and among them the one which maximises the score statistic is selected. Therefore, T_S converges in distribution to the maximum of p generally dependent χ^2 variables with 1 degree of freedom. This distribution does not depend on the maximal dimensions $d_1, \ldots, d_p$, and thus the power of the test is expected to be stable with respect to the choice $d_1, \ldots, d_p$. (This contrasts with the strategy based on all subsets where the asymptotic distribution is the maximum of $\bar{d}$ chi-square variables.) The asymptotic distribution is easily approximated by simulations.

The small sample accuracy of the asymptotic max-chi-square approximation is a question. In a limited set of simulations (not presented here) for models with two covariates (the same models as in Section 6.3) the approximation seemed to perform well (the true level under the null hypothesis was not far from the nominal level). However, the special case of a model with one covariate coincides with the situation of Table 4.1. In that case the asymptotic approximation was found quite unreliable. Therefore, in general I expect a similar degree of inaccuracy in models with multiple covariates.

Rather, we may consider a two-term approximation similar to that derived in Section 1.4 for nested subsets. The idea is analogous: write

$$\Pr[T_S \le x] = \Pr[T_S \le x, \bar{S} = 1] + \Pr[T_S \le x, \bar{S} = 2] + \Pr[T_S \le x, \bar{S} > 2].$$

Instead of ignoring the last two terms (which tend to zero under the hypothesis), we neglect only the last term.

Let us further approximate the first two summands on the right-hand side of the above identity. The event $[\bar{S} = 1]$ means that some of one-dimensional alternative models wins over all models of higher dimension. It will be approximated by the event that some one-dimensional model wins against all two-dimensional models. Thus, denoting

$$T_1^* = \max_{(k_1,\ldots,k_p):\bar{k}=1} T_{(k_1,\ldots,k_p)}, \qquad T_2^* = \max_{(k_1,\ldots,k_p):\bar{k}=2} T_{(k_1,\ldots,k_p)},$$

I use

$$[\bar{S} = 1] \doteq [T_1^* - \log n \ge T_2^* - 2\log n] = [T_2^* - T_1^* \le \log n].$$

Analogously, $[\bar{S} = 2]$ is approximately the event that there exists a two-dimensional model which beats all models of dimension 1 (instead of all models of dimension different from 2), i.e.,

$$[\bar{S} = 2] \doteq [T_1^* - \log n < T_2^* - 2\log n] = [T_2^* - T_1^* > \log n].$$

Hence, finally, the distribution of the test statistic is approximated by

$$\Pr[T_S \le x] \doteq \Pr[T_1^* \le x, T_2^* - T_1^* \le \log n] + \Pr[T_2^* \le x, T_2^* - T_1^* > \log n]. \qquad (6.1)$$

The variables T_1^*, T_2^* are more complicated than T_1, T_2 in Section 1.4 and we are not able to obtain an explicit formula like (1.6). But the above probabilities are easily estimated by simulation from the asymptotic multivariate normal distribution of the score.

6.3 Simulation results

I carried out a small simulation study to see the performance of the procedures. I used the same simulation design as in Section 5.2: one model (eq. (5.3)) satisfies the proportional

Table 6.1: Estimated rejection probabilities of global tests of proportional hazards. Correlation ρ between covariates Z_1, Z_2. Nominal level 5 %. Estimates based on 10 000 repetitions under H_0 and 5000 repetitions under alternatives.

	Constant β_1 (H_0)		Monotonic $\beta_1(t)$		Nonmonotonic $\beta_1(t)$	
	$\rho = 0$	$\rho = 0.7$	$\rho = 0$	$\rho = 0.7$	$\rho = 0$	$\rho = 0.7$
$T_{(4,4)}$	0.0591	0.0597	0.592	0.387	0.546	0.401
T_S	0.0622	0.0624	0.732	0.521	0.630	0.469
KS	0.0531	0.0554	0.626	0.481	0.358	0.489
GT	0.0429	0.0406	0.744	0.535	0.296	0.331

hazards assumption, in the second one (eq. (5.4)) the coefficient $\beta_1(t)$ is monotonic and in the third model (5.5) the function $\beta_1(t)$ is nonmonotonic. The distribution of covariates, the censoring pattern and the sample sizes ($n = 200$) were the same as in Section 5.2. The same randomly generated data sets were used. The smooth tests were performed with $d = (d_1, d_2) = (4, 4)$. Table 6.1 reports a part of these results. p-values for the Kolmogorov–Smirnov test were computed from 1000 simulated paths, for the data-driven smooth test the two-term approximation (6.1) was used.

All the tests seem to have size reasonably close to the nominal level. Under the monotonic alternative, the data-driven test has higher power than the test with the fixed number of smooth functions. The monotonic function $\beta_1(t)$ is well described by one basis function (a linear function) and the inclusion of many other functions leads to a decay of the power. In this situation, the data-driven smooth test and the GT test have comparable power. Under the nonmonotonic alternative, the behaviour of the GT test is worse than the behaviour of the smooth tests. This is not surprising as the GT test models each coefficient by one monotonic function which reveals the time-varying nature of $\beta_1(t)$ only partly. The power of the smooth tests and the GT test decreases as the correlation between the covariates increases. For the Kolmogorov–Smirnov test this effect is not seen.

Concluding remarks

We have seen how Neyman's idea of smooth tests and Ledwina's idea of data-driven tests can be extended to the field of survival analysis. Smooth tests and their data-driven versions can be considered as a serious competitor to many existing procedures, especially when one seeks a test without a clear advance idea of the alternative. Certainly, methods of this type can be developed for many other branches of statistics, though now they do not seem to be used often.

Here I outline possible directions for future work in the context of survival analysis.

Two-sample comparisons may be extended to K samples. The construction of Neyman's smooth test would be straightforward. To compare K survival distributions or K cumulative incidence functions for a particular cause of failure, one sample would be taken as a refernce group and the difference of the other samples would be modelled by several basis functions. For the semiparametric proportional rate transformation model, the possibly time-varying coefficients of all but the reference group would be expressed as combinations of basis functions. In $K > 2$ samples, the possible alternatives are not nested (naturally ordered) but data-driven tests could be preformed using one of the classes of subsets discussed in Chapter 6.

In Chapter 1, permutations and the bootstrap were used. It may be worthwhile to develop resampling techniques for smooth tests for other two-sample or regression situations. In the first chapter, the resampling was obvious because the survival distributions were equal under the null hypothesis. This is not the case for the other two-sample situations: the null hypothesis does not imply the full equivalence of the two distributions, and thus the null distribution cannot be approximated by random sampling (with or without replacement) from the pooled sample. In these situations or in the regression context, the permutation approach seems to be applicable in a different way. The score vectors involved in all of the score tests in this thesis are in fact the empirical covariance between the artificial covariates describing the smooth alternative and (martingale) residuals from the null model without these covariates. Under the null hypothesis of insignificance, these covariates are unrelated to the residuals, and thus permutations are a natural way to assess the association between the smooth covariates and the residuals. I have limited experience with this approach in the two-sample proportional rate models of Chapter 3 where it seems to work.

Theoretical asymptotic properties of data-driven smooth tests for censored data could be investigated. It would be interesting to know whether some optimality results similar to those of Ducharme and Ledwina (2003) could be obtained.

A Software implementation

I provide a software implementation. I have collected my programs into two add-on libraries for the R system for statistical computing (R Development Core Team, 2007).

These packages are made available to the public through CRAN (Comprehensive R Archive Network, `http://cran.r-project.org/`). They can be installed directly from a CRAN server by a standard procedure depending on the user's computer environment (e.g., from within R by a command or through a menu item in a graphical user interface, or from the system command-line).

A.1 Package 'surv2sample'

This package covers topics of Chapters 1, 2 and 3. It provides functions for performing Neyman's smooth tests proposed in those chapters as well as other two-sample tests mentioned there. Some plotting routines are provided too. A webpage for the package is located at `http://www.davidkraus.net/surv2sample/`.

A.2 Package 'proptest'

The package implements the proportional hazards tests of Chapters 4, 5 and 6, i.e., data-driven smooth tests and score process based tests for global and individual covariate tests. A webpage for this package may be found at `http://www.davidkraus.net/proptest/`.

Bibliography

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.

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Statist.*, 23, 193–212.

Antoch, J., Hušková, M., Janic, A. and Ledwina, T. (2007). Data driven rank test for the change point problem. *Metrika*. To appear.

Ash, R. B. and Gardner, M. F. (1975). *Topics in Stochastic Processes*. Academic Press, New York.

Bagdonavičius, V., Levuliene, R. J., Nikulin, M. and Zdorova-Cheminade, O. (2004). Tests for equality of survival distributions against non-location alternatives. *Lifetime Data Anal.*, 10, 445–460.

Bagdonavičius, V. and Nikulin, M. (1999). Generalized proportional hazards model based on modified partial likelihood. *Lifetime Data Anal.*, 5, 329–350.

Bagdonavičius, V. and Nikulin, M. (2000). On goodness-of-fit for the linear transformation and frailty models. *Statist. Probab. Lett.*, 47, 177–188.

Bagdonavičius, V. and Nikulin, M. (2001). *Accelerated life models. Modeling and Statistical Analysis*. Chapman & Hall/CRC, Boca Raton.

Bajorunaite, R. and Klein, J. P. (2007). Two-sample tests of the equality of two cumulative incidence functions. *Comput. Statist. Data Anal.*, 51, 4269–4281.

Chen, K., Jin, Z. and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika*, 89, 659–668.

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

Collett, D. (2003). *Modelling Survival Data in Medical Research*. Chapman & Hall/CRC, Boca Raton.

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

Dabrowska, D. M. and Doksum, K. A. (1988). Estimation and testing in a two-sample generalized odds-rate model. *J. Amer. Statist. Assoc.*, 83, 744–749.

Dauxois, J.-Y. and Kirmani, S. N. U. A. (2003). Testing the proportional odds model under random censoring. *Biometrika*, 90, 913–922.

Ducharme, G. R. and Ledwina, T. (2003). Efficient and adaptive nonparametric test for the two-sample problem. *Ann. Statist.*, 31, 2036–2058.

Durbin, J. and Knott, M. (1972). Components of Cramér–von Mises statistics. I. *J. Roy. Statist. Soc. Ser. B*, 34, 290–307.

Durbin, J., Knott, M. and Taylor, C. C. (1975). Components of Cramér–von Mises statistics. II. *J. Roy. Statist. Soc. Ser. B*, 37, 216–237.

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

Fleming, T. R., Harrington, D. P. and O'Sullivan, M. (1987). Supremum versions of the log-rank and generalized Wilcoxon statistics. *J. Amer. Statist. Assoc.*, 82, 312–320.

Gill, R. D. (1980). *Censoring and stochastic integrals.* Mathematical Centre Tracts 124. Mathematisch Centrum, Amsterdam.

Gill, R. D. and Schumacher, M. (1987). A simple test of the proportional hazards assumption. *Biometrika*, 74, 289–300.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81, 515–526.

Gray, R. J. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *Ann. Statist.*, 16, 1141–1154.

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, 69, 553–566.

Heller, G. and Venkatraman, E. S. (1996). Resampling procedures to compare two survival distributions in the presence of right-censored data. *Biometrics*, 52, 1204–1213.

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

Janic-Wróblewska, A. and Ledwina, T. (2000). Data driven rank test for two-sample problem. *Scand. J. Statist.*, 27, 281–397.

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

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data.* Wiley, New York.

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.

Khmaladze, E. V. (1981). Martingale approach in the theory of goodness-of-fit tests. *Teor. Veroyatnost. i Primenen.*, 26, 246–265. In Russian. English translation in *Theory Probab. Appl.*, 26, 240–257.

Kraus, D. (2004). Goodness-of-fit inference for the Cox–Aalen additive–multiplicative regression model. *Statist. Probab. Lett.*, 70, 285–298.

Kraus, D. (2007a). Adaptive Neyman's smooth tests of homogeneity of two samples of survival data. Research Report 2187, Institute of Information Theory and Automation, Prague. Submitted.

Kraus, D. (2007b). Checking proportional rates in the two-sample transformation model. Research Report 2203, Institute of Information Theory and Automation, Prague. Submitted.

Kraus, D. (2007c). Data-driven smooth tests of the proportional hazards assumption. *Lifetime Data Anal.*, 13, 1–16.

Kraus, D. (2007d). Smooth tests of equality of cumulative incidence functions in two samples. Research Report 2197, Institute of Information Theory and Automation, Prague. Submitted.

Kraus, D. (2008). Identifying nonproportional covariates in the Cox model. *Comm. Statist. Theory Methods*, 37. To appear.

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.

Lee, J. W. (1996). Some versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics*, 52, 721–725.

Lin, D. Y. (1997). Non-parametric inference for cumulative incidence functions in competing risks studies. *Stat. Med.*, 16, 901–910.

Lin, D. Y. and Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *J. Amer. Statist. Assoc.*, 84, 1074–1078.

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. and Scheike, T. H. (2006). *Dynamic Regression Models for Survival Data.* Springer, New York.

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.

Murphy, S. A., Rossini, A. J. and van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *J. Amer. Statist. Assoc.*, 92, 968–976.

Murphy, S. A. and Sen, P. K. (1991). Time-dependent coefficients in a Cox-type regression model. *Stochastic Process. Appl.*, 39, 153–180.

Neuhaus, G. (1993). Conditional rank tests for the two-sample problem under random censorship. *Ann. Statist.*, 21, 1760–1779.

Neyman, J. (1937). "Smooth" test for goodness of fit. *Skandinavisk Aktuarietidskrift*, 20, 149–199.

Pecková, M. and Fleming, T. R. (2003). Adaptive test for testing the difference in survival distributions. *Lifetime Data Anal.*, 9, 223–238.

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.

Peng, L. and Fine, J. P. (2007). Nonparametric quantile inference with competing-risks data. *Biometrika*, 94, 735–744.

Pepe, M. S. (1991). Inference for events with dependent risks in multiple endpoint studies. *J. Amer. Statist. Assoc.*, 86, 770–778.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Rayner, J. C. W. and Best, D. J. (1989). *Smooth Tests of Goodness of Fit*. Oxford University Press, New York.

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.

Schumacher, M. (1984). Two-sample tests of Cramér–von Mises- and Kolmogorov–Smirnov-type for randomly censored data. *Int. Stat. Rev.*, 52, 263–281.

Sengupta, D., Bhattacharjee, A. and Rajeev, B. (1998). Testing for the proportionality of hazards in two samples against the increasing cumulative hazard ratio alternative. *Scand. J. Statist.*, 25, 637–647.

Stablein, D. M. and Koutrouvelis, I. A. (1985). A two-sample test sensitive to crossing hazards in uncensored and singly censored data. *Biometrics*, 41, 643–652.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika*, 73, 363–369.

Stute, W. (1997). Nonparametric model checks for regression. *Ann. Statist.*, 25, 613–641.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model.* Springer, New York.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics.* Springer, New York.

Wei, L. J. (1984). Testing goodness of fit for proportional hazards model with censored observations. *J. Amer. Statist. Assoc.*, 79, 649–652.

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

Yang, S. and Prentice, R. (2005). Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. *Biometrika*, 92, 1–17.