

Goodness-of-fit testing in survival models

M Cockeran



orcid.org / 0000-0002-3990-8345

Thesis accepted for the degree *Doctor of Philosophy in Statistics* at the North-West University

Promoter:	Prof JS Allison
Co-promoter:	Prof L Santana
Assistant Promoter:	Prof SG Meintanis

Graduation May 2020
21102007

Bedankings

Aan my Hemelse Vader vir sy liefde, krag en genade.

Aan die volgende persone, my opregte dank:

- My studieleiers, Proff. J.S. Allison, S.G. Meintanis en L. Santana, vir u bekwame en professionele leiding.
- My man, Henry, vir jou oneindige liefde, belangstelling, motivering en geduld.
- My ouers, vir julle deurlopende opofferings, ondersteuning en bemoediging deur al my studiejare.

Abstract

The statistical analysis of lifetime data is an important topic in many areas, such as the biomedical and engineering sciences. The main objective in many of these studies is to examine the relationship between lifetimes and their associated covariates by means of a survival model. In these models it is sometimes appropriate to make specific distributional assumptions about the shape of the survival function. Therefore, one of the main objectives of this thesis is to evaluate a variety of tests for testing the goodness-of-fit (GOF) of parametric survival models (in particular the proportional hazards and accelerated failure time models) and compare their performance to one another. The approach adopted in this thesis, to conduct GOF, involves exploiting the property that, if the survival model is correctly specified, the errors follow a unit exponential distribution.

Aspects of the models under consideration include the shape of the survival function and time-invariance of the covariates. In addition, interest lies in testing GOF of the proposed model against a more general model. These aspects are investigated by first evaluating the GOF of survival models in the complete sample case by considering various tests for exponentiality. The results of an extensive Monte Carlo study show that tests based on the characteristic function and those based on the Laplace transform have the best overall power performance. However, in lifetime studies, it is not always possible to obtain complete information on lifetimes for all the subjects or units under study. This kind of data is called *censored* and a special case of this is Type-II right censored data which commonly arises in the testing of, for example, the lifetimes of equipment. The GOF testing procedures considered for Type-II right censored data can be divided into two categories: (i) an approach involving transforming the data to a complete sample, and then, on the newly constructed complete sample, employ any of the numerous tests for exponentiality designed for complete samples, and (ii) an approach where the test statistic is modified to accommodate Type-II right censoring. It was found, after performing a thorough Monte Carlo study, that both of these two approaches are useful for conducting GOF testing for exponentiality if the censoring proportion in a data set is lower than 30%. However, for higher censoring proportions, we would recommend using the approach that first transforms the sample to a complete sample. Additionally, it was found that tests based on the characteristic function or the Laplace transform generally outperformed the other tests for many of the settings considered. Lastly, we present a new way of modifying three well-known tests for exponentiality based on either the characteristic function or the Laplace transform to accommodate

Type-II right censored data. In a limited Monte Carlo study we found that these newly modified test statistics outperformed both the existing modified tests as well as the tests based on transformed data.

Key words:

bootstrap, exponentiality, goodness-of-fit, type-II right censoring, survival analysis.

Preface

Chapter 1 contains a general overview, as well as the objectives of the study. This thesis is written in article format and consists of one published article and two submitted articles. The three articles can be found in Chapters 2, 3 and 4, entailing abstracts, introductions, methods, results, conclusions and appropriate referencing formats according to the specific peer-reviewed journals' guidelines. Chapter 5 consists of the main findings and conclusions, as well as recommendations for future research.

The first article, namely: *Goodness-of-fit tests in the Cox proportional hazards model*, has been published in the journal *Communications in Statistics - Simulation and Computation*.

The second article, namely: *Goodness-of-fit testing of survival models in the presence of Type-II right censoring*, has been submitted to the journal *Lifetime Data Analysis*.

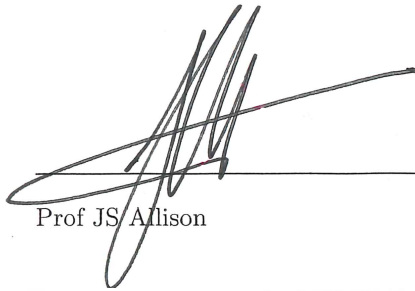
The third article, namely: *Modified tests for exponentiality in the presence of Type-II right censored data*, has been submitted to the journal *Stat*.

The promoters agreed on co-authorship and gave consent for the use of these articles as part of the final thesis. The first author was solely responsible for the initial planning and proposal of the thesis, literature searches, writing all programs used in the Monte Carlo studies, interpretation of results, as well as planning and writing of the articles and the entire thesis.

STATEMENT BY CO-AUTHORS

Herewith is a statement of the co-authors giving permission that the articles may form part of the thesis.

I hereby approve the articles and give my consent that these articles may be published as part of the thesis for the degree Doctor of Philosophy in Statistics of Mrs M Cockeran.



Prof JS Allison

18 November 2019

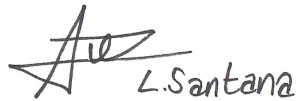
Date



Prof SG Meintanis

18 November 2019

Date



Prof L Santana

18 November 2019

Date

Contents

Bedankings	i
Abstract	ii
Preface	iv
1 Introduction	1
1.1 Overview	1
1.2 Objectives	2
1.3 Thesis outline	3
2 Article 1: Goodness-of-fit tests in the Cox proportional hazards model	5
Addendum to Article 1: Consistency results	18
3 Article 2: Goodness-of-fit testing in survival models in the presence of Type-II right censoring	23
4 Article 3: Modified tests for exponentiality in the presence of Type-II right censored data	58
5 Conclusion	68
5.1 Introduction	68
5.2 Results	69
5.3 Concluding remarks and future research	71

Chapter 1

Introduction

1.1 Overview

The statistical analysis variously referred to as lifetime, survival time, or failure time data is an important topic in many areas, including the biomedical, engineering, and social sciences. Applications of lifetime distribution methodology range from investigations of the durability of manufactured items to studies of human diseases and their treatment. In these fields of study it is not always possible to obtain complete information on survival or failure times for all the subjects or units under study. This kind of data is called *censored* and a special case of this is Type-II right censored data which commonly arises in the testing of equipment lifetimes. In this type of study all units are put on test at the same time, and the study is terminated when r of the n units have failed. Now, the main objective in many survival and reliability studies is to examine the effect of a set of explanatory variables on these time-to-event outcomes. Survival models, which are commonly used to model these relationships between time-to-event outcomes and explanatory variables can be formulated in many ways, specifically there are two approaches that have become popular in the statistical literature.

These are the accelerated failure time (AFT) model in which the survival function is modeled by means of covariates, and the Cox proportional hazards (PH) model in which one models the cumulative hazard function, where it is occasionally appropriate to make specific distributional assumptions about the shape of the survival function. If these parametric models provide a good fit to the data, the models tend to give more precise estimates of the quantities of interest, since these estimates are based on fewer parameters. Of course, if the model is wrong, this can lead to incorrect conclusions. Therefore testing goodness-of-fit (GOF) of parametric survival models is of great importance. However, deviations from the proposed shape of the survival function is not the only violation that needs to be investigated, but also time-dependency of the covariates as well as testing GOF of the proposed model against a more general model.

There exists several methods of testing GOF of PH and AFT models. Such methods include graphical

analyses of residuals which are, however, based on subjective assessment and should therefore be employed only at a preliminary data analysis level. On the other hand, one may opt for classical chi-squared tests, but this approach is extremely sensitive to the type of data-grouping, which again is chosen by the practitioner. Due to the limitations of the above mentioned methods one would prefer to determine model adequacy based on formal hypothesis testing methods. The approach adopted in this thesis, to conduct GOF using statistical inference, involves using the fact that if the model is correctly specified the distribution of the errors should be unit exponential. The GOF then reduces to a simple comparison of the properly estimated model-based residuals with that of a unit exponential variate. Furthermore, in the case of Type-II right censored data there are two approaches to this GOF testing based on the exponential distribution. The first approach is to employ test statistics for exponentiality which were developed for complete samples and modify them in order to make them applicable for Type-II right censored data. The second approach is to first transform the censored sample to a complete sample and then, on the newly constructed complete sample, employ any of the numerous tests for exponentiality designed for complete samples. The methods employed here are omnibus, meaning that they are able to detect arbitrary deviations from the null hypothesis provided that the sample size is sufficiently large.

1.2 Objectives

The overall aim of this thesis is to conduct Monte Carlo simulation studies to evaluate a variety of tests for testing the GOF of parametric survival models (in particular the PH and AFT models) and compare their performance to one another. Aspects of the models under consideration include the shape of the survival function and time-invariance of the covariates. In addition, interest lies in testing GOF of the proposed model against a more general model. These aspects are investigated by first evaluating the GOF of survival models in the complete sample case by considering various tests for exponentiality and then considering Type-II right censored data. The main objectives of the thesis can be summarized as follows:

- Review the existing literature on goodness-of-fit testing in survival models.
- Develop new parametric bootstrap algorithms for approximating the null distribution of the various test statistics used to test the adequacy of the survival models.
- Evaluate the performance of various GOF tests in order to determine the adequacy of the PH model with respect to the three scenarios stated above in the complete samples case.
- Derive some theory relating to the consistency of one of the test statistics under consideration.
- Compare two different strategies that are often employed to conduct GOF testing in the presence of Type-II right censoring.

- Evaluate the performance of various GOF tests in order to determine the adequacy of PH and AFT models in the presence of Type-II right censoring.
- Suitably modify test statistics based on either the characteristic function or the Laplace transform to accommodate Type-II right censored data.
- Evaluate the effectiveness of the newly modified test statistics against existing modified tests.
- Make recommendations of which tests to use under which circumstances when testing the adequacy of a specific survival model.

1.3 Thesis outline

In this section we present the general themes, titles, abstracts and connections between the three articles that comprise this thesis. The three articles presented all contributed to the literature on GOF of survival models, with specific emphasis on PH models. The first article (Chapter 2) starts by examining the GOF of the PH model in the complete sample case by considering various tests for exponentiality. We also include some proofs pertaining to the consistency of one of the Laplace transform based tests in the parametric Cox proportional hazards setup. Article two (Chapter 3) extends the study done in article one to testing the fit of both PH and AFT models in the presence of Type-II right censoring. Here, both transforming the sample to an approximate independent identically distributed (i.i.d) complete sample and existing modified test statistics are used. The third and final article (Chapter 4) presents a new way of modifying three well-known tests for exponentiality based on either the characteristic function or the Laplace transform to accommodate Type-II right censored data. Lastly, in Chapter 5 we provide some overall conclusions, recommendations and suggestions on possible future research in the area of goodness-of-fit testing in the model-based setup in the presence of random censoring.

The titles and abstracts of the three articles, presented in Chapters 2, 3 and 4 respectively, are given below:

Goodness-of-fit tests in the Cox proportional hazards model

We consider a variety of tests for testing goodness-of-fit in a parametric Cox proportional hazards (PH) model and compare their performance. Aspects of the model under test include the baseline distribution and time-invariance of covariates. We also test for the PH model itself against a certain generalization. This is done through an extensive Monte Carlo study where we simulate the performance of the tests for these three paired hypotheses. The results show that the tests based on the empirical characteristic function and those based on the empirical Laplace transform have the best overall power performance. It is also found that the distributions of the considered test statistics do not depend on the specific functional form of the covariate function.

Goodness-of-fit testing of survival models in the presence of Type-II right censoring

We consider a variety of tests for testing goodness-of-fit in a parametric Cox proportional hazards (PH) model and accelerated failure time (AFT) model in the presence of Type-II right censoring. The testing procedures considered can be divided into two categories: an approach involving transforming the data to a complete sample, and an approach where the test statistic is modified to accommodate Type-II right censoring. The power of the proposed tests are compared through a Monte Carlo study.

Modified tests for exponentiality in the presence of Type-II right censored data

The exponential distribution is one of the most popular choices of model, especially in survival analysis and reliability theory. Typically, to test the hypothesis that the underlying distribution of lifetimes are exponentially distributed based on censored data, the standard goodness-of-fit tests used in the complete sample case are modified. However, none of the characteristic or Laplace function based tests have been adapted to incorporate censored data. In this paper we modify three well-known tests for exponentiality based on either the characteristic function or Laplace transform to accommodate Type-II right censored data. We compare the power of the three newly changed tests to some existing modified tests in an i.i.d. and model-based setup through a Monte Carlo study, where it was found that the newly adjusted Baringhaus and Henze test outperformed all other tests considered.

Chapter 2

Article 1: Goodness-of-fit tests in the Cox proportional hazards model

The first article, namely: *Goodness-of-fit tests in the Cox proportional hazards model*, has been published in the journal *Communications in Statistics - Simulation and Computation*. A summary of the guidelines to authors from the journal is now presented.

Manuscript	A maximum of 25 pages inclusive of the abstract, tables, references, figure captions.
Title	A full title and a short title of up to 50 characters should be provided.
Abstract and keywords	A maximum of 200 words succinctly describing the article and between 3–5 keywords.
Tables	Tables should present new information rather than duplicating what is in the text. Readers should be able to interpret the table without reference to the text.
References	References should be prepared according to the Taylor & Francis US Chicago (B) author-data style.
General formatting	A LaTeX template is provided for submission.
Additional information	https://tinyurl.com/y37atv3z



Goodness-of-fit tests in the Cox proportional hazards model

Marika Cockeran^a , Simos George Meintanis^{a,b}, and James S. Allison^a

^aUnit for Business Mathematics and Informatics, North-West University, Potchefstroom, South Africa;

^bDepartment of Economics, National and Kapodistrian University of Athens, Athens, Greece*

ABSTRACT

We consider a variety of tests for testing goodness-of-fit in a parametric Cox proportional hazards (PH) model and compare their performance. Aspects of the model under test include the baseline distribution and time-invariance of covariates. We also test for the PH model itself against a certain generalization. This is done through an extensive Monte Carlo study where we simulate the performance of the tests for these three paired hypotheses. The results show that the tests based on the empirical characteristic function and those based on the empirical Laplace transform have the best overall power performance. It is also found that the distributions of the considered test statistics do not depend on the specific functional form of the covariate function.

ARTICLE HISTORY

Received 17 December 2018

Accepted 28 June 2019

KEYWORDS

Test for exponentiality;
Parametric bootstrap;
Survival model

1. Introduction

The main objective in many clinical studies is to examine the effect of a set of explanatory variables on time-to-event outcomes. Regression models for time-to-event outcomes can be formulated in many ways. According to Klein and Moeschberger (2006) two approaches to the modeling of covariate effects on time-to-event outcomes have become popular in the statistical literature. These are the accelerated failure time (AFT) model in which the survival function is modeled by means of covariates and the Cox proportional hazards (PH) model in which we model the cumulative hazard function.

In some cases it is appropriate to make specific distributional assumptions about the shape of the survival function. If these parametric models provide a good fit to the data, the models tend to give more precise estimates of the quantities of interest since these estimates are based on fewer parameters. Of course if the model is wrong, this leads to incorrect conclusions. Therefore testing goodness-of-fit of parametric regression models is of great importance. There are several methods for testing parametric regression models including informal model-checking procedures such as probability and residual plots. However, the interpretation of these plots can be subjective. In this study the determination of model adequacy is based on formal hypothesis testing.

CONTACT Simos George Meintanis  simosmei@econ.uoa.gr  Unit for Business Mathematics and Informatics, North-West University, Potchefstroom, South Africa.

*On sabbatical leave from the University of Athens.

© 2019 Taylor & Francis Group, LLC

In this connection we note that Balakrishnan et al. (2013) investigated competing hypotheses concerning the parameterization of the baseline distribution and considered the classical goodness-of-fit statistics based on the empirical distribution function (EDF), i.e. the Kolmogorov-Smirnov, Cramér-von Mises and Anderson-Darling test statistics. Here we investigate competing hypotheses concerning the parameterization of the baseline distribution, but also time-dependency of the covariates as well as testing goodness-of-fit of the PH model against a more general model. Moreover, we also include, apart from the classical tests based on the EDF, several recent goodness-of-fit tests based on the empirical Laplace transform and on the empirical characteristic function. Although in this paper we present results just for the proportional hazards model, we have also run an analogous study for the AFT model. These results were, however, quite similar and therefore they are not reported here.

To fix notation, let T be a non-negative random variable representing the time until some specified event with probability density function $f(t)$ and cumulative distribution function $F(t)$. We model the cumulative hazard rate of the random variable T by means of the proportional hazards (PH) model

$$\Lambda(t) = r(\mathbf{x})H(t) \quad (1)$$

where $r(\cdot)$ is a regression function that depends on the covariate $\mathbf{x} = (x_1, x_2, \dots, x_m)$ and $H(\cdot)$ is some unspecified baseline cumulative hazard rate function. On the basis of independent observations $(T_j, \mathbf{x}_j), j = 1, \dots, n$, we wish to test the null hypothesis

$$\mathcal{H}_0 : H(t) = H_0(t; \boldsymbol{\theta}), \text{ for some } \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k \quad (2)$$

where $H_0(\cdot; \boldsymbol{\theta})$ is a parametric baseline cumulative hazard rate function depending on some unknown parameter $\boldsymbol{\theta}$. We assume that $H_0(\cdot; \boldsymbol{\theta})$ is strictly increasing.

Although it will be seen in the ensuing study that the specific functional form of the regression function $r(\cdot)$ does not affect the performance of the considered tests, for simplicity, we will adopt a specified regression function $r(\mathbf{x}) = r(\mathbf{x}; \beta)$ that depends on the parameter $\beta \in \mathcal{B} \subset \mathbb{R}^m$. Under this presumption the cumulative hazard function of the j^{th} individual can be described by the PH model as

$$\Lambda_j(t; \psi) = r(\mathbf{x}_j; \beta)H_0(t; \boldsymbol{\theta}), t \geq 0 \quad (3)$$

where $\psi = (\beta, \boldsymbol{\theta})$. Now let $\hat{\psi}_n = (\hat{\beta}_n, \hat{\boldsymbol{\theta}}_n)$ be some consistent estimator of $\psi = (\beta, \boldsymbol{\theta})$, (throughout the paper we make use of maximum likelihood estimation). Then under the null hypothesis, the (so-called) Cox-Snell residuals from the PH model defined in Eq. (1) should (approximately) follow a unit exponential distribution. These residuals are defined as

$$\hat{\varepsilon}_j = r(\mathbf{x}_j; \hat{\beta}_n)H_0(T_j; \hat{\boldsymbol{\theta}}_n) \quad (4)$$

where $T_j, j = 1, 2, \dots, n$, are the observed survival times. Hence any exponentiality test on the basis of $\hat{\varepsilon}_j, j = 1, \dots, n$, constitutes in effect a test for the PH model itself. Important as it is, the type of baseline hazard is only one of the elements of survival-modeling. For instance such models are often assumed to involve time-dependent covariates, i.e. $\mathbf{x} := \mathbf{x}(t)$. This is a very important assumption and therefore we plan to also address this aspect of the PH model in our Monte Carlo study.

Another possible model violation relates to more general configurations of the PH model implied by Eq. (3) such as the Hsieh model; see Hsieh and Lavori (2000). The Hsieh model is a generalization of the proportional hazards model and is given by

$$\Lambda_j(t; \psi, \gamma) = r(\mathbf{x}_j; \beta) H_0(t; \theta) e^{\gamma' \mathbf{x}_j}$$

where $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_m) \in \mathbb{R}^m$ is a set of unknown positive parameters. Clearly if $\gamma = 0$ then the Hsieh model reduces to the PH model.

The rest of the article is organized as follows: In Sec. 2 we provide details of the various tests for exponentiality included in the simulation study. In Sec. 3 a parametric resampling procedure for the actual implementation of the tests is outlined. In Sec. 4 we present the Monte Carlo design on the basis of which we subsequently simulate the performance of the tests considered for different pairs of competing hypotheses. The paper concludes with a discussion in Sec. 5. In this paper we report only complete-data scenarios. The case of censoring requires a separate study. We nevertheless include a supplement with a small simulation study for test-performance under type-II right censored data.

2. Tests for exponentiality

To test the hypothesis that the observed residuals $\hat{e}_1, \hat{e}_2, \dots, \hat{e}_n$ are realizations from a standard exponential distribution, we use a number of goodness-of-fit tests for exponentiality. The approaches used to construct these tests can broadly be categorized into two groups: Classical tests based on the EDF and new tests based on empirical transforms. We also consider a recent moment-based test for the exponential distribution.

2.1. Tests based on the EDF

The EDF-based tests are expressed in terms of distances between the EDF

$$G_n(t) = \frac{1}{n} \sum_{j=1}^n I(\hat{e}_j \leq t)$$

of the residuals $\hat{e}_1, \hat{e}_2, \dots, \hat{e}_n$ defined in Eq. (4) and the corresponding population quantity for the standard exponential distribution $G(t) = 1 - e^{-t}$. The following EDF-based test statistics are considered: The Kolmogorov-Smirnov (KS) test statistic

$$\widetilde{KS}_n = \sup_{t \geq 0} |G_n(t) - G(t)|$$

the Cramér-von Mises (CM) test statistic

$$CM_n = \int_0^\infty [G_n(t) - G(t)]^2 dG(t)$$

and the Anderson-Darling test (AD) statistic

$$AD_n = \int_0^\infty \frac{[G_n(t) - G(t)]^2}{G(t)[1 - G(t)]} dG(t).$$

There are computationally efficient formulae for the EDF-based statistics that can be found in D'Agostino and Stephens (1986). Specifically, the KS, the CM and the AD test statistics simplify to

$$\widetilde{KS}_n = \max\{KS_n^+, KS_n^-\}$$

where

$$\begin{aligned} KS_n^+ &= \max_{1 \leq j \leq n} \left[\frac{j}{n} - (1 - e^{-\hat{\varepsilon}_{(j)}}) \right] \\ KS_n^- &= \max_{1 \leq j \leq n} \left[(1 - e^{-\hat{\varepsilon}_{(j)}}) - \frac{j-1}{n} \right] \\ CM_n &= \frac{1}{12n} + \sum_{j=1}^n \left[(1 - e^{-\hat{\varepsilon}_{(j)}}) - \frac{2j-1}{2n} \right]^2 \end{aligned}$$

and

$$AD_n = -n - \sum_{j=1}^n \frac{2j-1}{n} \left[\ln(1 - e^{-\hat{\varepsilon}_{(j)}}) - \hat{\varepsilon}_{(n+1-j)} \right]$$

respectively, and where $\hat{\varepsilon}_{(1)} \leq \dots \leq \hat{\varepsilon}_{(n)}$ denotes the ordered residuals. For the KS test statistic we use the Bolshev correction that rejects the null hypothesis for large values of

$$KS_n = \frac{6n \cdot \widetilde{KS}_n + 1}{6\sqrt{n}}.$$

2.2. Tests based on transforms

The tests based on empirical transforms that we have in mind are tests based on the empirical characteristic function (ECF)

$$\phi_n(t) = \frac{1}{n} \sum_{j=1}^n e^{it\hat{\varepsilon}_j}$$

tests based on the empirical Laplace transform (ELT)

$$\psi_n(t) = \frac{1}{n} \sum_{j=1}^n e^{-t\hat{\varepsilon}_j}$$

as well as tests based on the so-called probability weighted empirical characteristic function (PWEFCF) defined as

$$\chi_n(t; a) = \frac{1}{n} \sum_{j=1}^n W(\hat{\varepsilon}_j; at) e^{it\hat{\varepsilon}_j}, t \in \mathbb{R}$$

where $W(\hat{\varepsilon}_j; \beta) = [G(\hat{\varepsilon}_j)(1 - G(\hat{\varepsilon}_j))]^{|\beta|}$. The corresponding population quantities for the standard exponential distribution are $\phi(t) = (1 - it)^{-1}$, $\psi(t) = (1 + t)^{-1}$, and

$$\chi(t; a) = \Gamma(1 + a|t|) \frac{\Gamma(1 + a|t| - it)}{\Gamma(2 + 2a|t| - it)}$$

respectively.

The transform-based criteria considered here are the test proposed in Epps and Pulley (1986)

$$EP_n = \int_{-\infty}^{\infty} [\phi_n(t) - \phi(t)] \frac{1}{2\pi} \phi(-t) dt$$

based on the ECF, the test of Baringhaus and Henze (1991)

$$BH_{n,a} = n \int_0^{\infty} [(1+t)\psi'_n(t) + \psi(t)]^2 \exp(-at) dt$$

and the test of Henze and Meintanis (2005)

$$HM_{n,a} = n \int_0^{\infty} [\psi_n(t) - \psi(t)]^2 (1+t)^2 e^{-at} dt$$

both based on the ELT, as well as the test of Meintanis, Swanepoel, and Allison (2014)

$$PW_{n,a} = n \int_{-\infty}^{\infty} |\chi_n(t; a) - \chi(t; a)|^2 dt$$

based on the PWECF, where $a > 0$ is a constant (tuning) parameter.

Convenient computational formulae for these test statistics are

$$\begin{aligned} EP_n &= \sqrt{48n} \left[\frac{1}{n} \sum_{j=1}^n e^{-\hat{e}_j} - \frac{1}{2} \right] \\ BH_{n,a} &= \frac{1}{n} \sum_{j=1}^n \sum_{k=1}^n \left(\frac{(1-\hat{e}_j)(1-\hat{e}_k)}{\hat{e}_j + \hat{e}_k + a} - \frac{\hat{e}_j + \hat{e}_k}{(\hat{e}_j + \hat{e}_k + a)^2} \right. \\ &\quad \left. + \frac{2\hat{e}_j\hat{e}_k}{(\hat{e}_j + \hat{e}_k + a)^2} + \frac{2\hat{e}_j\hat{e}_k}{(\hat{e}_j + \hat{e}_k + a)^3} \right) \\ HM_{n,a} &= \frac{1}{n} \sum_{j=1}^n \sum_{k=1}^n \left[\frac{1 + (\hat{e}_j + \hat{e}_k + a + 1)^2}{(\hat{e}_j + \hat{e}_k + a)^3} \right] - 2 \sum_{j=1}^n \left[\frac{1 + \hat{e}_j + a}{(\hat{e}_j + a)^2} \right] + \frac{n}{a} \end{aligned}$$

and

$$\begin{aligned} PW_{n,a} &= -\frac{2}{n^2} \sum_{j=1}^n \sum_{k=1}^n \frac{a \ln [(1 - Z_j)Z_j(1 - Z_k)Z_k]}{(\hat{e}_j - \hat{e}_k)^2 + a^2 \ln^2 [(1 - Z_j)Z_j(1 - Z_k)Z_k]} \\ &\quad + \frac{2}{n} \sum_{j=1}^n \int_0^1 \frac{a \ln [(1 - Z_j)Z_j(1 - u)u]}{[\hat{e}_j + \ln(1 - u)]^2 + a^2 \ln^2 [(1 - Z_j)Z_j(1 - u)u]} du \end{aligned}$$

where $Z_j = 1 - e^{-\hat{e}_j}$, $j = 1, \dots, n$. The computational form of $PW_{n,a}$ is from Meintanis, Swanepoel, and Allison (2014), where the ‘constant term’ is dropped. In what follows we often write EP, BH, HM and PW for these criteria.

2.3. A moment-based test

Clearly transforms such as the characteristic function and the Laplace transform are moment-related quantities and so are the corresponding test statistics involving the ECF or the ELT. Another exponentiality test based on moments was first put forward by Lee, Locke, and Spurrier (1980). This test utilizes the fact that if X is exponentially distributed then

$$\frac{E(X^a)}{[E(X)]^a} = \Gamma(1 + a), \text{ for } a > 0.$$

The test was further modified by Mimoto and Zitikis (2008) by means of the relation

$$\frac{E[X^a]^{\frac{1}{a}}}{E(X)} = \Gamma(1 + a)^{\frac{1}{a}}, \text{ for } -1 < a < 1, a \neq 0.$$

The resulting test criterion yields the so-called Atkinson (AT) test based on

$$AT_{n,a} = \sqrt{n} |R_n(a) - \Gamma(1 + a)^{\frac{1}{a}}|$$

where

$$R_{n,a} = \left[\frac{1}{n} \sum_{j=1}^n \hat{e}_j^a \right]^{\frac{1}{a}}.$$

Extensive Monte Carlo power studies were done in Mimoto and Zitikis (2008) and they found that a value of a close to 0 or 1 produced the highest power for the majority of alternatives considered.

3. Parametric resampling procedure

Typically the asymptotic distributions of complicated statistics, even when available, are most often highly non-standard, making it extremely difficult to apply and actually carry out in practice. One main reason is that the null hypothesis (see for instance $\mathcal{H}_0$ in Eq. (2)) often involves unknown parameters, and therefore the tests statistics should be implemented in a way that takes into account not only the model under test, but also the variability of the corresponding parameter estimates. In such situations we appeal to resampling techniques in order to calculate critical values. These techniques are typically conditioned on the residuals $\{\hat{e}_j\}_{j=1}^n$ defined in Eq. (4) and on the resulting test statistic, say $S = S(\hat{e}_1, \hat{e}_2, \dots, \hat{e}_n)$. The most often-used resampling method is the parametric bootstrap procedure that runs as follows:

1. Generate a bootstrap sample of survival times $T_1^*, T_2^*, \dots, T_n^*$, where

$$T_j^* = H_0^{-1} \left[-\log(U_j) \frac{1}{r(\mathbf{x}_j; \hat{\beta}_n)}; \hat{\theta}_n \right]$$

2. Using the bootstrap survival times, $T_1^*, T_2^*, \dots, T_n^*$, fit a PH model defined under the null hypothesis $\mathcal{H}_0$.

3. Calculate the Cox-Snell residuals for the PH model fitted in step 2,

$$\hat{e}_j^* = r(\mathbf{x}_j, \hat{\beta}_n^*) H_0(T_j^*, \hat{\theta}_n^*)$$

where $\hat{\psi}_n^* = (\hat{\beta}_n^*, \hat{\theta}_n^*)$ are the estimators of the model parameters based on the bootstrap sample.

4. Based on $\hat{e}_1^*, \hat{e}_2^*, \dots, \hat{e}_n^*$ compute the test statistic $S^* = S(\hat{e}_1^*, \hat{e}_2^*, \dots, \hat{e}_n^*)$.
5. Repeat steps 1 to 4 a number of times say B , and obtain $\{S_m^*\}_{m=1}^B$.
6. The null hypothesis is rejected at a prescribed significance level α if $S > S_{(B-\alpha B)}^*$, where $S_{(1)}^* \leq \dots \leq S_{(B)}^*$ denote the order statistics.

Given that the above algorithm might be computationally demanding, we employ the warp-speed method proposed by Giacomini, Politis, and White (2013) in order to calculate critical values of all the test statistics considered in the Monte Carlo study. This method circumvents the need for time-consuming bootstrap replications for each Monte Carlo sample. With the warp-speed method, only one resample is generated for each Monte Carlo sample and we compute the bootstrap test statistic S^* for that resample. If we then denote the number of Monte Carlo replications by MC , the size- α critical value is determined similarly as in step 6, by computing the $(1-\alpha)^{\text{th}}$ quantile of the bootstrap statistics $\{S_m^*\}_{m=1}^{MC}$.

4. Simulation design and results

In this section we compare the finite-sample performance of all the tests for exponentiality discussed in Sec. 2. We evaluate the empirical powers of the tests for the following pairs of competing hypotheses for complete samples (henceforth referred to as Scenarios A, B and C, respectively):

- A. The null hypothesis corresponds to the PH model with a certain baseline distribution against an alternative with the same model but different baseline distribution;
- B. The null hypothesis corresponds to the PH model with a time-independent covariate against the same PH model, i.e. with same baseline distribution, but with time-varying covariate;
- C. The null hypothesis corresponds to the PH model and the alternative is a Hsieh model (with same baseline distribution).

4.1. Simulation setting

All size and power estimates were calculated for 10 000 independent Monte Carlo samples for sample sizes $n=100$ and $n=200$. Tests in which a tuning parameter, a , appears, were evaluated for different values of a . The distributions used for the baseline cumulative hazard function as well as their parameterizations are given in Table 1. These distributions include models with a constant hazard rate, decreasing hazard rate, increasing hazard rate and non-monotone hazard rate. The unknown parameters are estimated by maximum likelihood. Throughout the study the covariate function takes the form $r(\mathbf{x}_j; \beta) = e^{\beta x_j}$ with $\beta = 0.1$. In case of the PH model with a time-independent

Table 1. Densities of baseline distributions.

Distribution	Density function	Parameter values	Type of hazard
Exponential	$f(t) = \lambda e^{-\lambda t}$	$\lambda = 0.05$	Constant
Weibull	$f(t) = \frac{a}{b} \left(\frac{t}{b}\right)^{a-1} e^{-\left(\frac{t}{b}\right)^a}$	$a = 1.1, b = 0.8$	Increasing
Lognormal	$f(t) = \frac{1}{\sqrt{2\pi\sigma t}} e^{-\frac{(\log t - \mu)^2}{2\sigma^2}}$	$\mu = 1.3, \sigma = 1.15$	Non-Monotone
Gompertz	$f(t) = be^{at} e^{-\frac{b}{a(e^{at}-1)}}$	$a = 1.1, b = 2.0$	Increasing
Gamma	$f(t) = \frac{1}{s^a \Gamma(a)} x^{a-1} e^{-\frac{t}{s}}$	$a = 0.5, s = 1.1$	Decreasing

covariate, a sample of observations was generated with a single covariate, which can take on one of four values: 0, 1, 2, or 3. Equal numbers of observations are generated corresponding to the different values of the covariate. Under the alternative a single time-varying covariate $x(t)$ is used. It is a dichotomous time-varying covariate with at most one change from untreated to treated. Let t_0 denote the time at which the time-varying covariate changes from the value c_1 to the value c_2 . The values used are $c_1 = 0, c_2 = 5$, and $t_0 = 0.1$. For the Hsieh model $\gamma \in \{0.25, 0.5\}$ is considered. All calculations are done in R (The R Development Core Team 2016).

4.2. Simulation results

Tables 2–8 show the estimated powers of the various tests discussed in Sec. 2 for the various pairs of competing hypotheses discussed at the beginning of this section. Tables 2–6 show the estimated powers for the hypothesis stated in scenario (A). Similarly, Tables 7 and 8 show the estimated powers for the hypotheses in scenario (B) and (C), respectively, for different choices of the baseline distribution. The entries in these tables are the percentage of 10 000 independent Monte Carlo samples that resulted in the rejection of the null hypothesis rounded to the nearest integer. For ease of comparison, the highest power for each alternative has been highlighted.

We will now discuss the performance of the tests for each scenario separately and then conclude with some overall findings.

4.2.1. Scenario A

Overall, AT produced some of the highest powers followed closely by BH and HM. For the specific case where the baseline cumulative hazard H_0 is gamma AT outperforms the majority of the other tests and, to a lesser degree, when H_0 is lognormal. Note, however, AT displays rather poor performance when a Gompertz baseline hazard is used. For this reason we also investigated the performance of AT when $a = 0.1$ and found that the test produced some power, but still underperformed when H_0 is Gompertz. When H_0 is Weibull, the three traditional tests (AD, CM, KS) do not compare favorably to the more modern tests such as EP, BH, HM and AT. PW appears to consistently have the lowest powers, appearing to only display reasonable powers for the lognormal alternative.

Table 2. Estimated powers for scenario A where H_0 is exponential.

$\mathcal{H}_a$	n	AD	CM	KS	EP	BH _{0.1}	BH _{2.5}	BH _{5.0}	HM _{0.1}	HM _{2.5}	HM _{5.0}	PW _{0.025}	PW _{0.1}	PW _{0.5}	AT _{0.99}
Exponential	100	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Gamma		100	100	99	100	100	100	99	100	100	99	100	100	100	100
Gompertz		57	60	49	69	39	69	71	17	66	70	6	17	44	68
Lognormal		61	58	50	46	53	51	59	81	44	61	84	78	26	54
Weibull		20	20	17	22	19	22	20	14	22	21	7	13	27	0
Exponential	200	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Gamma		100	100	100	100	100	100	100	100	100	100	87	0	0	100
Gompertz		90	91	81	95	72	95	97	32	94	96	4	22	68	96
Lognormal		95	90	82	71	94	78	84	79	70	82	99	98	48	79
Weibull		35	35	28	39	35	39	36	24	40	37	7	20	41	0

Table 3. Estimated powers for scenario A where H_0 is gamma.

$\mathcal{H}_a$	n	AD	CM	KS	EP	BH _{0.1}	BH _{2.5}	BH _{5.0}	HM _{0.1}	HM _{2.5}	HM _{5.0}	PW _{0.025}	PW _{0.1}	PW _{0.5}	AT _{0.99}
Exponential	100	4	4	4	2	4	2	1	4	2	1	4	4	4	6
Gamma		5	5	5	5	5	5	5	5	5	5	5	5	5	5
Gompertz		27	23	19	24	25	21	13	24	21	15	0	0	3	45
Lognormal		84	79	68	86	88	85	81	88	85	82	83	79	58	93
Weibull		5	4	4	3	5	2	1	5	3	1	3	3	4	8
Exponential	200	4	4	4	3	4	2	1	5	2	1	5	4	5	7
Gamma		5	5	5	5	5	5	5	5	5	5	5	5	5	5
Gompertz		51	45	35	49	47	46	37	45	45	39	2	23	41	76
Lognormal		99	98	94	100	100	99	98	13	99	99	99	97	82	100
Weibull		5	5	5	3	5	2	1	6	3	2	2	2	4	9

4.2.2. Scenario B

In this scenario PW produced the highest estimated powers for almost all alternatives considered. Although HM, EP and AT do not produce the highest powers, they can still be regarded as competitive alternatives to PW. It seems that when the baseline distribution is lognormal, then all the tests have some difficulty distinguishing between time-varying and time-independent covariates.

4.2.3. Scenario C

PW outperforms the majority of the tests, followed by HM when $\gamma = 0.25$, and by AD when $\gamma = 0.5$. BH is also a viable option, since its power performance is comparable to AD and PW for both choices of gamma. Again note the poor performance of AT for the Gompertz baseline hazard.

A final overview of our findings is that the AT test produced some of the highest powers uniformly over all testing scenarios, except for the cases when a Gompertz distribution is involved. In addition, and although the BH and HM tests did not achieve the highest powers, they are consistently good performers, typically producing estimated powers within the top three. In this connection we note that all these three methods involve a parameter which should be chosen by the practitioner. Now for the AT test we have followed the suggestion in Mimoto and Zitakis (2008) that recommend a value of a close to the boundaries of the interval $(0, 1)$. Here we only present results for $a = 0.99$ but additional simulations with a close to zero lead to pretty much the same conclusions.

Turning to the BH and HM tests and based on close inspection of our simulation results we make the following suggestions: For both tests under scenario A a value of a away from $a = 0$ should be used for alternatives with constant or increasing hazard rate.

Table 4. Estimated powers for scenario A where H_0 is Gompertz.

$\mathcal{H}_a$	n	AD	CM	KS	EP	BH _{0.1}	BH _{2.5}	BH _{5.0}	HM _{0.1}	HM _{2.5}	HM _{5.0}	PW _{0.025}	PW _{0.1}	PW _{0.5}	AT _{0.99}
Exponential	100	7	7	7	19	7	15	20	6	15	21	6	6	4	0
Gamma		100	99	99	100	100	100	100	100	100	100	99	99	95	0
Gompertz		5	5	5	5	5	5	5	5	5	5	5	5	5	5
Lognormal		58	46	41	31	69	34	42	90	33	83	64	61	54	3
Weibull		9	9	9	16	8	14	16	6	14	16	11	14	17	0
Exponential	200	8	8	9	19	7	15	21	6	15	22	6	6	7	0
Gamma		100	100	100	100	100	100	100	100	100	100	13	0	0	0
Gompertz		5	5	5	5	5	5	5	5	5	5	5	5	5	5
Lognormal		93	79	74	35	98	47	47	99	45	79	94	89	79	2
Weibull		15	14	12	23	15	21	23	12	22	24	15	20	26	0

Table 5. Estimated powers for scenario A where H_0 is lognormal.

$\mathcal{H}_a$	n	AD	CM	KS	EP	BH _{0.1}	BH _{2.5}	BH _{5.0}	HM _{0.1}	HM _{2.5}	HM _{5.0}	PW _{0.025}	PW _{0.1}	PW _{0.5}	AT _{0.99}
Exponential	100	87	81	68	95	83	95	95	67	94	94	6	20	58	95
Gamma		99	97	91	100	97	100	100	83	100	100	19	48	82	100
Gompertz		98	97	90	100	96	100	100	82	99	100	17	44	80	100
Lognormal		5	5	5	5	5	5	5	5	5	5	5	5	5	5
Weibull		86	80	68	94	83	94	94	67	94	93	6	22	58	95
Exponential	200	99	98	94	100	99	100	100	92	100	100	5	30	81	100
Gamma		100	100	100	100	100	100	100	98	100	100	28	70	97	71
Gompertz		100	100	100	100	100	100	100	97	100	100	22	64	96	100
Lognormal		5	5	5	5	5	5	5	5	5	5	5	5	5	5
Weibull		99	98	93	100	99	100	100	92	100	100	5	30	81	100

Table 6. Estimated powers for scenario A where H_0 is Weibull.

$\mathcal{H}_a$	n	AD	CM	KS	EP	BH _{0.1}	BH _{2.5}	BH _{5.0}	HM _{0.1}	HM _{2.5}	HM _{5.0}	PW _{0.025}	PW _{0.1}	PW _{0.5}	AT _{0.99}
Exponential	100	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Gamma		20	16	13	27	23	28	29	24	27	30	1	1	0	28
Gompertz		20	16	13	24	22	26	28	22	25	28	1	0	0	27
Lognormal		83	75	59	82	92	91	91	93	90	92	85	83	75	90
Weibull		5	5	5	5	5	5	5	5	5	5	5	5	5	5
Exponential	200	6	5	5	6	6	5	5	5	5	5	6	6	5	5
Gamma		35	28	22	44	41	49	51	42	44	50	0	0	0	51
Gompertz		36	29	22	40	40	48	51	39	41	49	0	0	0	49
Lognormal		99	98	90	98	100	100	100	100	100	100	99	99	94	100
Weibull		5	5	5	5	5	5	5	5	5	5	5	5	5	5

On the other hand a value of this parameter close to zero should be preferred for alternatives with non-monotone hazard rate. The same is true under decreasing hazard rate alternatives for the BH test, while the HM seems to have a more robust performance under such alternatives. Under scenario B, again small values are supported for the BH test and this choice persists under scenario C with non-monotone and decreasing hazard rate alternatives for this test, while under constant or increasing hazard rates a value of a away from $a = 0$ should be chosen. For the HM test under scenario B a small value of a is only supported under constant hazard rates while in the remaining cases this value should be chosen away from $a = 0$, but it is often true that in those cases the exact value of this parameter does not have a significant impact on the power of the test. Finally under scenario C, the HM test performs better for small a only under decreasing hazard rates, while a larger value should be used for all other cases, i.e. for increasing, constant or non-monotone hazards.

It becomes clear from the preceding discussion that a ‘good’ parameter-value for a (rendering the test with high power) depends on the underlying alternative, which

Table 7. Estimated powers for scenario B for different baseline distributions.

Baseline	n	AD	CM	KS	EP	BH _{0.1}	BH _{2.5}	BH _{5.0}	HM _{0.1}	HM _{2.5}	HM _{5.0}	PW _{0.025}	PW _{0.1}	PW _{0.5}	AT _{0.99}
Exponential	100	23	20	18	20	28	19	16	22	20	17	21	27	35	18
Gamma		20	20	19	17	19	15	13	16	17	14	5	8	15	16
Gompertz		23	24	23	13	17	17	13	10	17	12	32	38	35	0
Lognormal		6	6	6	8	6	8	7	6	8	7	3	4	6	8
Weibull		11	10	8	8	11	10	10	9	9	10	17	15	13	9
Exponential	200	46	39	39	31	53	32	26	42	36	28	34	47	57	30
Gamma		34	36	37	27	34	23	18	28	27	20	5	12	25	23
Gompertz		46	47	46	24	36	33	25	21	34	23	55	65	57	0
Lognormal		7	7	6	10	8	9	9	7	10	9	3	4	6	10
Weibull		17	15	11	13	19	15	15	14	12	15	29	24	18	14

Table 8. Estimated powers for scenario C, for sample size $n=100$, for different baseline distributions.

Baseline	γ	AD	CM	KS	EP	BH _{0.1}	BH _{2.5}	BH _{5.0}	HM _{0.1}	HM _{2.5}	HM _{5.0}	PW _{0.025}	PW _{0.1}	PW _{0.5}	AT _{0.99}
Exponential	0.25	98	98	95	98	96	98	96	75	98	97	63	89	99	97
Gamma		19	16	13	15	21	13	9	23	15	10	1	3	11	19
Gompertz		32	32	27	45	36	46	49	32	44	46	23	39	55	0
Lognormal		24	21	17	28	27	27	23	23	29	24	18	14	13	24
Weibull		11	10	8	13	10	13	10	9	14	10	3	7	16	10
Exponential	0.50	100	100	100	100	100	100	100	97	100	100	100	100	100	100
Gamma		82	77	64	54	76	48	28	71	55	33	14	39	73	50
Gompertz		80	81	72	81	49	86	88	18	82	83	54	79	89	0
Lognormal		87	83	71	75	83	77	66	75	79	64	21	52	79	57
Weibull		71	68	53	54	48	62	36	28	66	33	15	41	67	25

however is unknown. Nevertheless the suggestions made can be of help if there exists some prior information regarding possible departures from the null hypothesis. Apart from prior experience such information may also be drawn by performing certain preliminary statistical analysis on the data at hand. In addition, running the test for several values of a with, say, $a \in (0, 5)$ could prove reassuring.

We also conducted a simulation study using various forms of the regression function $r(\mathbf{x}_j; \beta)$ with different parameter values, but the results are not reported here since we found that the distribution of the considered goodness-of-fit statistics do not depend on the specific functional form of the regression function $r(\mathbf{x}_j; \beta)$ or on the parameters featuring in $r(\mathbf{x}_j; \beta)$. The same distribution however is strongly dependent on the chosen parameterization of the baseline cumulative hazard $H_0(t; \theta)$.

5. Conclusion

We compare the performance of a variety of tests for exponentiality when testing goodness-of-fit for the PH model. The testing procedures considered included tests based on the EDF, tests based on empirical transforms, as well as a criterion based on a moment-type characterization of the exponential distribution. Based on the results of the Monte Carlo study conducted in this paper, it was found that the test of Epps and Pulley outperformed the other tests under most alternatives. Also the BH test based on the empirical Laplace transform and the HM test using the empirical characteristic function although they did not achieve the highest powers, they are consistently good performers. We conclude by noting that an analogous Monte Carlo study was carried out for AFT models. The performance of the tests under the AFT model was however similar to the performance observed for the PH models and it is therefore not reported.

ORCID

Marike Cockeran  <http://orcid.org/0000-0002-3990-8345>

References

- Balakrishnan, N., E. Chimitova, N. Galanova, and M. Vedernikova. 2013. Testing goodness of fit of parametric AFT and PH models with residuals. *Communications in Statistics-Simulation and Computation* 42 (6):1352–67. doi:[10.1080/03610918.2012.659824](https://doi.org/10.1080/03610918.2012.659824).
- Baringhaus, L., and N. Henze. 1991. A class of consistent tests for exponentiality based on the empirical Laplace transform. *Annals of the Institute of Statistical Mathematics* 43 (3):551–64. doi:[10.1007/BF00053372](https://doi.org/10.1007/BF00053372).
- D'Agostino, R., and M. Stephens. 1986. *Goodness-of-fit techniques. Series Statistics: Textbooks and Monographs*, 68. Boca Raton, FL: CRC Press.
- Epps, T., and L. Pulley. 1986. A test of exponentiality vs. monotone-hazard alternatives derived from the empirical characteristic function. *Journal of the Royal Statistical Society. Series B (Methodological)* 48 (2):206–13. doi:[10.1111/j.2517-6161.1986.tb01403.x](https://doi.org/10.1111/j.2517-6161.1986.tb01403.x).
- Giacomini, R., D. N. Politis, and H. White. 2013. A warp-speed method for conducting Monte Carlo experiments involving bootstrap estimators. *Econometric Theory* 29 (3):567–89. doi:[10.1017/S0266466612000655](https://doi.org/10.1017/S0266466612000655).
- Henze, N., and S. G. Meintanis. 2005. Recent and classical tests for exponentiality: A partial review with comparisons. *Metrika* 61 (1):29–45. doi:[10.1007/s001840400322](https://doi.org/10.1007/s001840400322).
- Hsieh, F., and P. W. Lavori. 2000. Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. *Controlled Clinical Trials* 21 (6):552–60. doi:[10.1016/S0197-2456\(00\)00104-5](https://doi.org/10.1016/S0197-2456(00)00104-5).
- Klein, J. P., and M. L. Moeschberger. 2006. *Survival analysis: Techniques for censored and truncated data*. New York, NY: Springer Science & Business Media.
- Lee, S., C. Locke, and J. D. Spurrier. 1980. On a class of tests of exponentiality. *Technometrics* 22 (4):547–54. doi:[10.2307/1268192](https://doi.org/10.2307/1268192).
- Meintanis, S. G., J. Swanepoel, and J. Allison. 2014. The probability weighted characteristic function and goodness-of-fit testing. *Journal of Statistical Planning and Inference* 146 (1):122–32. doi:[10.1016/j.jspi.2013.09.011](https://doi.org/10.1016/j.jspi.2013.09.011).
- Mimoto, N., and R. Zitikis. 2008. The Atkinson index, the Moran statistic, and testing exponentiality. *Journal of the Japan Statistical Society* 38 (2):187–205. doi:[10.14490/jjss.38.187](https://doi.org/10.14490/jjss.38.187).
- The R Development Core Team. 2016. *R: A language and environment for statistical computing [Computer software manual]*. Vienna, Austria: The R Foundation. <http://www.R-project.org>.

Addendum to Article 1: Consistency results

Suppose we observe the survival times $T_1, \dots, T_n$ of n independent individuals and that the cumulative hazard rate function of the k th individual can be described by the model

$$H(\mathbf{x}_k) = -\log S(\mathbf{x}_k) = r(\mathbf{x}_k; \boldsymbol{\beta}_0) H_0(\cdot; \boldsymbol{\theta}_0), \quad (2.1)$$

where r is some positive, real-valued regression function, $\mathbf{x}_k$ is a p -vector containing the values of the p co-variates associated with the k th individual (assumed fixed), $\boldsymbol{\beta}_0$ is a p -vector containing the true (unknown) regression parameters, H_0 denotes the baseline cumulative hazard rate function, $\boldsymbol{\theta}_0$ is a q -vector of parameters (unknown), and $S(t|\mathbf{x}_k)$ is the conditional survival function of T_k given $\mathbf{x}_k$. Clearly, if the above model is correct and $S(t|\mathbf{x}_k)$ is the true survival function of T_k , then, by the probability integral transform, $S(T_k|\mathbf{x}_k)$ has a uniform(0,1) distribution and consequently $H(T_k|\mathbf{x}_k) = -\log S(T_k|\mathbf{x}_k)$ a standard exponential distribution. Moreover, $H(T_1|\mathbf{x}_1), \dots, H(T_n|\mathbf{x}_n)$ are mutually independent. Notice also that under this model the observations have the representation

$$T_k = H_0^{-1} \left(\frac{-\log U_k}{r(\mathbf{x}_k; \boldsymbol{\beta}_0)}; \boldsymbol{\theta}_0 \right), \quad k = 1, \dots, n, \quad (2.2)$$

where $U_1, \dots, U_n$ are independent uniform(0,1) random variables. Given estimators $\hat{\boldsymbol{\beta}}_n$ and $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$, we may calculate the Cox–Snell residuals as

$$H(T_k|\mathbf{x}_k) = r(\mathbf{x}_k; \hat{\boldsymbol{\beta}}_n) H_0(T_k; \hat{\boldsymbol{\theta}}_n), \quad k = 1, \dots, n,$$

The test is based on the criterion

$$H_{n,a} = n \int_0^\infty \left[(1+t) \psi_n(t; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_n) - 1 \right]^2 w(t) e^{-at} dt,$$

where

$$\psi_n(t; \boldsymbol{\theta}, \boldsymbol{\beta}) = \frac{1}{n} \sum_{k=1}^n e^{-tr(\mathbf{x}_k; \boldsymbol{\beta}) H_0(T_k; \boldsymbol{\theta})}.$$

Our results rely on the following assumptions:

(A.1) There exist parameters $\boldsymbol{\theta}_A$ and $\boldsymbol{\beta}_A$ such that $\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_A$ and $\hat{\boldsymbol{\beta}}_n \xrightarrow{P} \boldsymbol{\beta}_A$ as $n \rightarrow \infty$.

(A.2) The baseline cumulative hazard function H_0 is strictly positive and, for all $u > 0$, satisfies the Lipschitz condition $|H_0(u; \boldsymbol{\theta}) - H_0(u; \boldsymbol{\theta}_A)| \leq \|\boldsymbol{\theta} - \boldsymbol{\theta}_A\| d_1(u)$ for all $\boldsymbol{\theta}$ in an open neighborhood of $\boldsymbol{\theta}_A$, where d_1 is a non-negative real-valued function satisfying $\limsup_{n \rightarrow \infty} \sup_{1 \leq k \leq n} E d_1(T_k) < \infty$.

(A.3) $\mathbf{x}_k \in \mathcal{R}_X$ for all $k = 1, \dots, n$, where $\mathcal{R}_X$ is a compact subset of $\mathbb{R}^p$. The regression function r is strictly positive and $r(\mathbf{x}, \boldsymbol{\beta}_A) < \infty$ for all $\mathbf{x} \in \mathcal{R}_X$. The regression function also satisfies the Lipschitz condition $|r(\mathbf{x}; \boldsymbol{\beta}) - r(\mathbf{x}; \boldsymbol{\beta}_A)| \leq \|\boldsymbol{\beta} - \boldsymbol{\beta}_A\| d_2(\mathbf{x})$ in a neighborhood of $\boldsymbol{\beta} = \boldsymbol{\beta}_A$, where d_2 is a non-negative real-valued function satisfying $\limsup_{n \rightarrow \infty} \max_{1 \leq k \leq n} d_2(\mathbf{x}_k) < \infty$.

(A.4) $w(\cdot)$ is a non-negative weight function with $w(t) = O(t^m)$ as $t \rightarrow \infty$ for some non-negative integer m .

Lemma 1. *Under Assumptions (A.1)–(A.3) it holds for each fixed $t \in \mathbb{R}^+$ that*

$$\left| \psi_n(t; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_n) - \mathbb{E} \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A) \right| \xrightarrow{\mathbb{P}} 0 \quad (2.3)$$

as $n \rightarrow \infty$. Furthermore, for any natural number ℓ one has for each $t \in \mathbb{R}^+$ that

$$\mathbb{E} \left| \psi_n(t; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_n) - \mathbb{E} \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A) \right|^\ell \rightarrow 0. \quad (2.4)$$

Proof. Using the fact that, for each fixed $t > 0$, the function $v \mapsto e^{-tv}$ is Lipschitz continuous on $[0, \infty)$, there exists a finite constant $M > 0$ such that

$$\left| e^{-tr(\mathbf{x}; \hat{\boldsymbol{\beta}}_n)H_0(u; \hat{\boldsymbol{\theta}}_n)} - e^{-tr(\mathbf{x}; \boldsymbol{\beta}_A)H_0(u; \boldsymbol{\theta}_A)} \right| \leq M \left| r(\mathbf{x}; \hat{\boldsymbol{\beta}}_n)H_0(u; \hat{\boldsymbol{\theta}}_n) - r(\mathbf{x}; \boldsymbol{\beta}_A)H_0(u; \boldsymbol{\theta}_A) \right|$$

for all $\mathbf{x} \in \mathcal{R}_X$ and all $u > 0$. Adding and subtracting terms and applying the triangle inequality one obtains

$$\begin{aligned} & \left| r(\mathbf{x}_k; \hat{\boldsymbol{\beta}}_n)H_0(T_k; \hat{\boldsymbol{\theta}}_n) - r(\mathbf{x}_k; \boldsymbol{\beta}_A)H_0(T_k; \boldsymbol{\theta}_A) \right| \\ & \leq \left| r(\mathbf{x}_k; \hat{\boldsymbol{\beta}}_n) - r(\mathbf{x}_k; \boldsymbol{\beta}_A) \right| \left| H_0(T_k; \hat{\boldsymbol{\theta}}_n) - H_0(T_k; \boldsymbol{\theta}_A) \right| + \left| r(\mathbf{x}_k; \hat{\boldsymbol{\beta}}_n) - r(\mathbf{x}_k; \boldsymbol{\beta}_A) \right| H_0(T_k; \boldsymbol{\theta}_A) \\ & \quad + \left| H_0(T_k; \hat{\boldsymbol{\theta}}_n) - H_0(T_k; \boldsymbol{\theta}_A) \right| r(\mathbf{x}_k; \boldsymbol{\beta}_A) \\ & \leq \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_A\| d_2(\mathbf{x}_k) \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_A\| d_1(T_k) + \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_A\| d_2(\mathbf{x}_k) H_0(T_k; \boldsymbol{\theta}_A) \\ & \quad + \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_A\| d_1(T_k) r(\mathbf{x}_k; \boldsymbol{\beta}_A) \end{aligned}$$

for each $k = 1, \dots, n$, where we have made use of the Lipschitz conditions in (A.2) and (A.3). Hence, $\psi_n(t; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_n) - \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A)$ is bounded from above by

$$\begin{aligned} |\psi_n(t; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_n) - \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A)| & \leq M \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_A\| \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_A\| \frac{1}{n} \sum_{k=1}^n d_1(T_k) d_2(\mathbf{x}_k) \\ & \quad + M \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_A\| \frac{1}{n} \sum_{k=1}^n H_0(T_k; \boldsymbol{\theta}_A) d_2(\mathbf{x}_k) \\ & \quad + M \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_A\| \frac{1}{n} \sum_{k=1}^n r(\mathbf{x}_k; \boldsymbol{\beta}_A) d_1(T_k), \end{aligned}$$

which converges to 0 in probability under Assumptions (A.1)–(A.3). Therefore,

$$\psi_n(t; \widehat{\boldsymbol{\theta}}_n, \widehat{\boldsymbol{\beta}}_n) = \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A) + o_P(1).$$

Now noting that, for each fixed $t \in \mathbb{R}^+$, the quantity $\psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A)$ is a mean of independent random variables with mean and variance bounded by 1 and 2, respectively. Hence, by Kolmogorov's law of large numbers (see, e.g., Serfling 1980, p. 27) we have

$$\left| \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A) - \mathbb{E} \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A) \right| = \left| \frac{1}{n} \sum_{k=1}^n e^{-tr(\mathbf{x}_k; \boldsymbol{\beta}_A) H_0(T_k; \boldsymbol{\theta}_A)} - \frac{1}{n} \sum_{k=1}^n \mathbb{E} e^{-tr(\mathbf{x}_k; \boldsymbol{\beta}_A) H_0(T_k; \boldsymbol{\theta}_A)} \right| \xrightarrow{P} 0.$$

To verify the second result of the theorem, notice that $|\psi_n(t; \cdot, \cdot)| \leq 1$ together with the inequality $|x + y|^\ell \leq 2^{\ell-1}|x|^\ell + 2^{\ell-1}|y|^\ell$ imply

$$|\tilde{\psi}_n(t)|^\ell := |\psi_n(t; \widehat{\boldsymbol{\theta}}_n, \widehat{\boldsymbol{\beta}}_n) - \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A)|^\ell \leq 2^{\ell-1} |\psi_n(t; \widehat{\boldsymbol{\theta}}_n, \widehat{\boldsymbol{\beta}}_n)|^\ell + 2^{\ell-1} |\psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A)|^\ell \leq 2^\ell.$$

From this and Chebyshev's inequality it follows that

$$\mathbb{E} |\tilde{\psi}_n(t)| \mathbf{1}(|\tilde{\psi}_n(t)| > c) \leq 2^\ell \mathbb{E} \mathbf{1}(|\tilde{\psi}_n(t)| > c) = 2^\ell \mathbb{P}(|\tilde{\psi}_n(t)| > c) \leq \frac{2^\ell}{c} \mathbb{E} |\tilde{\psi}_n(t)| \leq \frac{4^\ell}{c},$$

for any constant $c > 0$, where $\mathbf{1}(\cdot)$ denotes the indicator function. Hence

$$\lim_{c \rightarrow \infty} \sup_n \mathbb{E} |\tilde{\psi}_n(t)| \mathbf{1}(|\tilde{\psi}_n(t)| > c) = 0.$$

This, together with (2.3), implies (2.4); see Serfling (1980, p. 15). □

Theorem 1. *Under assumptions (A.1)–(A.4) it follows that*

$$\frac{1}{n} H M_{n,a} = \int_0^\infty \left[(1+t) \mathbb{E} \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A) - 1 \right]^2 w(t) e^{-at} dt + o_P(1)$$

as $n \rightarrow \infty$. Moreover,

$$\frac{1}{n} H M_{n,a} \xrightarrow{P} H M_a := \int_0^\infty \left[(1+t) \psi_A(t) - 1 \right]^2 w(t) e^{-at} dt,$$

where $\psi_A(t) = \lim_{n \rightarrow \infty} \mathbb{E} \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A)$, provided that this limit exists.

Proof. Write $n^{-1}HM_{n,a} = n^{-1}HM_{n,a}^* + A_n + B_n$, where

$$\begin{aligned} HM_{n,a}^* &= n \int_0^\infty [(1+t) E \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A) - 1]^2 w(t) e^{-at} dt, \\ A_n &= \int_0^\infty [\psi_n(t; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_n) - E \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A)]^2 (1+t)^2 w(t) e^{-at} dt, \\ B_n &= 2 \int_0^\infty [\psi_n(t; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_n) - E \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A)][(1+t) E \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A) - 1](1+t) w(t) e^{-at} dt. \end{aligned}$$

We first show that $A_n \xrightarrow{P} 0$. Notice that $[\psi_n(t; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_n) - E \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A)]^2 \leq 4$ and $w(t) = O(t^m)$ imply

$$E \left(\int_0^\infty [\psi_n(t; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_n) - E \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A)]^2 (1+t)^2 w(t) e^{-at} dt \right) \leq 4 \int_0^\infty (1+t)^2 w(t) e^{-at} dt < \infty.$$

We may therefore invoke Fubini's theorem and the dominated convergence theorem to obtain

$$\lim_{n \rightarrow \infty} E A_n = \int_0^\infty \lim_{n \rightarrow \infty} E [\psi_n(t; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_n) - E \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A)]^2 (1+t)^2 w(t) e^{-at} dt = 0,$$

where we have made use of (2.4) with $\ell = 2$. Since $A_n \geq 0$, it follows by Chebyshev's inequality that $A_n \xrightarrow{P} 0$.

To see that $B_n \xrightarrow{P} 0$, first observe that by the boundedness of $\psi_n(\cdot)$ one has

$$\begin{aligned} & E \int_0^\infty |[\psi_n(t; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_n) - E \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A)][(1+t) E \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A) - 1]| (1+t) w(t) e^{-at} dt \\ & \leq 2 \int_0^\infty (2+t)(1+t) w(t) e^{-at} dt < \infty, \end{aligned}$$

so that, by Fubini's theorem,

$$E B_n = 2 \int_0^\infty E [\psi_n(t; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_n) - E \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A)][(1+t) E \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A) - 1](1+t) w(t) e^{-at} dt \rightarrow 0.$$

Moreover, by the Cauchy-Schwarz inequality,

$$\begin{aligned} B_n^2 &= 4 \left(\int_0^\infty [\psi_n(t; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_n) - E \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A)][(1+t) E \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A) - 1](1+t) w(t) e^{-at} dt \right)^2 \\ &\leq 4 \int_0^\infty [\psi_n(t; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_n) - E \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A)]^2 w(t) e^{-at} dt \\ &\quad \times \int_0^\infty [(1+t) E \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A) - 1]^2 (1+t)^2 w(t) e^{-at} dt. \end{aligned} \tag{2.5}$$

As before, we have by Fubini's theorem that we may interchange the order of the expectation and the integral

in the first factor above, so that, by the dominated convergence theorem,

$$\mathbb{E} \int_0^\infty [\psi_n(t; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_n) - \mathbb{E} \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A)]^2 e^{-at} dt = \int_0^\infty \mathbb{E} [\psi_n(t; \hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_n) - \mathbb{E} \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A)]^2 e^{-at} dt \rightarrow 0,$$

where we have made use of (2.4). As the remaining integral in (2.5) is finite, it follows that $\mathbb{E} B_n^2 \rightarrow 0$. Hence, $B_n \xrightarrow{\mathbb{P}} 0$. The second claim in the theorem follows by similar arguments. $\square$

We now demonstrate how the result in Theorem 1 may be used to establish consistency of the test $H_{n,a}$ against certain alternatives. Recall that if the model in (2.1) is indeed the true underlying model, then $\hat{\boldsymbol{\beta}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\theta}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_0$ as $n \rightarrow \infty$ and by Theorem 1 we have

$$\frac{1}{n} H_{n,a} = \int_0^\infty [(1+t) \mathbb{E} \psi_n(t; \boldsymbol{\theta}_0, \boldsymbol{\beta}_0) - 1]^2 w(t) e^{-at} dt + o_{\mathbb{P}}(1). \quad (2.6)$$

Under the null hypothesis the expectation in the expression above becomes

$$\mathbb{E} \psi_n(t; \boldsymbol{\theta}_0, \boldsymbol{\beta}_0) = \mathbb{E} e^{-tr(x_k; \boldsymbol{\beta}) H_0(T_k; \boldsymbol{\theta})} = \mathbb{E} U_k^t,$$

where we have made use of the representation in (2.2). Since the U_k are independent uniform(0,1) random variables, one has that

$$\mathbb{E} \psi_n(t; \boldsymbol{\theta}_0, \boldsymbol{\beta}_0) = \frac{1}{1+t}. \quad (2.7)$$

Clearly the integral in (2.6) will be zero if and only if (2.7) holds for all $t > 0$, provided that $w(t) > 0$ for all $t > 0$.

Now suppose that the null hypothesis is not satisfied. Then we may write the statistic as

$$\mathbb{E} \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A) = \frac{1}{n} \sum_{k=1}^n \mathbb{E} e^{-tr(x_k; \boldsymbol{\beta}_A) H_0(T_k; \boldsymbol{\theta}_A)} = \frac{1}{n} \sum_{k=1}^n \mathbb{E} \tilde{U}_k^t,$$

where $\tilde{U}_1, \dots, \tilde{U}_n$ are mutually independent random variables taking values in the interval (0,1). Now clearly $(1+t)\psi_A(t) - 1 = 0$ if and only if $\psi_A(t) = \lim_{n \rightarrow \infty} \mathbb{E} \psi_n(t; \boldsymbol{\theta}_A, \boldsymbol{\beta}_A) = (1+t)^{-1}$. If this is not the case for all $t > 0$, then

$$HM_a = \int_0^\infty [(1+t)\psi_A(t) - 1]^2 w(t) e^{-at} dt > 0.$$

Therefore, by Theorem 1, one has that $HM_{n,a} \xrightarrow{\mathbb{P}} \infty$ as $n \rightarrow \infty$ and hence the test is consistent.

Chapter 3

Article 2: Goodness-of-fit testing in survival models in the presence of Type-II right censoring

The second article, namely: *Goodness-of-fit testing of survival models in the presence of Type-II right censoring*, has been submitted to the journal *Lifetime Data Analysis*. A summary of the guidelines to authors from the journal is now presented.

Manuscript	A maximum length of the manuscript is not provided.
Title	A concise and informative title should be provided.
Abstract and keywords	A maximum of 150–200 words succinctly describing the article and between 4–6 keywords.
Tables	All tables are to be numbered using Arabic numerals. All tables should be referred to sequentially in the text.
References	The list of references should only include works that are cited in the text and that have been published or accepted for publication.
General formatting	A LaTeX template is provided for submission.
Additional information	https://tinyurl.com/yyzeaybu

Goodness-of-fit testing of survival models in the presence of Type-II right censoring

M. Cockeran · S. G. Meintanis · L. Santana · J. S. Allison

Received: date / Accepted: date

Abstract We consider a variety of tests for testing goodness-of-fit in a parametric Cox proportional hazards (PH) and accelerated failure time (AFT) model in the presence of Type-II right censoring. The testing procedures considered can be divided in two categories: an approach involving transforming the data to a complete sample and an approach where the test statistic is modified to accommodate Type-II right censoring. The power of the proposed tests are compared through a Monte Carlo study.

Keywords Bootstrap · Goodness-of-fit · Type-II right censoring · Survival analysis

1 Introduction

Consider a life-testing experiment in which n identical and independent units are placed on a life test. Let $T_1, \dots, T_n$ be independent, non-negative random variables representing the lifetimes of these experimental units with unknown probability density function $f(t)$ and unknown cumulative distribution function $F(t)$. Suppose the experiment is carried out until the time of the r^{th} failure, and so the data arising from such a life test would be of the form $T_{1:n} < T_{2:n} < \dots < T_{r:n}$, $r < n$, with the remaining $n - r$ lifetimes being greater than $T_{r:n}$. This situation is referred to as Type-II right censoring and is a useful technique for economical use of resources in these types of life-testing experiments. In these experiments it is of interest to examine the effect of a set of covariates on time-to-event outcomes, for which two approaches have become popular in the statistical literature, the Cox proportional hazards (PH) model and the accelerated failure time (AFT) model. For an in-depth discussion and comparison of these two models, the reader is referred to Qi (2009) and Klein and Moeschberger (2006). Novak (2010) also discusses various methods of checking the goodness-of-fit of these models to observed censored data. In the first approach the cumulative hazard rate

M. Cockeran
Unit for Business Mathematics and Informatics, North-West University, Potchefstroom, South Africa
E-mail: Marike.Cockeran@nwu.ac.za

S. G. Meintanis
Department of Economics, National and Kapodistrian University of Athens, Athens, Greece (On sabbatical leave from the University of Athens)

L. Santana
Unit for Business Mathematics and Informatics, North-West University, Potchefstroom, South Africa

J. S. Allison
Unit for Business Mathematics and Informatics, North-West University, Potchefstroom, South Africa

$A(t)$ of the random variable T is modeled by means of the PH model

$$A(t) = r(\mathbf{x})H(t), \quad (1)$$

where $r(\cdot)$ is a regression function that depends on the covariate $\mathbf{x} = (x_1, x_2, \dots, x_m)$ and $H(\cdot)$ is some unspecified baseline cumulative hazard rate function. In some cases it is appropriate to make specific distributional assumptions about the shape of the survival function. In this study the determination of model adequacy is based on formal hypothesis testing. To fix notation let $Y_1, Y_2, \dots, Y_n$ denote the failure times in a sample of size n . Let $r < n$ be fixed and denote by $Y_{1:n} < Y_{2:n} < \dots < Y_{r:n}$ the first r ordered values of $Y_1, Y_2, \dots, Y_n$. Under Type-II right censoring, we only observe $T_{1:n}, T_{2:n}, \dots, T_{r:n}$, where $T_{j:n} = Y_{j:n}$, $j = 1, 2, \dots, r$. For ease of notation define these observed values along with their corresponding covariates by $(T_j, \mathbf{x}_j^\top)$, $j \in \{k : T_k \leq T_{r:n}\}$. On the basis of the observations $(T_j, \mathbf{x}_j)$, $j \in \{k : T_k \leq T_{r:n}\}$ we wish to test the null hypothesis

$$\mathcal{H}_0 : H(t) = H_0(t; \boldsymbol{\theta}), \text{ for some } \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k, \quad (2)$$

where $H_0(\cdot; \boldsymbol{\theta})$ is a parametric baseline cumulative hazard rate function depending on some unknown parameter $\boldsymbol{\theta}$. We will adopt a specified regression function $r(\mathbf{x}) = r(\mathbf{x}; \boldsymbol{\beta})$ that depends on the parameter $\boldsymbol{\beta} \in \mathcal{B} \subset \mathbb{R}^m$. Under this assumption, the cumulative hazard function of the j^{th} individual can be described by the PH model as

$$A_j(t; \boldsymbol{\psi}) = r(\mathbf{x}_j; \boldsymbol{\beta})H_0(t; \boldsymbol{\theta}), \quad t \geq 0, \quad (3)$$

where $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta})$. Now let $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ be some estimator of $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta})$ (throughout the paper we make use of maximum likelihood estimation). Then if the Cox PH model is true and under the null hypothesis, the so-called Cox-Snell residuals from the PH model defined in (1) should (approximately) follow a unit exponential distribution. These residuals are defined as

$$\hat{\varepsilon}_j = r(\mathbf{x}_j; \hat{\boldsymbol{\beta}})H_0(T_j; \hat{\boldsymbol{\theta}}), \quad j = 1, 2, \dots, r. \quad (4)$$

Hence any exponentiality test on the basis of $\hat{\varepsilon}_j$, $j = 1, 2, \dots, r$, constitutes in effect a test for the PH model itself. Another possible model violation relates to more general configurations of the PH model implied by (3) such as the Hsieh model; see Hsieh and Lavori (2000). The Hsieh model is a generalization of the PH model and is given by

$$A_j(t; \boldsymbol{\psi}, \boldsymbol{\gamma}) = r(\mathbf{x}_j; \boldsymbol{\beta})H_0(t; \boldsymbol{\theta})e^{\boldsymbol{\gamma}^\top \mathbf{x}_j},$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_m) \in \mathbb{R}^m$ is a set of unknown positive parameters. Clearly if $\boldsymbol{\gamma} = 0$ then the Hsieh model reduces to the PH model. It will be shown that our tests pick up this type of violation, from a PH model with a given hazard towards a Hsieh model with the same baseline hazard.

The next model we consider is the AFT model. In terms of the cumulative hazard rate the AFT model for individual j is given by

$$A_j(t) = H\left(\frac{t}{r(\mathbf{x}_j; \boldsymbol{\beta})}\right), \quad (5)$$

with $r(\mathbf{x}_j; \boldsymbol{\beta}) = e^{\boldsymbol{\beta}^\top \mathbf{x}_j}$. Alternatively, the AFT model can be represented as a log-linear model given by

$$\log(T_j) = \boldsymbol{\beta}^\top \mathbf{x}_j + \varepsilon_j,$$

where ε_j is a random variable assumed to have a particular distribution. Note that the cumulative hazard function of e^{ε_j} is H . The Cox-Snell residuals for the AFT model are given by

$$\hat{\varepsilon}_j = \log(T_j) - \hat{\boldsymbol{\beta}}^\top \mathbf{x}_j.$$

Now, under $\mathcal{H}_0$, $F_0(\hat{\varepsilon}_j)$, $j = 1, \dots, r$, should approximately be uniformly distributed on $(0,1)$, where $F_0(\cdot)$ is the distribution function of ε_j under the null hypothesis. Hence, the distribution of

$$\hat{\varepsilon}_j^T = -\log(1 - F_0(\hat{\varepsilon}_j))$$

should be approximately unit exponential if the model is specified correctly.

In what follows we consider two approaches for testing the hypothesis in (2) with Type-II right censored data. The first approach is to employ test statistics which were developed for complete samples (i.e., $r = n$), and modify them accordingly in order to make them applicable for Type-II right censored data. The second approach is to transform a censored sample of size $r < n$, to a complete sample of size r . Based on the complete sample of size r , any of the numerous tests for exponentiality can be used. See, e.g., the review papers by Allison, Santana, Smit, and Visagie (2017), and Henze and Meintanis (2005).

It should be mentioned at the outset that there are alternative methods of testing goodness-of-fit to PH (and AFT models); see for instance Caroni (2004). Such methods include graphical analyses of residuals, which however are based on subjective assessment and they should therefore be employed only at a preliminary data-analysis level. On the other hand one may opt for classical chi-squared tests but this approach is extremely sensitive to the type of data-grouping, which again is chosen by the practitioner. In this paper we will consider and compare a variety of tests for testing goodness-of-fit to the above mentioned models under Type-II right censoring by means of comparing the distribution of properly estimated model-based residuals with that of a unit exponential variate. Compared to the aforementioned alternative procedures, and apart from removing subjective elements, the methods employed here are omnibus, meaning that they are able to detect arbitrary deviations from the null hypothesis provided that the sample size is sufficiently large.

The rest of the article is organized as follows: In Section 2 we present the transformations that lead from a censored sample to a complete sample. Section 3 provides an outline of bootstrap procedures used, both under the Cox PH model as well as under the AFT model, in order to obtain the critical value of any given test. In Section 4 we provide details of the various tests for exponentiality, with complete and censored samples, included in the simulation study. Section 5 presents the results of an extensive Monte Carlo simulation which compares the performance of the tests for different pairs of competing hypotheses. The paper concludes with final remarks and a discussion on the performance of the tests in Section 6.

2 Transformations to complete samples

As previously noted, one approach for testing the problem in (2) is to apply goodness-of-fit statistics that have specifically been tailored to handle the case of censored data. These test statistics are discussed in Section 4.6. A second strategy is to transform a censored sample of size $r < n$, to a complete sample of size r and then apply any of the numerous tests for exponentiality for complete samples of size r . Implementing this strategy necessitates the use of uniform $(0,1)$ random values, obtained by the transformation $U_j = G(\hat{\varepsilon}_j)$, where $G(x) = 1 - e^{-x}$, so that, if the conditions stipulated in $\mathcal{H}_0$ are satisfied, then the resulting order statistics, $U_{1:n} < \dots < U_{r:n}$, may approximately be viewed as the first r order statistics in a random sample of size n from a $U(0,1)$ distribution. Examples of such transformations appear in Michael and Schucany (1979), O'Reilly and Stephens (1988), Lin, Huang, and Balakrishnan (2008), and Fischer and Kamps (2011). Specifically, we consider the following two transformations, as they have shown to produce the highest empirical powers in a preliminary Monte Carlo study conducted by the authors:

- Transformation by O'Reilly and Stephens (1988)

$$v_{i:r} = 1 - \prod_{j=1}^i \left[\frac{1 - U_{j:n}}{1 - U_{j-1:n}} \right]^{\frac{n-j+1}{r-j+1}}, \quad i = 1, 2, \dots, r, \quad \text{with } U_{0:n} \equiv 0.$$

– Transformation by Fischer and Kamps (2011)

$$v_{i:r} = \prod_{j=i}^r \left[1 - \left(\frac{1 - U_{j:n}}{1 - U_{j-1:n}} \right)^{n-j+1} \right]^{1/j}, \quad i = 1, 2, \dots, r, \text{ with } U_{0:n} \equiv 0.$$

For independent identically distributed (i.i.d.) data, these transformations are structured in such a way that $v_{1:r} < \dots < v_{r:r}$, represents a complete random sample of size r from a $U(0, 1)$ distribution. In what follows we write OS and FK for these two transformations, respectively. The final step is to then transform the complete sample of uniform values using $Z_j = G^{-1}(v_{j:r})$, to produce $Z_1, \dots, Z_r$, which is not only considered to be a complete sample of size r , but will also have an approximate standard exponential distribution if the conditions of the null hypothesis are satisfied. The above three steps are summarized in the following algorithm where we apply the transformations to the residuals of the PH model.

1. Calculate $U_{i:n} = G(\widehat{\varepsilon}_{i:n})$, $i = 1, 2, \dots, r$, where $\widehat{\varepsilon}_{1:n} < \widehat{\varepsilon}_{2:n} < \dots < \widehat{\varepsilon}_{r:n}$ denote the ordered residuals.
2. Transform to $v_{i:r} = \tau(U_{i:n})$, $i = 1, 2, \dots, r$, where $\tau(\cdot)$ denotes either the OS or FK transformation.
3. Calculate $Z_j = G^{-1}(v_{j:r})$, $i = 1, 2, \dots, r$.

Goldmann, Klar, and Meintanis (2015) successfully used these transformations to test the goodness-of-fit for various distributions in the i.i.d. case. In the current context however, these test statistics should be implemented in a way that takes into account the model being tested, as well as the variability of the corresponding parameter estimates present in the null hypothesis $\mathcal{H}_0$ in (2). In the next section we provide details of parametric bootstrap resampling schemes which are appropriate for approximating the limit distribution of the test statistics.

3 Bootstrap estimation of critical values

Write $S := S(Z_1, Z_2, \dots, Z_r)$, for an arbitrary test statistic. We now illustrate the combined transformation and the application of the bootstrap procedure in order to calculate the critical values of S under the PH model.

1. Generate a bootstrap sample of failure times $Y_1^*, Y_2^*, \dots, Y_n^*$, where

$$Y_j^* = H_0^{-1} \left[-\log(Q_j) r(\mathbf{x}_j; \widehat{\boldsymbol{\beta}})^{-1}; \widehat{\boldsymbol{\theta}} \right],$$

and where Q_j , $j = 1, \dots, n$, are independent random variates from a uniform distribution on $(0, 1)$. Denote by $Y_{1:n}^* < Y_{2:n}^* < \dots < Y_{r:n}^*$, the first r ordered values of $Y_1^*, Y_2^*, \dots, Y_n^*$.

2. Now, based on a Type-II right censored sample of size r from these failure times, fit a PH model under $\mathcal{H}_0$ using the observations $(T_j^*, \mathbf{x}_j)$, $j \in \{k : T_k^* \leq T_{r:n}^*\}$, where $T_{j:n}^* = Y_{j:n}^*$, $j = 1, 2, \dots, r$.
3. Calculate the Cox-Snell residuals for the PH model fitted in step 2,

$$\widehat{\varepsilon}_j^* = r(\mathbf{x}_j, \widehat{\boldsymbol{\beta}}^*) H_0(T_j^*, \widehat{\boldsymbol{\theta}}^*), \quad j = 1, 2, \dots, r,$$

where $\widehat{\boldsymbol{\psi}}^* = (\widehat{\boldsymbol{\beta}}^*, \widehat{\boldsymbol{\theta}}^*)$ are the estimators of the model parameters based on the censored bootstrap sample in step 2.

4. Calculate $U_{j:r}^* = G(\widehat{\varepsilon}_{j:n}^*) = 1 - e^{-\widehat{\varepsilon}_{j:n}^*}$, where $\widehat{\varepsilon}_{1:n}^* < \widehat{\varepsilon}_{2:n}^* < \dots < \widehat{\varepsilon}_{r:n}^*$ denote the ordered values of $\widehat{\varepsilon}_1^*, \widehat{\varepsilon}_2^*, \dots, \widehat{\varepsilon}_r^*$.
5. Transform to $v_{j:r}^* = \tau(U_{j:r}^*)$ where $\tau(\cdot)$ denotes either the OS or the FK transformation.
6. Calculate $Z_j^* = G^{-1}(v_{j:r}^*)$ and compute the test statistic, say $S^* := S(Z_1^*, Z_2^*, \dots, Z_r^*)$.
7. Repeat steps 1 to 6 a number of times, say B , and obtain $\{S_m^*\}_{m=1}^B$.

8. The null hypothesis is rejected at a prescribed significance level α if $S > S_{[B-\alpha B]}^*$, where $S_{(1)}^* \leq \dots \leq S_{(B)}^*$ denote the order statistics.

The algorithm to obtain the critical values under the AFT model is nearly identical to the one stated above, with the exception that the first three steps change as follows:

- 1*. Generate a bootstrap sample of failure times $Y_1^*, Y_2^*, \dots, Y_n^*$, where

$$Y_j^* = e^{\hat{\beta}_n^\top \mathbf{x}_j + \varepsilon_j},$$

where ε_j , $j = 1, \dots, n$, are i.i.d. and such that the random variable e^{ε_j} , has cumulative hazard rate H_0 .

- 2*. Now construct a Type-II right censored sample of these failure times and fit an AFT model under $\mathcal{H}_0$ using the observations $(T_j^*, \mathbf{x}_j)$, $j \in \{k : T_k^* \leq T_{r:n}^*\}$ and where $T_{j:n}^* = Y_{j:n}^*$, $j = 1, 2, \dots, r$.
- 3*. Calculate the residuals for the AFT model fitted in step 2*,

$$\hat{\varepsilon}_j^{T*} = -\log(1 - F_0(\hat{\varepsilon}_j^*)), \quad (6)$$

where $\hat{\varepsilon}_j^* = \log(T_j^*) - \hat{\beta}^{*\top} \mathbf{x}_j$ and $\hat{\beta}^*$ are the estimators of the model parameters based on the censored bootstrap sample in step 2* and $F_0(\cdot)$ is the distribution function of ε_j .

The implementation of a Monte Carlo simulation employing the algorithm described above will potentially incur a great cost in computational time due to the nested iteration structure involved with attempting to evaluate a bootstrap procedure in a Monte Carlo. To alleviate this computational burden, the so-called “warp speed” bootstrap procedure, given by Giacomini, Politis, and White (2013), will be used to approximate the bootstrap critical value in the Monte Carlo study. If we denote the number of Monte Carlo replications by MC , then this bootstrap procedure capitalizes on the repetition inherent in the Monte Carlo simulation to produce bootstrap replications, rather than relying on a separate “bootstrap loop”. By calculating the bootstrap test statistic S^* for only one bootstrap sample generated for each Monte Carlo iteration, we ultimately obtain MC many bootstrap sample test statistics at the end of the simulation, i.e., we obtain the set $\{S_m^*\}_{m=1}^{MC}$. The size- α critical value can then easily be obtained (employing a method similar to the one described in step 8) by simply finding the $(1 - \alpha)^{\text{th}}$ quantile of this set of bootstrap statistics.

4 Tests for exponentiality

To test the hypothesis that the observed transformed residuals $Z_1, Z_2, \dots, Z_r$, are realizations from a standard exponential distribution, we use a number of goodness-of-fit tests for exponentiality. Note that some of these tests were also included in an analogous Monte Carlo study for non-censored data conducted recently by Cockeran, Meintanis, and Allison (2019). The tests are arranged according to the characteristics of the exponential distribution that the test is based on.

4.1 Tests based on the distribution function

Most classical tests employ the empirical distribution function (EDF) and are expressed in terms of distances between the EDF of the transformed residuals $Z_1, Z_2, \dots, Z_r$, given by

$$G_r(t) = \frac{1}{r} \sum_{j=1}^r I(Z_j \leq t), \quad (7)$$

and the corresponding population quantity for the standard exponential distribution, $G(t) = 1 - e^{-t}$. The following EDF-based test statistics are considered: The Kolmogorov-Smirnov (KS) test statistic

$$KS_r = \sup_{t \geq 0} |G_r(t) - G(t)|,$$

the Cramér-von Mises (CM) test statistic

$$CM_r = \int_0^\infty [G_r(t) - G(t)]^2 dG(t),$$

and the Anderson-Darling test (AD) statistic

$$AD_r = \int_0^\infty \frac{[G_r(t) - G(t)]^2}{[1 - G(t)]G(t)} dG(t).$$

There are computationally efficient formulae for the EDF-based statistics that may be found in D'Agostino and Stephens (1986). Specifically, the KS, the CM and the AD test statistics simplify to

$$KS_r = \max \{KS_r^+, KS_r^-\},$$

where

$$KS_r^+ = \max_{1 \leq j \leq r} \left[\frac{j}{r} - (1 - e^{-Z_{(j)}}) \right],$$

$$KS_r^- = \max_{1 \leq j \leq r} \left[(1 - e^{-Z_{(j)}}) - \frac{j-1}{r} \right],$$

$$CM_r = \frac{1}{12r} + \sum_{j=1}^r \left[(1 - e^{-Z_{(j)}}) - \frac{2j-1}{2r} \right]^2,$$

and

$$AD_r = -r - \sum_{j=1}^r \frac{2j-1}{r} [\ln(1 - e^{-Z_{(j)}}) - Z_{(r+1-j)}],$$

respectively, and where $Z_{(1)} \leq \dots \leq Z_{(r)}$ denote the ordered transformed residuals. Large values of these test statistics lead to the rejection of the null hypothesis.

4.2 Tests based on the characteristic function and Laplace transform

The tests discussed in this section employ either the characteristic function or the Laplace transform, and the test statistics are constructed by comparing the empirical version of the transform to their corresponding theoretical version.

Recall that the characteristic function is defined by $\phi(t) = E(e^{itX})$, $t \in \mathbb{R}$, $i = \sqrt{-1}$. We note that the characteristic function of the standard exponential distribution is given by $(1 - it)^{-1}$. The corresponding empirical characteristic function (ECF) is defined by

$$\phi_r(t) = \frac{1}{r} \sum_{j=1}^r e^{itZ_j} = C_r(t) + iS_r(t),$$

where $C_r(t)$ and $S_r(t)$ are the real part and imaginary part, respectively, of the $\phi_r(t)$. One of the test statistics used in this paper is given by Epps and Pulley (1986):

$$EP_r = \int_{-\infty}^{\infty} [\phi_r(t) - \phi(t)] \frac{1}{2\pi} \phi(-t) dt.$$

A convenient computational formula for this test statistic is

$$EP_r = \sqrt{48r} \left[\frac{1}{r} \sum_{j=1}^r e^{-Z_j} - \frac{1}{2} \right].$$

A second test statistic, proposed by Henze and Meintanis (2002), is given by

$$HM_{r,a} = r \int_0^{\infty} [S_r(t) - tC_r(t)]^2 e^{-at} dt, \quad a > 0,$$

with computational form

$$HM_{r,a} = \frac{a}{2r} \sum_{j=1}^r \sum_{k=1}^r \left[\frac{1}{a^2 + (Z_j - Z_k)^2} - \frac{1}{a^2 + (Z_j + Z_k)^2} - \frac{4(Z_j + Z_k)}{a^2 + (Z_j + Z_k)^2} \right. \\ \left. + \frac{2a^2 - 6(Z_j - Z_k)^2}{(a^2 + (Z_j - Z_k)^2)^3} + \frac{2a^2 - 6(Z_j + Z_k)^2}{(a^2 + (Z_j + Z_k)^2)^3} \right].$$

The second set of test statistics are based on comparison of the Laplace transform $\psi(t) = E(e^{-tX})$, $t > 0$, to the empirical Laplace transform (ELT), defined by

$$\psi_r(t) = \frac{1}{r} \sum_{j=1}^r e^{-tZ_j}.$$

Note that for a standard exponential random variable, the Laplace transform is given by $(1+t)^{-1}$. The first of the tests that we will consider based on the ELT is the one proposed by Baringhaus and Henze (1991)

$$BH_{r,a} = r \int_0^{\infty} [(1+t)\psi_r'(t) + \psi_r(t)]^2 e^{-at} dt,$$

and the second, more recent test, by Henze and Meintanis (2005), is given by

$$L_{r,a} = r \int_0^{\infty} [\psi_r(t) - \psi(t)]^2 (1+t)^2 e^{-at} dt.$$

Computational formulae that allow these test statistics to be easily calculable are given by

$$BH_{r,a} = \frac{1}{r} \sum_{j=1}^r \sum_{k=1}^r \left[\frac{(1-Z_j)(1-Z_k)}{Z_j + Z_k + a} - \frac{Z_j + Z_k}{(Z_j + Z_k + a)^2} + \frac{2Z_j Z_k}{(Z_j + Z_k + a)^2} + \frac{2Z_j Z_k}{(Z_j + Z_k + a)^3} \right]$$

and

$$L_{r,a} = \frac{1}{r} \sum_{j=1}^r \sum_{k=1}^r \left[\frac{1 + (Z_j + Z_k + a + 1)^2}{(Z_j + Z_k + a)^3} \right] - 2 \sum_{j=1}^r \left[\frac{1 + Z_j + a}{(Z_j + a)^2} \right] + \frac{r}{a},$$

respectively.

4.3 A moment-based test

In addition to tests based on the characteristic function and Laplace transform, we also consider a test based on a specific moment property of the exponential distribution. Specifically, the test named next is based on the following property identified by Lee, Locke, and Spurrier (1980), which states that if X is an exponentially distributed random variable, then

$$\frac{E(X^a)}{[E(X)]^a} = \Gamma(1+a), \quad \text{for } a > 0.$$

The property was then altered by Mimoto and Zitikis (2008), where it was noted that the following property also holds for exponential distributed random variables:

$$\frac{E[X^a]^{\frac{1}{a}}}{E(X)} = \Gamma(1+a)^{\frac{1}{a}}, \quad \text{for } -1 < a < 1, a \neq 0.$$

The resulting test statistic that employs this property produces the so-called Atkinson (AT) test,

$$AT_{r,a} = \sqrt{n} |R_{r,a} - \Gamma(1+a)^{\frac{1}{a}}|,$$

where

$$R_{r,a} = \left[\frac{1}{r} \sum_{j=1}^r Z_j^a \right]^{\frac{1}{a}}.$$

4.4 Mean residual life-based test

Another test, proposed by Baringhaus and Henze (2000), makes use of the idea that the exponential distribution is characterized by a constant ‘mean residual life’ (MRL), as shown in Shanbhag (1970). We define the MRL of a non-negative random variable X at time t as the expected value of the amount of lifetime remaining after time t , that is,

$$m(t) = E[X - t | X > t] = \frac{\int_t^\infty (1 - F(x)) dx}{1 - F(t)}.$$

In particular, this characterization for the exponential distribution with parameter λ can be expressed as

$$m(t) = E(X) = \frac{1}{\lambda}, \quad \forall t > 0, \tag{8}$$

which can be shown to be equivalent to

$$E(\min\{X, t\}) = \frac{F(t)}{\lambda}, \quad \forall t > 0. \tag{9}$$

The approach in Baringhaus and Henze (2000) uses a Cramér-von Mises type test statistic and the form of the characterization shown in (9), and is given by

$$M_r = r \int_0^\infty \left[\frac{1}{r} \sum_{j=1}^r \min\{Z_j, t\} - G_r(t) \right]^2 e^{-t} dt.$$

After some straightforward algebra, one finds the following easy to calculate form of the test statistic,

$$M_r = \frac{1}{r} \sum_{j=1}^r \sum_{k=1}^r \left[2 - 3 \exp(-\min\{Z_j, Z_k\}) - 2 \min\{Z_j, Z_k\} (e^{-Z_j} + e^{-Z_k}) + 2 \exp(-\max\{Z_j, Z_k\}) \right].$$

4.5 Score function-based test

A test based on the score function, defined as the gradient of the log-likelihood function, was developed by Cox and Oakes (1984) to test for exponentiality. Although this test was originally developed for censored data, it may readily be applied to a complete sample of size r by means of the formula

$$CO_r = r + \sum_{j=1}^r (1 - Z_j) \ln(Z_j).$$

4.6 Modified statistics

We conclude this section with three modified statistics, that is, statistics which are modifications of the corresponding full-sample versions, and may be applied directly to the observed Type-II right censored sample, without requiring transformation to a complete sample. The first two of these are the Cramér-von Mises and Anderson-Darling tests, which were modified by Pettitt and Stephens (1976). The modifications resulted in the following tractable forms for these statistics.

$$\widetilde{CM}_n = \sum_{j=1}^r \left(\widehat{\varepsilon}_{j:n} - \frac{2j-1}{2n} \right)^2 + \frac{r}{12n^2} + \frac{n}{3} \left(\widehat{\varepsilon}_{r:r} - \frac{r}{n} \right)^3 \quad (10)$$

and

$$\begin{aligned} \widetilde{AD}_n = & -\frac{1}{n} \sum_{j=1}^r (2j-1) \left[\log\{\widehat{\varepsilon}_{j:n}\} - \log\{1 - \widehat{\varepsilon}_{j:n}\} \right] - 2 \sum_{j=1}^r \log\{1 - \widehat{\varepsilon}_{j:n}\} \\ & - \frac{1}{n} \left[(r-n)^2 \log\{1 - \widehat{\varepsilon}_{r:r}\} - r^2 \log\{\widehat{\varepsilon}_{r:r}\} + n^2 \widehat{\varepsilon}_{r:r} \right] \end{aligned} \quad (11)$$

Finally, we also include the modified Cox and Oakes test appropriate for testing exponentiality with type-II censored samples,

$$\widetilde{CO}_n = r + \sum_{j=1}^r \ln(\widehat{\varepsilon}_j) - r \frac{\sum_{j=1}^n \widehat{\varepsilon}_j \ln(\widehat{\varepsilon}_j)}{\sum_{j=1}^n \widehat{\varepsilon}_j}. \quad (12)$$

Note that in the case of the AFT model, $\widehat{\varepsilon}_{j:n}$ is replaced by $\widehat{\varepsilon}_{j:n}^T$ (or by $\widehat{\varepsilon}_j^T$ if ordering is irrelevant) in (10), (11) and (12).

5 Simulation design and results

In this section we compare the finite-sample performance of all the tests for exponentiality discussed in Section 4 in the presence of Type-II right censoring. We evaluate the empirical powers of the tests for the following pairs of competing hypotheses (henceforth referred to as Scenarios A, B and C, respectively):

- (A) The null hypothesis $\mathcal{H}_0$ corresponds to the PH model with a certain baseline distribution against an alternative, say $\mathcal{H}_1$, with the same model, but different baseline distribution;
- (B) The null hypothesis corresponds to the PH model and the alternative is a Hsieh model with the same baseline distribution.
- (C) The null hypothesis corresponds to the AFT model with log-normal baseline distribution against an alternative with the same model, but different baseline distribution.

5.1 Simulation setting

All size and power estimates were calculated based on 10 000 independent Monte Carlo samples for sample sizes $n = 100$ and $n = 200$. The censoring proportions considered are $p = 10\%, 20\%, 30\%, 40\%$ and 50% . Each test, in which a tuning parameter a appears was evaluated for different values of a . Table 1 provides the distributions used for the baseline cumulative hazard function for the PH model and the distributions used for the error term in the AFT model (along with their parameterizations) and includes models with constant, decreasing, increasing, and non-monotone hazard rates. It should also be noted that the unknown parameters, appearing in each of the distributions presented in this simulation study, are estimated by means of maximum likelihood. Throughout the study we utilize a single continuous covariate, x_j , randomly and independently generated from a uniform (0,1) distribution, with covariate function: $r(x_j; \beta) = e^{\beta x_j}$, for $\beta = 0.1$. All calculations and simulations were performed with R version 3.2.4 and PH models were fit using the `flexsurvreg` function in the R package `flexsurv`.

5.2 Simulation results

Tables 2-10 and 17-18 show results under Scenario (A). In particular Tables 2 to 6 show the estimated powers of the various tests based on τ -transformations to complete sample, for the null hypothesis with an exponential baseline distribution against an alternative with a different baseline distribution. Tables 7 to 10 show the estimated powers for a Weibull baseline distribution against an alternative with a different baseline distribution, again for test statistics based on τ -transformations. We note incidentally that both the FK and OS transformations are reported in these tables (and in all subsequent tables with results based on transformations), but only the results from the OS transformation will be discussed in this section as the estimated power results from the FK transformation were consistently lower than those from the OS transformation. Tables 17 and 18, on the other hand, show the estimated powers based on the modified statistics (without transformation).

Next, Tables 11-14 and Table 19 display results under Scenario (B). Specifically, Tables 11 and 12 display the estimated powers using the Hsieh model as alternative with an exponential baseline, whereas Tables 13 and 14 use the same alternative model, but with a Weibull baseline, all for test statistics based on τ -transformations to complete sample. Corresponding results for the modified statistics using the Hsieh alternative with exponential and Weibull baselines are shown in Table 19. Finally in Tables 15-16 and Table 20 we report results under Scenario (C). Namely, Tables 15 and 16 correspond to transformation-based tests, and show the estimated powers for the null hypothesis AFT model with a log-normal baseline distribution against an alternative from the same

model, but different baseline distributions, while Table 20 displays the estimated powers for the AFT model when using the modified statistics.

The entries in all tables are the percentage of 10 000 independent Monte Carlo samples that were declared significant at nominal level 5%, rounded to the nearest integer. For ease of comparison, the highest power for each alternative has been highlighted in the reported tables. We will now discuss the performance of the tests for each scenario separately and then conclude with some overall findings.

Scenario (A): Exponential baseline hazard: The first set of results that we discuss (shown in Tables 2 to 6) relate to the size and power performance of EDF-criteria and other “traditional” test statistics for exponentiality using the exponential baseline hazard (this includes the KS, CM, AD, CO and M tests), that use τ -transformations to a complete sample.

The empirical sizes of the tests are reported in Table 2, where one can see that almost all of the test statistics maintain the size of the test, with the exception of the tests that use a tuning parameter a , namely AT, BH, L and HM. For these tests, one finds that only some choices of a produce sizes close to the nominal specified level of 5%, whereas other choices of this tuning parameter tend to underestimate the size.

When considering the power performance of the traditional tests across the various alternatives investigated in this study, one finds that there is no single test statistic dominating all others in all cases. For example, for the Weibull alternative in Table 6, the CO test achieved the highest estimated powers, but was the worst performing test for every other alternative considered. On the other hand, the AD test performs best for only one of the alternatives (the gamma alternative, shown in Table 3), but fails to attain adequate powers for other alternatives relative to the other tests. Our finding, which in fact persists in all other instances considered in this study, is not surprising as it is well known that each nonparametric test attains competitive power only along a specific direction away from the null hypothesis; see Janssen (2000). This fact notwithstanding and although there is no overall “best” test in this group, it can be argued that the CM and M tests are “second best” in that, while they do not always achieve the best results for all alternatives, these tests always have powers that are very close to the best performing tests. Note, however, that the CM test displays better results for higher censoring proportion p , whereas the M test performs better for lower values of p . It should also be noted that the M test tends to have the largest drop in power over the censoring proportions. In addition, one can see that the KS statistic also produces reasonably good results and can be regarded as a competitive option for testing exponentiality against these alternatives. When one compares the magnitude of the powers for the various alternatives, one finds that under gamma and log-normal alternatives (Tables 3 and 5) we observe relatively higher powers when compared to the respective power under Gompertz and Weibull alternatives (Tables 4 and 6) for all test statistics and censoring proportions used. However, all tests have difficulty distinguishing between the null and alternative for $p = 50\%$ censoring proportion. For the remaining statistics based on a τ -transformation (namely, the EP, AT, BH, L and HM tests), one once again finds that no statistic is dominant for the exponential baseline hazard across all alternatives. The EP test produces the highest estimated powers against Gompertz and Weibull alternatives and achieves estimated powers that are very close to the best performing test for each of the other alternatives considered. Although the L and HM test statistics do not produce the highest estimated powers, they can still be regarded as competitive to the EP test statistic. At the other end, the powers of the AT test statistic fare poorly when compared to those of the other statistics, as they are the lowest empirical powers for all alternatives considered. It is interesting to note that all test statistics in this set-up also attain rather low estimated powers for $p = 40\%$ and $p = 50\%$ censoring proportions, except when testing against a gamma alternative, where these powers are more robust with respect to p (see Table 3). We now turn to modified test statistics for which corresponding results are displayed in Table 17. Here too there is no clear-cut best choice overall, since all three tests show comparable results for lower censoring proportions

at each alternative. However, the $\widetilde{CO}$ test does appear to retain the size of the test better than either $\widetilde{CM}$ or $\widetilde{AD}$. Nevertheless, the rate at which the powers decay across censoring proportions is notably more pronounced for the $\widetilde{CM}$ statistic, whereas the $\widetilde{CO}$ statistic is more stable in this regard. It is for these reasons that one might prefer the $\widetilde{CO}$ test in this case as it is more often the better performing statistic for larger censoring proportions. Additionally and similar to all other statistics discussed so far, the empirical power of the tests based on the modified statistics degrade rapidly with increasing censoring proportions. This type of behavior however is less pronounced for tests based on transformation to a complete sample.

Scenario (A): Weibull baseline hazard: For the Weibull baseline hazard under the null hypothesis (Tables 7 to 10) one finds that, when comparing the “traditional” tests to one another, the CM and M statistics show again good overall performance, with the M test performing better for lower censoring proportions and the CM test better for higher censoring proportions. The estimated sizes reported in Table 7 display the same general features as those for the tests employing the exponential baseline hazards discussed earlier. For the majority of alternatives, the AD statistic now produces reasonably good results whereas the CO statistic, in contrast, appears to consistently have the lowest estimated powers among all the tests and alternative distributions. When one compares the magnitude of the estimated powers for the various alternatives, one finds that only the log-normal alternative produces moderately high powers for the KS, CM, AD, and M tests, as seen in Table 10, while the remaining alternatives all have much lower power results. Finally, the M statistic once again is reported to have the least robust power values across censoring proportions.

As with the previous set-up using the exponential baseline hazard, when one compares the tests based on transformation to complete sample that do not use the EDF, one again finds that the EP statistic generally performs the best when using the Weibull baseline hazard and the L statistic can be considered to be a decent runner-up candidate. The worst performing statistic is once again the AT statistic, with powers routinely below the nominal significance level, even in the case of the log-normal alternative, where one generally observes the highest powers for this set-up.

In the case of the modified statistics, the $\widetilde{AD}$ and $\widetilde{CM}$ statistics outperform the $\widetilde{CO}$ test, which however does appear to produce sizes closer to the specified 5% nominal level.

Finally when comparing the modified statistics $\widetilde{CM}$, $\widetilde{AD}$, and $\widetilde{CO}$ to their counterparts based on transformations to complete samples, i.e., CM, AD and CO, similar powers are observed for lower censoring proportions. However as also reported earlier, the modified statistics tests have the tendency to lose power much more rapidly as the censoring proportion p increases.

Scenario (B): Exponential baseline hazard: For the first set of results comparing the PH model to the Hsieh model, the null hypothesis corresponds to the PH model with exponential baseline and the Hsieh model alternative has parameter $\gamma = 0.25$ (see Table 11). Looking at the results for the “traditional” tests we find that the CO test statistic achieved the highest estimated powers, with the M and CM tests being second best. For $\gamma = 0.50$ (Table 12), the M test displays the highest estimated powers for low censoring proportions whereas the CM test performs better for higher censoring proportions. When considering the performance across the different censoring proportions, the AD test displays the largest drop in power. For the remaining test statistics based on a τ -transformation, the L test is the dominant test, producing the highest estimated powers for $\gamma = 0.25$ and 0.50 . Once again AT is the worst performing test.

For the modified test statistics in Table 19, $\widetilde{CO}$ outperforms $\widetilde{CM}$ and $\widetilde{AD}$, with the $\widetilde{CM}$ test displaying the greatest drop in power over the censoring proportions.

When one compares the power performance between the modified versions of the statistics to the versions based on τ -transformations to complete samples, one notes that $\widetilde{CO}$ performs much

better than its counterpart. The other two modified tests, however, tend to have very similar performance to the corresponding transformation-based tests for lower censoring proportions, but degrade in performance when the censoring proportion p increases. In addition, it should be noted that the $\widetilde{CO}$ statistic does not suffer from the same rate of degradation as observed for either $\widetilde{CM}$ or $\widetilde{AD}$.

Scenario (B): **Weibull baseline hazard:** When considering the Hsieh model alternative with Weibull baseline and $\gamma = 0.5$ in Table 13, one finds that all transformation-based tests considered produced low estimated powers. Similarly, the estimated power results for the modified statistics, shown in Table 19, also display dismally low powers. However, for $\gamma = 1$, somewhat larger powers are observed for the Hsieh model alternative with Weibull baseline, with the best performing tests being the tests based on the L and EP statistics. It is interesting to note that the more modern tests tend to outperform the traditional EDF-based statistics in this setting.

Scenario (C): In Table 15 the estimated powers for the AFT model with a log-normal cumulative hazard rate are compared to an alternative that also uses the AFT model, but with a log-logistic cumulative hazard rate. In this table we find that the EDF-based and the M tests display similar powers, with the AD test producing the highest estimated powers. For the Weibull alternative (Table 16) the M test produces the highest estimated powers closely followed by the CM and AD tests. For the remaining transformation-based tests, the L test has the largest estimated powers for both alternatives considered. As with the results from the previous scenarios, AT produces the lowest estimated power relative to the other tests.

In the case of the modified tests, the $\widetilde{CO}$ test performs best, with all three tests producing high estimated powers for the Weibull alternative.

Turning attention to the differences between the modified statistics and statistics based on data transformed to complete samples, one finds that $\widetilde{CO}$ outperforms CO, whereas $\widetilde{CM}$ and $\widetilde{AD}$ are relatively similar to their counterparts for lower censoring proportions. In addition, the $\widetilde{CM}$ and $\widetilde{AD}$ statistics show a high rate of decreasing powers with increasing censoring proportions, whereas this behavior is not present in $\widetilde{CO}$.

6 Conclusion

In this paper we considered goodness-of-fit tests for PH and AFT models in the presence of Type-II right censoring by means of testing exponentiality of corresponding residuals. In doing so we investigated the behavior of a number of exponentiality tests in a variety of settings. In particular, interest lay in the comparison of the two different strategies that are often employed to conduct these tests, namely (i) first transforming the censored sample of size $r < n$ to a complete sample of size r and then using full sample versions of the test statistics, and (ii) the alternative where the statistics themselves are modified to accommodate Type-II right censoring. The test statistics considered under strategy (i) include tests based on the EDF, score function, characteristic function, Laplace transform, a criterion based on a moment-type characterization of the exponential distribution, as well as a test based on the mean residual life. The modified statistics (strategy (ii)) include the Anderson-Darling and Cramér-von Mises EDF-based tests, as well as the corresponding modified test based on a score function. The performance of these test statistics when testing exponentiality was evaluated in a comprehensive simulation study in three distinct scenarios: (A) being able to distinguish between a number of alternative baseline hazards when using the PH model, (B) being able to distinguish between the PH and Hsieh models with the same baseline hazard, and (C) the ability to distinguish between the AFT model with a given baseline hazard against the same model with a different hazard rate.

Based on the results of the Monte Carlo study conducted in this paper, it was found that no single test dominated all other tests for all scenarios considered, but that the Epps and Pulley test based on the characteristic function, and the test of Henze and Meintanis based on the Laplace transform generally outperformed the other tests for *many* of the settings considered, while the test of Baringhaus and Henze based on the mean residual life is also consistently a good performer.

Two transformations to complete samples were considered in this paper, namely those suggested by O'Reilly and Stephens (1988) and by Fischer and Kamps (2011). In the finite sample study it was found that the former transformation consistently produced higher powers than the latter for all scenarios investigated. Hence it seems reasonable to recommend the use of the O'Reilly and Stephens (1988) transformation if one opts for strategy (i) above.

Last we present suggestions based on the performance of strategies (i) and (ii). With lower censoring proportions the difference in powers between the tests that make use of statistics modified to accommodate Type-II right censoring and the tests that transform the censored sample to a complete sample is not all that great. However, this difference becomes much more pronounced with higher censoring proportions. Specifically, the empirical powers for tests employing modified statistics display a rapid decline as the censoring proportions increase and consequently these tests have much lower powers for higher censoring proportions than the tests using transformations to complete samples. Therefore either of the two approaches is useful for conducting goodness-of-fit testing for exponentiality if the censoring proportion in a data set is lower than 30%, but we would recommend using the approach that first transforms the sample to a complete sample for higher censoring proportions.

References

- Allison, J., Santana, L., Smit, N., & Visagie, I. (2017). An ‘apples to apples’ comparison of various tests for exponentiality. *Computational Statistics*, 32(4), 1241-1283. doi: 10.1007/s00180-017-0733-3
- Baringhaus, L., & Henze, N. (1991). A class of consistent tests for exponentiality based on the empirical Laplace transform. *Annals of the Institute of Statistical Mathematics*, 43(3), 551-564.
- Baringhaus, L., & Henze, N. (2000). Tests of fit for exponentiality based on a characterization via the mean residual life function. *Statistical Papers*, 41(2), 225-236. doi: 10.1007/BF02926105
- Caroni, C. (2004). Diagnostics for Cox’s proportional hazards model. In Nikulin, M.S., Balakrishnan, N., Mesbah M., & Limnios, N. (Eds.), *Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life* (pp. 27-38). Boston, MA: Birkhäuser.
- Cockeran, M., Meintanis, S., & Allison, J. (2019). Goodness-of-fit tests in the Cox proportional hazards model. *Communications in Statistics-Simulation and Computation*, 1-12. doi: 10.1080/03610918.2019.1639738
- Cox, D. R., & Oakes, D. (1984). *Analysis of Survival Data*. New York: Chapman and Hall.
- D’Agostino, R. B., & Stephens, M. (1986). *Goodness-of-fit Techniques, Series Statistics: Textbooks and Monographs*, 68. New York, NY: Marcel Dekker.
- Epps, T., & Pulley, L. (1986). A test of exponentiality vs. monotone-hazard alternatives derived from the empirical characteristic function. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(2), 206-213. doi: 10.1111/j.2517-6161.1986.tb01403.x
- Fischer, T., & Kamps, U. (2011). On the existence of transformations preserving the structure of order statistics in lower dimensions. *Journal of Statistical Planning and Inference*, 141(1), 536-548. doi: 10.1016/j.jspi.2010.06.028
- Giacomini, R., Politis, D. N., & White, H. (2013). A warp-speed method for conducting Monte Carlo experiments involving bootstrap estimators. *Econometric Theory*, 29(3), 567-589. doi: 10.1017/S0266466612000655
- Goldmann, C., Klar, B., & Meintanis, S. G. (2015). Data transformations and goodness-of-fit tests for type-II right censored samples. *Metrika*, 78(1), 59-83. doi: 10.1007/s00184-014-0490-z
- Henze, N., & Meintanis, S. G. (2002). Goodness-of-fit tests based on a new characterization of the exponential distribution. *Communications in Statistics-Theory and Methods*, 31(9), 1479-1497. doi: 10.1081/STA-120013007
- Henze, N., & Meintanis, S. G. (2005). Recent and classical tests for exponentiality: a partial review with comparisons. *Metrika*, 61(1), 29-45. doi: 10.1007/s001840400322
- Hsieh, F., & Lavori, P. W. (2000). Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. *Controlled Clinical Trials*, 21(6), 552-560. doi:

10.1016/S0197-2456(00)00104-5

- Janssen, A. (2000). Global power functions of goodness-of-fit tests. *Annals of Statistics*, 28(1), 239-253.
- Klein, J. P., & Moeschberger, M. L. (2006). *Survival Analysis: Techniques for Censored and Truncated Data*. New York, NY: Springer.
- Lee, S., Locke, C., & Spurrier, J. D. (1980). On a class of tests of exponentiality. *Technometrics*, 22(4), 547-554. doi: 10.1080/00401706.1980.10486204
- Lin, C.-T., Huang, Y.-L., & Balakrishnan, N. (2008). A new method for goodness-of-fit testing based on type-II right censored samples. *IEEE Transactions on Reliability*, 57(4), 633-642. doi: 10.1109/TR.2008.2005860
- Michael, J. R., & Schucany, W. R. (1979). A new approach to testing goodness of fit for censored samples. *Technometrics*, 21(4), 435-441.
- Mimoto, N., & Zitikis, R. (2008). The Atkinson index, the Moran statistic, and testing exponentiality. *Journal of the Japan Statistical Society*, 38(2), 187-205. doi: 10.14490/jjss.38.187
- Novak, P. (2010). Checking goodness-of-fit of the accelerated failure time model for survival data. *WDS'10 Proceedings of Contributed Papers, I*, 189-194.
- O'Reilly, F., & Stephens, M. (1988). Transforming censored samples for testing fit. *Technometrics*, 30(1), 79-86. doi: 10.1080/00401706.1988.10488325
- Pettitt, A. N., & Stephens, M. A. (1976). Modified Cramér-von Mises statistics for censored data. *Biometrika*, 63(2), 291-298.
- Qi, J. (2009). *Comparison of proportional hazards and accelerated failure time models* (Unpublished master's thesis). Department of Mathematics and Statistics, Saskatoon, Canada.
- Shanbhag, D. (1970). The characterizations for exponential and geometric distributions. *Journal of the American Statistical Association*, 65(331), 1256-1259. doi: 10.1080/01621459.1970.10481161

Table 1 Densities of baseline distributions.

Distribution	Density function	Parameter values	Hazard	Model
Exponential	$f(t) = \lambda e^{-\lambda t}$	$\lambda = 0.05$	Constant	PH
Weibull	$f(t) = \frac{a}{b} \left(\frac{t}{b}\right)^{a-1} e^{-\left(\frac{t}{b}\right)^a}$	$a = 1.1 \ b = 0.8$	Increasing	PH and AFT
Log-normal	$f(t) = \frac{1}{\sqrt{2\pi}\sigma t} e^{-\frac{(\log t - \mu)^2}{2\sigma^2}}$	$\mu = 1.3 \ \sigma = 1.5$	Non-Monotone	PH
Gompertz	$f(t) = be^{at} e^{-\frac{b}{a(e^{at}-1)}}$	$a = 1.1 \ b = 2.0$	Increasing	PH
Gamma	$f(t) = \frac{1}{s^a \Gamma(a)} x^{a-1} e^{-\frac{t}{s}}$	$a = 0.5 \ s = 1.1$	Decreasing	PH
Log-logistic	$f(t) = \frac{\gamma(t/\theta)^\gamma}{t[1+(t/\theta)^\gamma]^2}$	$\gamma = 0 \ \theta = 1$	Non-monotone	AFT

Table 2 Empirical rejection rates for Scenario A with exponential hazard both under $\mathcal{H}_0$ and $\mathcal{H}_1$.

n	τ	p	KS	CM	AD	EP	CO	M	AT _{0.01}	AT _{0.99}	BH _{0.1}	BH _{2.5}	BH _{5.0}	L _{0.1}	L _{2.5}	L _{5.0}	HM _{0.1}	HM _{2.5}	HM _{5.0}
100	OS	10%	5	5	4	5	4	5	4	2	5	3	2	5	5	4	5	5	4
		20%	5	5	4	5	4	5	3	2	4	2	1	5	4	3	5	5	4
		30%	5	5	3	4	3	4	2	1	4	1	1	5	4	2	5	4	2
		40%	5	4	2	4	3	4	2	1	4	1	1	5	3	1	5	4	2
		50%	4	4	2	4	4	3	2	1	4	1	1	5	2	1	5	5	2
	FK	10%	5	4	4	4	4	4	4	2	4	4	3	5	4	4	4	4	4
		20%	4	4	3	3	4	3	4	3	3	2	2	5	3	3	5	4	3
		30%	3	3	2	3	4	3	3	2	3	2	2	5	3	2	5	3	2
		40%	3	2	2	2	3	2	3	2	2	2	2	5	2	2	5	3	2
		50%	3	2	2	2	3	2	4	3	2	2	2	5	2	2	5	3	2
200	OS	10%	5	5	5	5	4	5	4	2	4	3	2	4	4	4	5	5	4
		20%	5	5	4	5	4	5	3	1	5	2	1	5	5	3	5	5	3
		30%	5	5	4	5	4	5	2	1	5	1	1	5	4	2	6	5	3
		40%	5	5	3	4	3	4	2	1	4	1	1	5	4	1	5	5	2
		50%	5	5	2	4	4	4	2	1	4	1	1	5	3	2	5	4	2
	FK	10%	4	4	3	4	4	4	3	2	4	3	3	5	4	3	5	4	4
		20%	4	4	3	3	4	3	3	2	3	2	2	5	3	2	5	4	3
		30%	3	3	2	3	4	3	3	2	2	2	2	4	3	2	5	3	3
		40%	3	3	2	3	4	2	3	2	2	2	2	4	2	2	5	3	2
		50%	3	3	2	3	4	2	3	3	2	2	2	5	2	2	4	3	2

Table 3 Empirical rejection rates for Scenario A where $\mathcal{H}_0$ is exponential and $\mathcal{H}_1$ is gamma.

n	τ	p	KS	CM	AD	EP	CO	M	$AT_{0.01}$	$AT_{0.99}$	$BH_{0.1}$	$BH_{2.5}$	$BH_{5.0}$	$L_{0.1}$	$L_{2.5}$	$L_{5.0}$	$HM_{0.1}$	$HM_{2.5}$	$HM_{5.0}$
100	OS	10%	99	99	100	99	99	99	100	96	100	99	94	100	99	97	92	96	92
		20%	98	99	100	98	89	98	100	82	100	95	68	100	98	93	91	95	88
		30%	96	98	99	97	70	97	99	69	100	86	37	100	97	83	87	93	82
		40%	95	97	99	94	49	95	99	59	99	67	13	99	94	65	83	90	74
		50%	89	93	96	89	32	88	97	50	98	41	4	99	87	40	77	86	60
	FK	10%	39	40	64	21	49	35	89	10	77	34	35	85	26	33	51	29	36
		20%	33	32	51	15	29	27	86	9	67	24	25	82	17	23	48	24	28
		30%	26	24	37	11	23	20	83	8	57	17	17	80	12	16	42	19	20
		40%	21	18	28	7	18	15	79	7	47	13	13	76	9	12	38	14	15
		50%	15	13	19	5	13	10	72	7	36	9	9	69	6	9	31	11	11
200	OS	10%	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
		20%	100	100	100	100	100	100	100	97	100	100	98	100	100	100	100	100	99
		30%	100	100	100	100	96	100	100	87	100	99	86	100	100	99	99	100	98
		40%	100	100	100	100	82	100	100	76	100	97	57	100	100	97	98	99	96
		50%	100	100	100	100	63	100	100	67	100	91	30	100	100	92	97	99	93
	FK	10%	63	68	90	38	68	59	99	9	97	54	55	98	44	53	87	51	58
		20%	57	60	82	31	48	51	98	7	94	42	42	98	36	42	84	46	50
		30%	50	50	69	22	31	39	96	6	88	30	29	97	26	30	79	36	39
		40%	41	40	55	17	24	30	94	6	80	22	21	95	19	23	72	29	30
		50%	32	30	41	11	19	22	91	6	69	16	16	92	13	16	63	23	22

Table 4 Empirical rejection rates for Scenario A where $\mathcal{H}_0$ is exponential and $\mathcal{H}_1$ is Gompertz.

n	τ	p	KS	CM	AD	EP	CO	M	AT _{0,01}	AT _{0,99}	BH _{0,1}	BH _{2,5}	BH _{5,0}	L _{0,1}	L _{2,5}	L _{5,0}	HM _{0,1}	HM _{2,5}	HM _{5,0}
100	OS	10%	34	42	37	50	39	48	26	3	26	38	20	12	47	43	4	47	33
		20%	25	29	21	31	25	30	11	0	17	12	3	9	30	20	4	27	11
		30%	15	18	10	20	16	18	4	0	10	3	2	7	17	8	5	16	3
		40%	12	12	5	13	10	11	2	1	7	2	1	6	11	4	5	9	1
		50%	8	8	2	9	9	6	1	1	5	1	1	5	6	3	5	6	1
			12	12	9	10	9	10	5	1	10	6	5	7	9	7	9	11	9
	FK	20%	6	5	3	4	8	4	3	2	4	2	1	6	3	2	8	6	3
		30%	4	3	2	3	6	2	2	2	3	1	1	6	2	2	7	4	2
		40%	3	2	1	2	5	2	2	2	2	1	1	6	2	1	6	3	2
		50%	2	2	1	2	4	2	3	2	2	1	1	5	2	2	6	2	2
200	OS	10%	61	72	69	80	67	79	43	6	51	73	54	22	78	79	6	79	69
		20%	40	48	40	54	42	53	16	0	31	29	6	16	53	43	5	51	30
		30%	25	30	21	36	28	33	5	0	19	8	2	11	33	20	5	30	9
		40%	18	21	10	23	17	21	2	0	12	2	1	8	21	8	5	17	3
		50%	12	12	5	14	12	12	1	0	7	2	1	6	12	5	5	8	1
	FK	10%	18	20	15	15	14	17	5	1	16	10	8	10	14	12	11	19	17
		20%	8	8	5	6	10	6	2	1	6	2	1	9	4	3	9	9	5
		30%	5	4	2	3	8	2	1	1	3	1	1	7	2	1	7	5	2
		40%	3	3	1	2	6	2	1	1	2	1	1	6	2	1	7	4	2
		50%	2	2	1	2	6	2	2	2	2	1	1	6	2	1	6	3	2

Table 5 Empirical rejection rates for Scenario A where $\mathcal{H}_0$ is exponential and $\mathcal{H}_1$ is log-normal.

n	τ	p	KS	CM	AD	EP	CO	M	AT _{0,01}	AT _{0,99}	BH _{0,1}	BH _{2,5}	BH _{5,0}	L _{0,1}	L _{2,5}	L _{5,0}	HM _{0,1}	HM _{2,5}	HM _{5,0}
100	OS	10%	87	91	90	93	61	93	70	85	83	91	85	38	92	91	13	92	92
		20%	65	71	66	72	14	73	33	41	52	56	30	13	69	58	11	73	73
		30%	36	41	34	40	4	42	12	17	23	17	4	4	34	19	10	46	41
		40%	18	21	13	18	2	19	4	8	8	3	1	2	13	4	10	24	19
		50%	9	9	4	7	4	6	2	5	3	1	1	3	4	1	9	12	7
	FK	10%	25	24	46	7	39	20	57	25	54	27	30	50	12	21	14	18	25
		20%	13	12	20	5	13	11	33	13	24	13	14	28	7	11	8	10	14
		30%	7	6	9	4	7	5	19	10	11	6	6	17	4	5	5	6	6
		40%	4	3	4	2	4	3	11	7	5	4	4	10	3	3	4	3	4
		50%	3	2	3	2	3	2	7	6	3	2	3	7	2	2	4	2	3
200	OS	10%	99	100	100	100	89	99	94	98	98	99	99	65	99	99	26	99	99
		20%	90	94	93	94	25	95	58	62	84	88	69	22	93	89	19	95	94
		30%	65	71	65	71	5	73	17	20	50	43	12	5	65	51	18	75	72
		40%	34	39	28	33	2	37	4	8	18	7	1	5	27	11	15	42	37
		50%	15	15	8	10	3	12	1	4	7	1	1	10	7	2	14	16	14
	FK	10%	48	53	75	12	53	41	83	28	84	44	47	79	22	37	37	36	46
		20%	28	27	38	9	18	22	50	12	49	22	24	48	13	20	16	20	26
		30%	15	12	16	5	6	10	26	7	21	10	10	26	7	9	9	10	13
		40%	7	6	6	3	4	5	14	5	9	4	5	14	4	4	5	5	6
		50%	4	3	3	2	3	3	8	5	4	3	3	9	2	2	4	3	3

Table 6 Empirical rejection rates for Scenario A where $\mathcal{H}_0$ is exponential and $\mathcal{H}_1$ is Weibull.

n	τ	p	KS	CM	AD	EP	CO	M	AT _{0,01}	AT _{0,99}	BH _{0,1}	BH _{2,5}	BH _{5,0}	L _{0,1}	L _{2,5}	L _{5,0}	HM _{0,1}	HM _{2,5}	HM _{5,0}
100	OS	10%	14	16	14	19	20	17	14	1	14	12	4	10	19	14	5	14	7
		20%	14	15	11	16	17	15	7	0	12	4	1	9	15	9	5	12	3
		30%	10	12	7	12	14	11	4	0	9	2	1	8	11	4	5	9	2
		40%	10	11	5	11	11	9	2	1	7	1	1	6	10	3	5	7	1
		50%	8	9	3	9	11	6	2	1	6	1	1	6	7	2	5	6	1
			6	5	4	5	5	4	3	2	5	3	2	6	4	3	6	5	4
	FK	20%	4	3	2	3	6	2	3	2	3	1	1	6	2	2	6	4	2
		30%	3	2	2	2	5	2	3	2	2	1	1	7	2	1	6	3	2
		40%	3	2	1	2	5	2	3	3	2	1	1	6	2	1	6	3	2
		50%	2	2	2	2	4	2	3	3	2	1	1	5	2	2	6	2	2
			23	28	28	31	33	29	21	0	28	24	11	20	31	27	6	25	14
		20%	19	23	20	25	26	23	11	0	23	9	2	18	25	16	6	20	7
200	OS	30%	16	18	13	20	22	18	4	0	18	3	1	15	19	10	6	14	3
		40%	15	18	10	18	18	16	2	0	15	2	1	12	17	5	5	12	1
		50%	12	13	5	14	15	11	1	1	10	1	1	10	12	4	5	8	1
		10%	6	6	4	5	6	5	3	2	5	3	2	7	4	3	7	7	5
		20%	4	4	2	3	5	3	2	2	3	1	1	7	2	1	7	5	2
		30%	4	3	2	3	6	2	2	2	2	1	1	7	2	1	6	4	2
	FK	40%	3	2	1	2	5	2	2	2	2	1	1	6	2	1	6	3	1
		50%	2	2	1	2	6	1	2	2	2	1	1	7	1	1	5	3	1
			23	28	28	31	33	29	21	0	28	24	11	20	31	27	6	25	14
		20%	19	23	20	25	26	23	11	0	23	9	2	18	25	16	6	20	7
		30%	16	18	13	20	22	18	4	0	18	3	1	15	19	10	6	14	3
		40%	15	18	10	18	18	16	2	0	15	2	1	12	17	5	5	12	1

Table 7 Empirical rejection rates for Scenario A with Weibull hazard both under $\mathcal{H}_0$ and $\mathcal{H}_1$.

n	τ	p	KS	CM	AD	EP	CO	M	AT _{0,01}	AT _{0,99}	BH _{0,1}	BH _{2,5}	BH _{5,0}	L _{0,1}	L _{2,5}	L _{5,0}	HM _{0,1}	HM _{2,5}	HM _{5,0}
100	OS	10%	5	5	4	5	3	5	1	1	5	1	1	5	4	2	5	5	4
		20%	5	5	2	5	3	4	1	1	4	1	1	5	2	1	5	4	3
		30%	5	5	2	4	4	3	1	1	3	1	1	4	2	1	5	3	2
		40%	4	4	1	3	3	2	1	1	2	1	1	5	1	1	5	3	2
		50%	4	3	1	3	3	2	1	1	1	1	1	5	1	1	5	3	1
	FK	10%	5	4	4	5	4	4	3	2	4	3	3	5	4	3	5	5	4
		20%	3	3	2	3	4	2	3	3	3	2	2	5	2	2	4	3	2
		30%	3	3	2	3	4	2	3	2	3	2	2	5	2	2	5	3	2
		40%	2	2	2	2	3	2	3	2	2	1	2	4	2	2	5	3	2
		50%	2	2	1	2	4	2	4	3	2	1	1	5	2	2	5	3	2
200	OS	10%	5	5	4	5	3	5	1	1	5	1	1	5	4	2	5	4	3
		20%	5	5	3	5	4	4	1	1	4	1	1	5	3	1	5	4	2
		30%	5	5	2	5	4	4	1	1	3	1	1	5	2	1	5	4	3
		40%	5	5	1	4	4	3	1	1	2	1	1	5	1	1	5	3	2
		50%	4	4	1	3	4	2	1	1	2	1	1	5	1	1	5	3	2
	FK	10%	4	4	3	4	4	4	3	2	3	3	3	5	4	3	5	5	4
		20%	4	3	3	3	4	3	3	2	3	2	2	5	3	2	5	4	3
		30%	4	3	2	3	4	2	3	2	3	2	2	5	2	2	5	4	2
		40%	3	2	2	2	4	2	3	2	2	2	2	5	2	2	5	4	2
		50%	3	2	2	2	4	2	3	3	2	2	2	4	2	2	4	3	2

Table 8 Empirical rejection rates for Scenario A where $\mathcal{H}_0$ is Weibull and $\mathcal{H}_1$ is gamma.

n	τ	p	KS	CM	AD	EP	CO	M	AT _{0,01}	AT _{0,99}	BH _{0,1}	BH _{2,5}	BH _{5,0}	L _{0,1}	L _{2,5}	L _{5,0}	HM _{0,1}	HM _{2,5}	HM _{5,0}
100	OS	10%	11	12	11	19	2	15	2	0	16	3	1	16	17	9	6	12	5
		20%	10	10	5	15	1	11	1	1	11	1	1	14	10	3	6	7	2
		30%	8	8	2	10	4	6	1	1	6	1	1	10	4	2	5	4	1
		40%	6	6	1	8	3	4	1	1	4	1	1	9	3	2	5	3	1
		50%	5	4	1	6	4	3	1	1	2	1	1	8	2	1	5	3	1
	FK	10%	6	5	3	4	6	4	3	1	4	2	2	4	3	2	7	4	3
		20%	4	3	2	3	6	2	3	2	2	2	1	4	2	2	6	3	2
		30%	3	3	2	3	4	2	3	2	3	2	2	4	2	2	6	3	2
		40%	3	2	2	3	4	2	2	2	2	2	2	4	2	2	6	3	2
		50%	2	2	2	2	4	2	3	2	2	2	1	4	2	2	5	3	2
200	OS	10%	17	21	21	34	4	27	2	0	27	6	2	29	32	21	8	23	13
		20%	14	16	10	24	6	18	1	0	18	2	1	21	19	7	6	13	3
		30%	11	11	5	19	6	12	1	0	11	1	1	16	10	4	6	8	2
		40%	8	9	2	12	6	8	1	1	7	1	1	12	5	2	5	4	1
		50%	7	6	1	8	5	4	1	1	4	1	1	10	3	2	6	3	1
	FK	10%	6	5	3	4	7	4	3	1	3	2	2	4	3	2	7	5	4
		20%	4	3	2	3	6	3	2	1	3	2	1	4	3	2	6	4	2
		30%	4	3	2	3	5	2	2	1	3	2	2	4	2	2	6	3	2
		40%	3	3	2	3	5	2	2	1	3	2	1	4	2	2	6	3	2
		50%	2	2	2	2	5	2	2	2	2	1	1	4	2	2	6	3	2

Table 9 Empirical rejection rates for Scenario A where $\mathcal{H}_0$ is Weibull and $\mathcal{H}_1$ is Gompertz.

n	τ	p	KS	CM	AD	EP	CO	M	AT _{0,01}	AT _{0,99}	BH _{0,1}	BH _{2,5}	BH _{5,0}	L _{0,1}	L _{2,5}	L _{5,0}	HM _{0,1}	HM _{2,5}	HM _{5,0}
100	OS	10%	10	11	10	16	2	13	2	0	14	2	1	14	15	8	6	11	5
		20%	9	9	5	13	1	9	1	1	10	1	1	12	8	3	6	6	2
		30%	7	7	2	8	4	5	1	1	6	1	1	9	3	1	5	3	1
		40%	5	5	1	6	3	3	1	1	5	1	1	8	3	1	5	3	1
		50%	4	4	1	5	4	2	1	1	3	1	1	7	2	1	5	2	1
			6	5	3	4	6	4	3	1	4	2	2	4	4	3	7	5	4
	FK	10%	4	3	2	3	6	2	3	2	3	2	2	4	2	2	6	3	2
		20%	3	3	2	3	4	3	3	2	3	2	2	4	3	2	6	3	2
		30%	3	3	2	3	4	2	2	2	2	2	2	4	2	2	6	3	2
		40%	2	2	2	2	4	2	2	2	2	2	2	4	2	2	5	3	2
		50%	15	19	19	29	4	25	2	0	24	5	2	25	27	19	7	20	13
			12	13	8	19	6	15	1	0	16	2	1	17	15	6	6	11	3
200	OS	10%	9	9	4	14	6	9	1	1	10	1	1	12	8	3	5	6	1
		20%	7	7	2	9	5	6	1	1	7	1	1	10	4	2	5	3	1
		30%	6	5	1	6	5	3	1	1	5	1	1	8	2	1	5	3	1
		40%	6	4	3	4	7	4	2	1	3	2	2	4	3	3	7	5	4
		50%	4	3	2	3	6	3	2	1	3	2	1	4	3	2	6	4	3
			4	3	2	3	5	2	2	1	3	2	2	4	3	2	6	4	2
	FK	10%	3	3	2	3	4	2	2	1	3	2	2	4	3	2	6	3	2
		20%	2	2	2	3	4	2	2	2	2	2	1	4	2	2	5	3	2
		30%	2	2	2	3	4	2	2	2	2	2	2	4	2	2	5	3	2
		40%	2	2	2	3	4	2	2	2	2	2	2	4	2	2	5	3	2
		50%	6	4	3	4	7	4	2	1	3	2	2	4	3	2	7	5	4
			4	3	2	3	6	3	2	1	3	2	1	4	3	2	6	4	3

Table 10 Empirical rejection rates for Scenario A where $\mathcal{H}_0$ is Weibull and $\mathcal{H}_1$ is Gompertz.

n	τ	p	KS	CM	AD	EP	CO	M	AT _{0,01}	AT _{0,99}	BH _{0,1}	BH _{2,5}	BH _{5,0}	L _{0,1}	L _{2,5}	L _{5,0}	HM _{0,1}	HM _{2,5}	HM _{5,0}
100	OS	10%	43	57	61	74	4	65	1	11	74	17	4	77	64	34	22	59	58
		20%	33	43	36	56	4	46	1	5	54	1	1	60	28	3	18	42	35
		30%	26	31	16	35	3	27	1	4	31	1	1	42	7	0	13	30	18
		40%	17	21	7	20	3	14	1	3	15	0	0	29	2	0	12	19	10
		50%	12	12	2	10	3	5	1	3	5	0	0	19	0	0	9	13	5
			8	7	12	4	7	7	18	11	13	7	7	15	5	6	4	7	8
	FK	10%	5	4	5	2	4	4	13	7	7	4	4	12	3	3	4	4	4
		20%	4	4	4	2	3	3	10	6	5	3	3	11	2	3	4	4	4
		30%	3	2	2	2	3	2	7	5	3	2	2	9	2	2	4	3	3
		40%	2	2	2	2	2	2	6	4	3	2	2	8	2	2	3	3	2
		50%																	
			76	90	93	98	4	95	1	16	98	67	24	98	97	85	53	90	88
200	OS	10%	61	75	73	89	2	81	1	7	88	6	1	92	78	27	37	76	66
		20%	47	60	48	76	2	63	1	5	70	1	1	80	44	3	27	60	46
		30%	34	45	22	55	1	43	0	3	47	1	0	66	16	0	19	41	26
		40%	25	29	10	34	2	23	1	3	25	1	1	47	3	0	15	29	15
		50%																	
			11	12	21	4	11	9	32	13	23	10	11	24	5	8	6	9	11
	FK	10%	10	8	11	4	5	7	19	7	15	6	7	17	4	6	5	7	8
		20%	8	5	7	3	3	5	14	5	9	4	5	14	3	4	5	5	6
		30%	5	4	4	2	2	3	11	4	5	3	3	11	2	3	4	4	4
		40%	4	3	3	2	2	3	8	4	3	2	2	9	2	2	4	3	3
		50%																	
			76	90	93	98	4	95	1	16	98	67	24	98	97	85	53	90	88

Table 11 Empirical rejection rates for Scenario B where $\mathcal{H}_0$ is the PH model and $\mathcal{H}_1$ is the Hsieh model with $\gamma = 0.25$, both for exponential baseline.

n	τ	p	KS	CM	AD	EP	CO	M	AT _{0.01}	BH _{0.05}	BH _{0.1}	BH _{0.5}	L _{0.75}	L _{1.0}	L _{1.5}	HM ₁	HM _{1.5}	HM ₂
100	OS	10%	22	27	25	31	30	28	23	21	24	29	30	31	31	25	26	26
		20%	21	24	18	24	25	23	11	17	19	22	26	26	26	21	22	21
		30%	16	19	12	19	20	17	7	13	16	15	21	21	20	17	18	16
		40%	14	16	7	17	16	14	4	10	12	10	19	19	18	15	15	13
		50%	11	12	4	13	15	9	2	7	8	6	16	16	15	13	12	10
	FK	10%	7	6	5	5	6	5	4	6	6	4	6	6	5	6	5	5
		20%	4	4	2	3	6	3	3	4	3	2	4	4	3	4	3	3
		30%	3	3	2	2	5	2	3	3	2	2	4	3	3	3	3	2
		40%	3	2	1	2	5	2	3	2	2	2	3	3	2	3	2	2
		50%	2	2	1	2	5	2	3	2	2	2	3	2	2	2	2	2
200	OS	10%	36	43	43	47	46	45	33	38	42	48	49	50	50	37	40	40
		20%	29	35	32	38	37	37	17	31	34	37	40	41	41	32	34	33
		30%	25	31	24	33	33	31	8	26	28	29	35	36	36	28	29	27
		40%	22	26	15	26	26	23	4	20	22	19	30	30	29	25	23	21
		50%	17	19	8	21	21	17	1	14	15	12	22	22	22	18	17	15
	FK	10%	8	8	6	7	7	6	3	8	7	5	6	6	6	6	6	6
		20%	6	5	3	4	6	4	2	5	4	3	5	5	4	5	4	4
		30%	4	3	2	3	7	2	1	3	3	2	4	4	3	4	3	2
		40%	3	2	1	2	5	2	1	2	2	1	3	3	2	3	2	2
		50%	2	2	1	2	6	1	2	2	2	1	3	2	2	2	2	1

Table 12 Empirical rejection rates for Scenario B where $\mathcal{H}_0$ is the PH model and $\mathcal{H}_1$ is the Hsieh model with $\gamma = 0.50$, both for exponential baseline.

n	τ	p	KS	CM	AD	EP	CO	M	AT _{0.01}	BH _{0.05}	BH _{0.1}	BH _{0.5}	L _{0.75}	L _{1.0}	L _{1.5}	HM ₁	HM _{1.5}	HM ₂
100	OS	10%	67	77	75	82	78	79	72	67	73	81	81	82	83	72	75	74
		20%	60	68	62	71	66	69	47	58	63	69	73	74	74	62	65	63
		30%	49	58	46	60	54	56	30	47	52	55	63	63	63	53	55	52
		40%	41	48	31	49	43	44	18	36	40	39	55	55	53	45	45	41
		50%	32	38	17	38	36	29	10	26	29	24	46	46	43	37	35	31
	FK	10%	15	16	13	14	7	13	5	17	15	11	12	13	13	13	13	13
		20%	10	10	6	8	8	7	3	9	8	5	8	8	7	8	7	6
		30%	6	6	3	4	6	3	3	5	4	3	6	6	5	6	5	4
		40%	4	3	2	3	6	2	2	3	3	2	5	4	3	4	3	2
		50%	3	2	1	2	5	2	3	3	2	2	4	3	2	3	2	2
200	OS	10%	89	95	95	96	93	96	89	92	94	97	97	97	97	91	93	93
		20%	82	89	88	91	83	90	71	86	88	92	93	93	93	86	87	86
		30%	73	81	76	84	73	82	45	76	79	82	86	87	87	77	78	76
		40%	64	73	61	73	60	70	25	65	69	67	78	78	77	69	69	65
		50%	51	57	40	60	48	54	11	50	53	49	66	65	65	56	54	49
	FK	10%	26	29	24	24	9	25	4	34	28	20	19	21	23	21	23	22
		20%	17	18	12	14	8	13	2	20	15	10	13	14	14	14	14	12
		30%	11	11	5	8	9	6	1	11	8	5	9	9	8	9	8	6
		40%	7	6	3	5	7	3	1	6	4	2	6	6	5	6	5	4
		50%	4	3	1	2	7	2	2	3	3	1	4	4	3	4	3	2

Table 13 Empirical rejection rates for Scenario B where $\mathcal{H}_0$ is the PH model and $\mathcal{H}_1$ is the Hsieh model with $\gamma = 0.50$, both for Weibull baseline.

n	τ	p	KS	CM	AD	EP	CO	M	AT _{0.01}	BH _{0.05}	BH _{0.1}	BH _{0.5}	L _{0.05}	L _{0.75}	L ₁	HM ₁	HM _{1.5}	HM ₂
100	OS	10%	6	5	5	7	3	6	3	6	6	4	7	6	7	5	5	5
		20%	6	6	3	7	4	6	1	6	5	2	7	7	6	5	5	4
		30%	5	5	2	5	4	4	1	4	3	1	6	5	4	6	5	4
		40%	4	4	1	4	5	2	1	4	2	1	6	4	4	5	4	3
		50%	4	3	1	3	5	2	2	3	2	1	6	4	3	5	4	3
	FK	10%	5	5	4	5	4	4	4	4	4	4	5	5	5	5	5	5
		20%	4	3	2	3	4	3	3	3	3	3	5	4	4	5	5	4
		30%	3	3	2	3	4	3	4	3	3	2	5	4	3	5	4	4
		40%	2	2	2	2	4	2	4	2	2	2	5	3	3	4	4	3
		50%	2	2	2	2	5	2	5	2	2	2	5	2	2	4	3	3
200	OS	10%	5	6	5	8	3	6	4	7	6	5	8	8	8	5	5	5
		20%	5	5	3	7	4	5	2	6	5	3	8	7	7	5	5	5
		30%	5	6	2	7	5	5	1	5	4	1	7	7	6	5	5	5
		40%	5	5	1	5	4	3	1	4	3	1	7	6	5	6	4	4
		50%	5	4	2	4	5	2	1	3	2	1	6	5	4	5	5	4
	FK	10%	4	4	4	4	4	4	3	4	4	4	4	4	4	5	5	5
		20%	4	3	3	4	4	3	3	4	4	3	4	4	4	5	4	4
		30%	4	3	2	3	4	2	2	3	3	2	4	4	3	5	5	4
		40%	3	2	2	2	4	2	3	2	2	2	4	3	3	5	4	4
		50%	2	2	2	2	4	2	4	2	2	2	4	3	2	4	3	3

Table 14 Empirical rejection rates for Scenario B where $\mathcal{H}_0$ is the PH model and $\mathcal{H}_1$ is the Hsieh model with $\gamma = 1$, both for Weibull baseline.

n	τ	p	KS	CM	AD	EP	CO	M	AT _{0.01}	BH _{0.05}	BH _{0.1}	BH _{0.5}	L _{0.05}	L _{0.75}	L ₁	HM ₁	HM _{1.5}	HM ₂
100	OS	10%	10	10	11	20	5	13	7	16	14	12	20	16	20	8	9	10
		20%	10	11	7	18	6	12	2	15	12	5	17	17	18	8	9	8
		30%	8	8	3	12	7	7	2	10	7	2	14	13	13	7	7	6
		40%	6	7	3	9	8	4	3	7	4	1	12	11	10	8	7	5
		50%	5	5	3	6	10	2	6	5	3	3	11	9	7	6	6	5
	FK	10%	6	5	4	5	4	5	4	5	5	5	4	5	5	6	6	5
		20%	4	3	2	3	5	3	3	3	3	3	4	4	4	6	5	4
		30%	3	3	2	3	5	2	5	3	3	2	5	4	3	5	5	4
		40%	2	2	2	2	6	2	5	2	2	2	4	3	3	4	4	3
		50%	3	3	3	3	7	3	9	3	3	3	5	3	3	4	4	3
200	OS	10%	13	15	16	30	4	20	12	25	22	17	28	23	28	9	12	14
		20%	13	14	10	25	6	17	2	21	18	8	25	22	26	9	12	12
		30%	11	12	6	19	8	12	1	15	11	2	20	19	20	9	12	11
		40%	8	9	3	11	7	7	2	10	6	1	15	15	15	10	11	9
		50%	7	7	7	8	8	5	7	7	6	5	11	13	12	10	10	9
	FK	10%	5	5	4	5	4	5	4	4	4	5	4	5	5	5	5	5
		20%	4	4	3	4	4	3	2	4	4	3	4	5	5	6	5	5
		30%	3	2	2	2	6	2	3	2	2	2	3	4	3	5	5	4
		40%	3	2	2	2	5	2	6	3	3	2	4	3	3	5	4	4
		50%	4	3	5	4	7	4	10	4	4	4	4	4	3	5	5	5

Table 15 Empirical rejection rates for Scenario C where $\mathcal{H}_0$ is log-normal and $\mathcal{H}_1$ is log-logistic.

n	τ	p	KS	CM	AD	EP	CO	M	AT _{0,01}	BH _{0,50}	BH _{0,75}	BH ₁	L _{0,75}	L ₁	L _{1,5}	HM ₁	HM _{1,5}	HM ₂
100	OS	10%	12	14	16	14	14	14	3	20	19	18	27	25	21	13	13	12
		20%	12	14	14	16	14	14	4	18	16	15	27	26	23	13	13	12
		30%	12	13	12	17	14	13	4	14	13	12	26	25	22	13	14	12
		40%	11	13	10	19	13	14	4	11	10	8	25	25	23	11	12	10
		50%	11	12	9	17	11	12	4	9	7	7	22	22	19	11	12	10
	FK	10%	5	5	5	5	5	5	6	5	5	5	5	5	5	6	6	6
		20%	5	6	5	5	5	5	6	5	4	4	4	5	5	6	5	6
		30%	5	5	4	5	5	4	6	4	4	4	4	5	5	5	5	6
		40%	4	4	4	4	5	4	5	4	4	4	4	5	5	4	5	5
		50%	5	5	4	4	4	4	5	4	4	4	4	5	5	6	5	5
200	OS	10%	17	22	25	19	19	21	3	33	31	29	41	38	31	21	20	18
		20%	18	23	24	25	19	22	3	30	27	25	42	40	35	21	22	20
		30%	19	22	22	28	19	22	4	26	23	20	42	41	38	20	22	20
		40%	18	20	18	28	17	22	4	21	18	15	39	39	36	18	20	20
		50%	16	19	14	27	16	20	4	14	12	10	36	35	33	17	18	17
	FK	10%	7	7	7	6	5	6	7	6	6	6	6	6	6	7	8	8
		20%	6	6	5	5	4	5	6	5	4	4	6	6	5	6	7	7
		30%	6	5	4	5	5	4	5	4	4	4	5	5	5	6	6	6
		40%	5	5	4	4	5	4	5	4	3	3	5	5	4	6	5	6
		50%	4	4	3	4	5	3	5	3	3	3	4	4	4	5	5	5

Table 16 Empirical rejection rates for Scenario C where $\mathcal{H}_0$ is log-normal and $\mathcal{H}_1$ is Weibull.

n	τ	p	KS	CM	AD	EP	CO	M	AT _{0,01}	BH _{0,50}	BH _{0,75}	BH ₁	L _{0,75}	L ₁	L _{1,5}	HM ₁	HM _{1,5}	HM ₂
100	OS	10%	59	70	72	86	53	78	9	83	83	84	81	84	86	58	70	74
		20%	48	56	58	75	33	66	14	70	71	70	73	75	78	45	54	56
		30%	37	44	46	64	26	53	15	55	54	52	63	65	67	34	40	39
		40%	29	34	34	55	20	45	14	40	38	35	53	56	58	25	29	28
		50%	22	27	25	42	17	33	13	29	27	25	44	45	46	19	21	19
	FK	10%	20	20	18	16	5	18	6	14	14	14	11	12	14	13	15	17
		20%	13	13	11	10	6	10	8	9	9	8	9	9	10	10	11	13
		30%	9	9	8	8	6	8	7	8	7	7	9	8	8	9	8	9
		40%	8	8	8	8	6	8	7	8	8	7	8	8	8	7	8	8
		50%	8	8	9	8	6	8	6	9	9	8	9	9	9	7	8	8
200	OS	10%	88	95	95	99	68	97	10	98	99	99	98	98	99	87	94	96
		20%	78	86	87	96	41	92	14	94	94	94	94	95	96	74	83	85
		30%	65	75	74	91	30	84	16	85	84	83	88	90	92	60	68	69
		40%	51	60	60	81	24	72	17	70	69	65	79	82	84	44	52	53
		50%	38	47	44	69	21	59	16	51	48	44	67	70	72	33	38	37
	FK	10%	38	42	39	31	5	36	7	29	29	29	19	22	27	25	32	36
		20%	20	21	18	15	5	16	8	13	12	12	11	12	13	16	18	20
		30%	13	12	11	10	6	10	8	9	8	8	9	9	9	11	12	13
		40%	9	9	9	8	7	8	7	8	8	8	9	9	9	10	9	10
		50%	9	9	10	9	6	9	6	9	9	9	10	9	9	9	9	9

Table 17 Empirical rejection rates for Scenario A where $\mathcal{H}_0$ is exponential.

n	$\mathcal{H}_1$ p	Exponential			Gamma			Gompertz			Log-normal			Weibull		
		$\widetilde{CM}$	$\widetilde{AD}$	$\widetilde{CO}$	$\widetilde{CM}$	$\widetilde{AD}$	$\widetilde{CO}$	$\widetilde{CM}$	$\widetilde{AD}$	$\widetilde{CO}$	$\widetilde{CM}$	$\widetilde{AD}$	$\widetilde{CO}$	$\widetilde{CM}$	$\widetilde{AD}$	$\widetilde{CO}$
100	10%	5	4	5	99	100	100	42	31	37	92	90	84	16	12	18
	20%	3	2	5	98	100	100	21	13	25	68	59	49	9	6	16
	30%	2	2	5	93	99	100	6	5	16	28	24	19	3	3	13
	40%	1	1	5	76	96	100	1	2	10	4	5	7	1	1	11
	50%	1	1	5	37	88	99	1	1	9	1	1	3	1	1	11
200	10%	4	4	5	100	100	100	71	62	65	100	100	99	26	24	32
	20%	4	3	5	100	100	100	41	31	42	93	90	79	16	14	25
	30%	2	2	5	100	100	100	15	13	28	58	53	37	7	8	23
	40%	1	1	5	99	100	100	2	4	18	12	14	10	2	4	19
	50%	1	1	5	90	100	100	1	2	12	1	2	3	1	2	17

Table 18 Empirical rejection rates for Scenario A where $\mathcal{H}_0$ is Weibull.

n	$\mathcal{H}_1$ p	Weibull			Gamma			Gompertz			Log-normal		
		$\widetilde{CM}$	$\widetilde{AD}$	$\widetilde{CO}$	$\widetilde{CM}$	$\widetilde{AD}$	$\widetilde{CO}$	$\widetilde{CM}$	$\widetilde{AD}$	$\widetilde{CO}$	$\widetilde{CM}$	$\widetilde{AD}$	$\widetilde{CO}$
100	10%	5	3	4	11	8	6	10	7	6	58	55	6
	20%	2	1	5	4	2	5	3	2	5	27	19	2
	30%	1	1	5	1	1	5	1	1	6	5	4	2
	40%	1	1	4	1	1	6	1	1	7	1	1	5
	50%	1	1	4	1	1	6	1	1	6	1	1	5
200	10%	4	2	4	20	16	5	18	13	5	90	91	6
	20%	2	1	4	7	5	6	5	3	5	62	58	3
	30%	1	1	4	1	1	6	1	1	6	19	20	2
	40%	1	1	4	1	1	6	1	1	7	3	4	4
	50%	1	1	5	1	1	7	1	1	7	1	1	5

Table 19 Empirical rejection rates for Scenario B where $\mathcal{H}_0$ is the PH model and $\mathcal{H}_1$ is the Hsieh model, both with the same baseline hazard.

n	p	$\mathcal{H}_0$ Exponential						$\mathcal{H}_0$ Weibull					
		$\gamma = 0.25$			$\gamma = 0.50$			$\gamma = 0.50$			$\gamma = 1.00$		
		$\widetilde{CM}$	$\widetilde{AD}$	$\widetilde{CO}$	$\widetilde{CM}$	$\widetilde{AD}$	$\widetilde{CO}$	$\widetilde{CM}$	$\widetilde{AD}$	$\widetilde{CO}$	$\widetilde{CM}$	$\widetilde{AD}$	$\widetilde{CO}$
100	10%	28	24	33	81	78	86	5	3	5	10	8	4
	20%	18	13	29	65	58	80	2	1	5	5	3	6
	30%	7	6	23	37	37	69	1	1	6	1	1	8
	40%	1	3	18	8	16	57	1	1	6	2	2	8
	50%	1	1	17	2	6	47	1	1	5	4	4	6
200	10%	44	40	51	96	95	97	6	4	6	15	14	3
	20%	29	25	42	88	86	94	2	2	5	7	5	5
	30%	14	15	36	65	69	89	1	1	7	1	2	9
	40%	3	7	30	27	46	81	0	1	5	2	2	7
	50%	1	3	24	5	23	70	1	1	4	6	6	4

Table 20 Empirical rejection rates for Scenario C where $\mathcal{H}_0$ is log-normal.

n	p	$\mathcal{H}_a$ Log-logistic			$\mathcal{H}_a$ Weibull		
		$\widetilde{CM}$	$\widetilde{AD}$	$\widetilde{CO}$	$\widetilde{CM}$	$\widetilde{AD}$	$\widetilde{CO}$
100	10%	14	15	18	70	71	87
	20%	12	14	21	56	59	76
	30%	9	12	21	43	47	63
	40%	6	9	23	27	33	53
	50%	5	6	19	18	24	40
200	10%	22	25	24	95	95	99
	20%	21	23	31	87	88	96
	30%	16	21	34	74	77	89
	40%	12	17	35	57	64	80
	50%	8	14	32	37	48	65

Chapter 4

Article 3: Modified tests for exponentiality in the presence of Type-II right censored data

The third article, namely: *Modified tests for exponentiality in the presence of Type-II right censored data*, has been published in the journal *Stat*. A summary of the guidelines to authors from the journal is now presented.

Manuscript	Maximum length of 10 pages (excluding references, figures, tables and other forms of exhibit, but including text in appendices).
Title	A full title and a short title of up to 70 characters should be provided.
Abstract and keywords	A maximum of 200 words succinctly describing the article and five or six keywords or key phrases from the list of keywords on the ScholarOne site.
Tables	Tables must be on separate pages after the reference list, and not be incorporated into the main text. All tables should be referred to sequentially in the text.
References	References should be prepared according to the <i>Publication Manual of the American Psychological Association</i> (6th edition).
General formatting	A LaTeX template is provided for submission.
Additional information	https://tinyurl.com/y4lu7f9f

ORIGINAL ARTICLE

Modified tests for exponentiality in the presence of Type II right censored data

Marike Cockeran^{ID} | Leonard Santana | James S. Allison

Unit for Business Mathematics and Informatics, North-West University, Potchefstroom, 2520, South Africa

Correspondence

Marike Cockeran, Unit for Business Mathematics and Informatics, North-West University, Potchefstroom 2520, South Africa.
Email: marike.cockeran@nwu.ac.za

The exponential distribution is one of the most popular choices of model, especially in survival analysis and reliability theory. Typically, to test the hypothesis that the underlying distribution of lifetimes are exponentially distributed based on censored data, the standard goodness-of-fit tests used in the complete sample case are modified. However, none of the characteristic or Laplace function based tests have been adapted to incorporate censored data. In this paper, we modify three well-known tests for exponentiality on the basis of either the characteristic function or Laplace transform to accommodate Type II right censored data. We compare the power of the three newly changed tests to some existing modified tests in an independent identically distributed (i.i.d.) and a model-based setup through a Monte Carlo study, where it was found that the newly adjusted Baringhaus and Henze test outperformed all other tests considered.

KEYWORDS

bootstrap, censored data, exponential family, survival analysis

1 | INTRODUCTION

The exponential distribution is one of the most popular choices of model, especially in survival analysis and reliability theory. In the complete sample case, numerous tests for testing the hypothesis that the observed data are realizations from the exponential distribution have been developed and studied. These include tests based on characteristic functions (Epps & Pulley, 1986; Meintanis, Swanepoel, & Allison, 2014), Laplace transforms (Baringhaus & Henze, 1991; Henze & Meintanis, 2002), mean residual life (Baringhaus & Henze, 2000; Jammalamadaka & Taufer, 2006), normalized spacings (Gnedenko, Belyayev, & Solov'yev, 1969), and entropy (Zardasht, Parsi, & Mousazadeh, 2015; Bara'tpour & Rad, 2012) to name just a few. Furthermore, there are a multitude of properties that characterize the exponential distribution, see, for example, the monographs of Galambos and Kotz (1978) and Kotz, Balakrishnan, and Johnson (2004). Many goodness-of-fit (GOF) tests for exponentiality have been developed based on these characterizations; Ahmad and Alwasel (1999) based a test on the memory-less property, Volkova and Nikitin (2013) based a test on order statistics, and Henze and Meintanis (2002) developed a test on an equation involving the real and imaginary part of the characteristic function, for which the exponential distribution is the unique solution. For an in-depth discussion on tests for exponentiality in the complete sample case, the interested reader is referred to the review papers by Ascher (1990), Henze and Meintanis (2005), Allison, Santana, Smit, and Visagie (2017), Torabi, Narges, and Grane (2018), and the references therein.

In survival analysis and reliability theory, we most often deal with the right censored data. Typically, to test the hypothesis that the underlying distribution of the lifetimes are exponentially distributed based on censored data, the standard GOF tests used in the complete sample case are modified. Early works include the modifications of the Cramér-von Mises and Kolmogorov-Smirnov test statistics. For the modification of the Cramér-von Mises statistic see Pettitt (1976) and Pettitt and Stephens (1976), and specifically for the exponential distribution where the scale parameter must be estimated, see Pettitt (1977). For the work on the Kolmogorov-Smirnov test statistic, the reader is referred to Barr and Davidson (1973) and Dufour and Maag (1978). For random censored data, Koziol and Green (1976) modified the Cramér-von Mises test statistic for a simple hypothesis on the basis of the Kaplan-Meier product limit distribution function. Kim (2012) generalized these results to the composite hypothesis for exponentiality with an unknown scale parameter. In the case of Type II right censored data, the reader is referred to Balakrishnan, Chimitova, and Vedernikova (2015), Castro-Kuriss (2011), and Castro-Kuriss, Kelmansky, Leiva, and Martinez (2010). However, none of the characteristic or Laplace function based test statistics have been modified in this way. Another approach that can be followed is to transform the censored sample to a complete sample, then any of the numerous tests for exponentiality can be used. This approach is discussed in Glen and

Foote (2009), and Goldmann, Klar, and Meintanis (2015) applied these data transformations for testing the GOF for various distributions on the basis of Type II right censored samples. For a discussion on earlier works on the approaches mentioned above, see D'Agostino & Stephens, 1986.

The aim of this paper is to modify three well-known tests for exponentiality based on either the characteristic function or Laplace transform to accommodate Type II right censored data. The remainder of the paper is organized as follows: In Section 2, we present the tests on the basis of the characteristic function and Laplace transform. In Section 3, we show how to modify the three test statistics discussed in Section 2 in order to apply them to the case of Type II right censored data. Section 4 presents the results of a Monte Carlo study that compares the three newly modified tests to some existing modified tests in the independent identically distributed (i.i.d.) case and a model-based setup. The paper concludes with final remarks and a discussion on the performance of the tests in Section 5.

2 | TESTS BASED ON THE CHARACTERISTIC FUNCTION AND THE LAPLACE TRANSFORM IN THE COMPLETE SAMPLE CASE

Let $X_1, X_2, \dots, X_n$ be the i.i.d. continuous realizations of a random variable X with unknown distribution F . Furthermore, let $\text{Exp}(\lambda)$ denote the exponential distribution with density $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$. The composite GOF hypothesis to be tested is

$$H_0 : \text{The law of } X \text{ is } \text{Exp}(\lambda), \quad (1)$$

for some $\lambda > 0$, against general alternatives. All the test statistics that we consider are based on the scaled values $Z_j = X_j \hat{\lambda}_n$ with $\hat{\lambda}_n = 1/\bar{X}_n$ the maximum likelihood estimator of λ and $\bar{X}_n = \frac{1}{n} \sum_{j=1}^n X_j$. We will specifically focus on three tests, on the basis of the characteristic function (CF) or Laplace transform (LT) and their empirical counterparts, to test the hypothesis in (1). Recall that the CF of the random variable X is given by

$$\phi(t) = E[e^{itX}] = \int_{-\infty}^{\infty} e^{itx} dF(x). \quad (2)$$

Now, if we estimate F by F_n , the empirical distribution function (EDF), then we obtain the empirical characteristic function

$$\phi_n(t) = \int_{-\infty}^{\infty} e^{itx} dF_n(x) = \frac{1}{n} \sum_{j=1}^n e^{itX_j},$$

which is an estimator for $\phi(t)$. Similarly, the LT of a random variable X is given by

$$\psi(t) = E[e^{-tX}] = \int_{-\infty}^{\infty} e^{-tx} dF(x), \quad (3)$$

and can be estimated by the empirical Laplace transform

$$\psi_n(t) = \int_{-\infty}^{\infty} e^{-tx} dF_n(x) = \frac{1}{n} \sum_{j=1}^n e^{-tX_j}.$$

The first of these tests is the one given by Epps and Pulley (1986),

$$EP_n = \int_{-\infty}^{\infty} [\phi_n(t) - \phi(t)] \frac{1}{2\pi} \phi(-t) dt, \quad (4)$$

the second by Baringhaus and Henze (1991),

$$BH_{n,a} = n \int_0^{\infty} [(1+t)\psi'_n(t) + \psi_n(t)]^2 e^{-at} dt, \quad (5)$$

and the third, more recent test, by Henze and Meintanis (2005), is given by

$$L_{n,a} = n \int_0^{\infty} [\psi_n(t) - \psi(t)]^2 (1+t)^2 e^{-at} dt, \quad (6)$$

with $a > 0$ a user defined tuning parameter. The above test statistics are only applicable in the full sample case; however, in the next section, we will discuss how these three statistics can be modified in order to accommodate Type II right censored data.

3 | MODIFICATION OF TEST STATISTICS

To accommodate censored data, which are commonly encountered in survival analysis and reliability theory, we now introduce some notation for right censored data in general and then specifically focus on Type II right censored data. Let $Y_1, Y_2, \dots, Y_n$ be i.i.d. lifetime variables with continuous distribution function F and survival function S , and let $C_1, C_2, \dots, C_n$ be i.i.d. censoring variables with distribution function G , independent of $Y_1, Y_2, \dots, Y_n$. We observe the random sample (T_i, δ_i) , $i = 1, 2, \dots, n$, where

$$T_i = \min(Y_i, C_i) \quad \text{and} \quad \delta_i = \begin{cases} 1, & \text{if } T_i = Y_i \\ 0, & \text{if } T_i = C_i. \end{cases}$$

Furthermore let $Y_{(1)} < Y_{(2)} < \dots < Y_{(n)}$ and $T_{(1)} < T_{(2)} < \dots < T_{(n)}$ denote the order statistics of $Y_1, Y_2, \dots, Y_n$, and $T_1, T_2, \dots, T_n$, respectively, with $\delta_{(i)}$ denoting the indicator variable associated with $Y_{(i)}$ and $T_{(i)}$. In this setup, we can use the Kaplan–Meier estimator $\tilde{F}_n$ to estimate F , where

$$1 - \tilde{F}_n(t) = \begin{cases} 1, & t \leq T_{(1)} \\ \prod_{j=1}^{k-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}}, & T_{(k-1)} < t \leq T_{(k)}, \quad k = 2, \dots, n. \\ 0, & t > T_{(n)}. \end{cases}$$

This product limit estimator has been studied in detail by, among others, Kaplan and Meier (1958), Efron (1967), and Breslow and Crowley (1974). If we let Δ_i denote the height of the jump of $\tilde{F}_n(t)$ at $Y_{(i)}$, we have that

$$\Delta_1 = \frac{\delta_{(1)}}{n}, \quad \Delta_n = \sum_{j=1}^{n-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}}$$

and, for $2 \leq i \leq n-1$,

$$\begin{aligned} \Delta_i &= \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}} - \prod_{j=1}^i \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}} \\ &= \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}}. \end{aligned}$$

If we now consider the characteristic function defined in (2), then a natural estimator of $\phi(t)$ in this censoring setup is

$$\tilde{\phi}_n(t) = \int_0^\infty e^{itx} d\tilde{F}_n(x) = \sum_{j=1}^n \Delta_j e^{itT_{(j)}}. \quad (7)$$

For some theoretical properties of this estimator, the interested reader is referred to Csorgo (1984). Similarly, an estimator for the Laplace transform $\psi(t)$ defined in (3) is given by

$$\tilde{\psi}_n(t) = \int_0^\infty e^{-tx} d\tilde{F}_n(x) = \sum_{j=1}^n \Delta_j e^{-tT_{(j)}}. \quad (8)$$

In the specific case of Type II right censoring, that is, where we only observed the first $r \leq n$ order statistics, we have that $T_{(i)} = Y_{(i)}$, for $i = 1, 2, \dots, r$, and $T_{(i)} = X_{(r)}$ for $i = r+1, \dots, n$. The expressions in (7) and (8) then simplify to

$$\tilde{\phi}_{n,r}(t) = \frac{1}{n} \sum_{j=1}^n e^{itT_{(j)}} \mathbb{1}(T_{(j)} \leq Y_{(r)}) = \frac{1}{n} \sum_{j=1}^r e^{itT_{(j)}}$$

and

$$\tilde{\psi}_{n,r}(t) = \frac{1}{n} \sum_{j=1}^n e^{-tT_{(j)}} \mathbb{1}(T_{(j)} \leq Y_{(r)}) = \frac{1}{n} \sum_{j=1}^r e^{-tT_{(j)}},$$

respectively, where

$$\mathbb{1}(T_{(j)} \leq Y_{(r)}) = \begin{cases} 1, & \text{if } T_{(j)} \leq Y_{(r)} \\ 0, & \text{else.} \end{cases}$$

We are now in a position to modify the three test statistics discussed in Section 2 in order to apply them to the case of Type II right censored data. The modified versions of EP_n , $BH_{n,a}$, and $L_{n,a}$ are obtained by simply replacing ϕ_n with $\tilde{\phi}_n$ in (4) and replacing ψ_n with $\tilde{\psi}_n$ in (5) and (6). After

some simple calculations, the resulting convenient computational formula for these modified statistics are then given by

$$\widetilde{\text{EP}}_{n,r} = \sqrt{48n} \left[\frac{1}{n} \sum_{j=1}^r e^{-\tilde{Z}_j} - \frac{1}{2} \right],$$

$$\widetilde{\text{BH}}_{n,r,a} = \frac{1}{n} \sum_{j=1}^r \sum_{k=1}^r \left[\frac{(1 - \tilde{Z}_j)(1 - \tilde{Z}_k)}{\tilde{Z}_j + \tilde{Z}_k + a} - \frac{\tilde{Z}_j + \tilde{Z}_k}{(\tilde{Z}_j + \tilde{Z}_k + a)^2} + \frac{2\tilde{Z}_j\tilde{Z}_k}{(\tilde{Z}_j + \tilde{Z}_k + a)^2} + \frac{2\tilde{Z}_j\tilde{Z}_k}{(\tilde{Z}_j + \tilde{Z}_k + a)^3} \right],$$

and

$$\tilde{\text{L}}_{n,r,a} = \frac{1}{n} \sum_{j=1}^r \sum_{k=1}^r \left[\frac{1 + (\tilde{Z}_j + \tilde{Z}_k + a + 1)^2}{(\tilde{Z}_j + \tilde{Z}_k + a)^3} \right] - 2 \sum_{j=1}^r \left[\frac{1 + \tilde{Z}_j + a}{(\tilde{Z}_j + a)^2} \right] + \frac{n}{a},$$

with $\tilde{Z}_j = T_j \hat{\lambda}_{n,r}$, where

$$\hat{\lambda}_{n,r} = \frac{r}{\sum_{j=1}^r Y_j + (n-r)Y_r}$$

is the maximum likelihood estimator of λ based on Type II right censored data.

4 | SIMULATION STUDY

Let $W_{(i)} = 1 - e^{-\tilde{Z}_{(i)}}$, where $\tilde{Z}_{(i)}$, $i = 1, 2, \dots, r$ denote the order statistics of $\tilde{Z}_1, \tilde{Z}_2, \dots, \tilde{Z}_r$. In this section, we now compare the power performance of our three newly modified test statistics to that of the existing modified Kolmogorov–Smirnov, Cramér–von Mises, Anderson–Darling, and Cox and Oakes test statistics given by

$$\text{KS}_{n,r} = \max_{1 \leq i \leq r} \left| \frac{i-0.5}{n} - \tilde{W}_{(i)} \right| + \frac{0.5}{n},$$

$$\text{CM}_{n,r} = \sum_{i=1}^r (\tilde{W}_{(i)} - \frac{2i-1}{2n})^2 + \frac{r}{12n^2} + \frac{n}{3} (\tilde{W}_{(r)} - \frac{r}{n})^3,$$

$$\begin{aligned} \text{AD}_{n,r} = & -\frac{1}{n} \sum_{j=1}^r (2j-1) [\log\{\tilde{W}_{(j)}\} - \log\{1 - \tilde{W}_{(j)}\}] - 2 \sum_{j=1}^r \log\{1 - \tilde{W}_{(j)}\} \\ & - \frac{1}{n} [(r-n)^2 \log\{1 - \tilde{W}_{(r)}\} - r^2 \log\{\tilde{W}_{(r)}\} + n^2 \tilde{W}_{(r)}], \end{aligned}$$

and

$$\text{CO}_{n,r} = r + \sum_{j=1}^r \ln(\tilde{Z}_j) - r \frac{\sum_{j=1}^n \tilde{Z}_j \ln(\tilde{Z}_j)}{\sum_{j=1}^n \tilde{Z}_j},$$

respectively. This comparison will be done in both the i.i.d. setup and in a model-based setup, where we test the adequacy of a fitted Cox proportional hazards (PH) model. All size and power estimates are calculated for 10,000 independent Monte Carlo samples for sample sizes $n = 40$ and $n = 80$ in the i.i.d. case and $n = 100$ and $n = 200$ in the model-based case. The censored sample size is given by $r = \lfloor np \rfloor$, with the notation $\lfloor x \rfloor$ denoting the largest integer less than x , and where p is the censoring proportion with values 10%, 20%, 30%, and 40%. Each test, in which a tuning parameter a appears, was evaluated for different values of a , but we only report the choice of a that produced reasonably good powers in a separate simulation study (not shown here); for the $\widetilde{\text{BH}}$ test, we chose $a = 0.5$, and for the $\tilde{\text{L}}$ test, we used $a = 0.05$. The entries in the tables are the percentage of 10,000 independent Monte Carlo samples that resulted in the rejection of the null hypothesis rounded to the nearest integer. For ease of comparison, the highest estimated power for each alternative has been highlighted in the reported tables. All calculations and simulations were performed with R version 3.2.4 (Development Core Team, R., 2016) and PH models were fit using the `flexsurvreg` function in the R package `flexsurv`.

4.1 | Simulation setting and results for the i.i.d. setup

Let $Y_1, Y_2, \dots, Y_n$ denote the failure times randomly sampled from one of the distributions listed in Table 1. Further, let $r < n$ be fixed and denote by $Y_{(1)} < Y_{(2)} < \dots < Y_{(r)}$ the first r ordered values of $Y_1, Y_2, \dots, Y_n$. The estimated sizes of the tests are reported in the first five rows of Tables 2 and 3 for $n = 40$ and $n = 80$, respectively, with data generated from $\text{Exp}(1)$. From these results, one can see that all the test statistics maintain the size of the test over all censoring proportions. The remaining entries in Tables 2 and 3 contain the estimated powers for data generated from the distributions $\Gamma(\theta)$, $\text{LN}(\theta)$, $\text{EL}(\theta)$, and $\text{EG}(\theta)$ in Table 1 for various choices of the parameter θ . When considering the power performance of these test statistics, one finds that the $\widetilde{\text{EP}}$ and $\widetilde{\text{BH}}$ tests are consistently the best performing tests followed by the $\tilde{\text{L}}$ test. All tests show low estimated powers for the alternatives that are close to the null hypothesis, namely, the $\text{EG}(0.2)$, $\text{EL}(0.5)$, and $\text{EL}(0.8)$ alternatives. On the other hand, all tests show high estimated powers for the $\Gamma(0.4)$ alternative, and for the $\Gamma(0.7)$ alternative, reasonable powers are observed, with the $\tilde{\text{L}}$ test attaining the highest estimated powers in both cases. For the $\text{LN}(1.5)$ alternative, the $\widetilde{\text{EP}}$ test displays better results for lower censoring proportions, whereas the $\widetilde{\text{BH}}$ test performs better for higher censoring proportions. The CO and $\tilde{\text{L}}$ tests are the worst performers for this alternative.

Alternative	$f(x)$	Notation
Exponential	e^{-x}	Exp(1)
Gamma	$\frac{1}{\Gamma(\theta)} x^{\theta-1} e^{-x}$	$\Gamma(\theta)$
Lognormal	$\exp\{-\frac{1}{2}(\log(x)/\theta)^2\} / \{\theta x \sqrt{2\pi}\}$	LN(θ)
Exponential logarithmic	$\frac{1}{-\ln \theta} \frac{(1-\theta)e^{-x}}{(1-\theta e^{-x})^2}$	EL(θ)
Exponential geometric	$\frac{(1-\theta)e^{-x}}{(1-\theta e^{-x})^2}$	EG(θ)

TABLE 1 Alternative distributions

p	Distribution	KS	CM	AD	CO	$\widetilde{EP}$	$\widetilde{BH}_{0.5}$	$\widetilde{L}_{0.05}$
0.1	Exp(1)	4	5	5	5	5	5	5
0.2	Exp(1)	5	5	5	5	5	5	5
0.3	Exp(1)	5	5	5	5	5	5	5
0.4	Exp(1)	5	5	5	6	5	5	5
0.1	EG(0.2)	5	6	6	6	10	10	7
0.2	EG(0.2)	6	6	7	6	9	9	7
0.3	EG(0.2)	5	5	6	5	6	8	6
0.4	EG(0.2)	5	5	6	6	4	8	8
0.1	EG(0.5)	13	17	18	11	27	27	13
0.2	EG(0.5)	11	12	14	9	21	20	12
0.3	EG(0.5)	7	9	10	6	12	16	10
0.4	EG(0.5)	6	7	8	7	2	13	9
0.1	EG(0.8)	41	49	50	29	64	63	30
0.2	EG(0.8)	28	33	35	18	47	47	23
0.3	EG(0.8)	17	20	22	12	27	33	17
0.4	EG(0.8)	11	12	15	10	1	24	15
0.1	EL(0.2)	22	26	29	19	41	41	21
0.2	EL(0.2)	17	20	23	13	33	33	18
0.3	EL(0.2)	12	14	16	10	20	25	14
0.4	EL(0.2)	9	10	13	9	1	21	14
0.1	EL(0.5)	7	8	9	7	15	14	9
0.2	EL(0.5)	7	7	9	6	13	13	9
0.3	EL(0.5)	6	6	7	5	8	10	7
0.4	EL(0.5)	6	6	6	6	3	10	8
0.1	EL(0.8)	5	6	6	5	7	7	6
0.2	EL(0.8)	5	5	6	5	7	7	6
0.3	EL(0.8)	5	5	5	5	5	6	6
0.4	EL(0.8)	5	5	6	6	4	6	7
0.1	$\Gamma(0.4)$	93	95	99	99	98	99	99
0.2	$\Gamma(0.4)$	90	93	98	99	97	98	99
0.3	$\Gamma(0.4)$	85	89	97	98	93	96	98
0.4	$\Gamma(0.4)$	81	85	96	97	0	95	98
0.1	$\Gamma(0.7)$	24	29	41	46	45	47	52
0.2	$\Gamma(0.7)$	23	27	40	41	42	44	49
0.3	$\Gamma(0.7)$	18	23	35	37	31	40	46
0.4	$\Gamma(0.7)$	17	20	33	36	0	38	45
0.1	LN(1.5)	58	65	64	23	75	74	19
0.2	LN(1.5)	36	41	40	8	53	51	8
0.3	LN(1.5)	19	23	20	3	27	32	4
0.4	LN(1.5)	11	12	11	3	1	19	2

TABLE 2 Percentage of 10,000 independent Monte Carlo samples (rounded to the nearest integer) that resulted in the rejection of H_0 at 5% significance level, for $n = 40$ and $r = \lfloor np \rfloor$

Note that the highest estimated power for each alternative has been highlighted using bold font.

4.2 | Simulation setting and results for the model-based setup

Let $Y_1, Y_2, \dots, Y_n$ be the failure times simulated from a PH model with corresponding covariate $x_1, x_2, \dots, x_n$. Under Type II right censoring, we only observe $T_{(1)}, T_{(2)}, \dots, T_{(r)}$, where $T_{(j)} = Y_{(j)}$, $j = 1, 2, \dots, r$. We will define these observed values along with their corresponding covariate x_j (for this simulation, we assume a single covariate) by (T_j, x_j) , $j \in \{k : T_{(k)} \leq T_{(r)}\}$. Recall that the cumulative hazard rate (CHR) of the j^{th} individual follows a PH model with a single covariate if

$$\Lambda_j(t) = q(x_j, \beta)H(t),$$

TABLE 3 Percentage of 10,000 independent Monte Carlo samples (rounded to the nearest integer) that resulted in the rejection of $\mathcal{H}_0$ at 5% significance level, for $n = 80$ and $r = \lfloor np \rfloor$

p	Distribution	KS	CM	AD	CO	$\tilde{EP}$	$\tilde{BH}_{0.5}$	$\tilde{L}_{0.05}$
0.1	Exp(1)	5	5	5	5	5	5	5
0.2	Exp(1)	4	4	4	5	5	5	5
0.3	Exp(1)	5	5	5	5	5	5	5
0.4	Exp(1)	5	5	5	5	5	5	5
0.1	EG(0.2)	7	7	8	6	13	13	8
0.2	EG(0.2)	6	6	7	5	11	10	7
0.3	EG(0.2)	5	6	7	6	8	9	7
0.4	EG(0.2)	5	5	6	5	3	8	6
0.1	EG(0.5)	23	27	28	15	42	42	17
0.2	EG(0.5)	16	18	19	12	30	30	15
0.3	EG(0.5)	12	14	15	9	21	24	13
0.4	EG(0.5)	8	9	11	7	1	17	10
0.1	EG(0.8)	70	77	78	47	87	87	46
0.2	EG(0.8)	48	55	56	30	70	70	33
0.3	EG(0.8)	30	36	37	19	48	52	24
0.4	EG(0.8)	18	21	23	12	0	36	18
0.1	EL(0.2)	39	46	49	29	63	63	30
0.2	EL(0.2)	29	35	37	22	51	51	25
0.3	EL(0.2)	21	26	28	17	37	41	21
0.4	EL(0.2)	15	17	20	11	0	30	17
0.1	EL(0.5)	10	12	13	8	21	21	11
0.2	EL(0.5)	9	9	10	7	17	17	9
0.3	EL(0.5)	7	8	9	6	12	14	8
0.4	EL(0.5)	6	7	8	5	2	12	8
0.1	EL(0.8)	6	6	6	6	8	9	6
0.2	EL(0.8)	5	5	5	5	7	7	6
0.3	EL(0.8)	5	5	6	5	6	7	6
0.4	EL(0.8)	5	5	6	5	4	7	6
0.1	$\Gamma(0.4)$	100	100	100	100	100	100	100
0.2	$\Gamma(0.4)$	99	100	100	100	100	100	100
0.3	$\Gamma(0.4)$	99	100	100	100	100	100	100
0.4	$\Gamma(0.4)$	98	99	100	100	0	100	100
0.1	$\Gamma(0.7)$	46	52	67	70	70	72	73
0.2	$\Gamma(0.7)$	39	46	61	64	64	66	71
0.3	$\Gamma(0.7)$	36	43	58	61	55	62	68
0.4	$\Gamma(0.7)$	31	37	54	56	0	58	64
0.1	LN(1.5)	85	89	88	37	94	94	30
0.2	LN(1.5)	61	68	64	13	77	76	12
0.3	LN(1.5)	37	42	38	4	50	51	4
0.4	LN(1.5)	19	21	18	2	1	28	2

Note that the highest estimated power for each alternative has been highlighted using bold font.

TABLE 4 Percentage of 10,000 independent Monte Carlo samples (rounded to the nearest integer) that resulted in the rejection of $\mathcal{H}_0$ at 5% significance level, where $\mathcal{H}_0$ is Weibull and $\mathcal{H}_1$ is Weibull

n	p	KS	CM	AD	CO	$\tilde{EP}$	$\tilde{BH}_{0.5}$	$\tilde{L}_{0.05}$
100	0.1	4	5	3	6	4	5	5
	0.2	1	2	1	6	5	6	5
	0.3	1	1	1	6	5	6	4
	0.4	1	1	1	2	4	5	5
200	0.1	3	4	3	5	5	5	5
	0.2	1	2	1	6	5	5	5
	0.3	1	1	1	6	4	6	5
	0.4	1	1	1	3	4	6	5

where $H(\cdot)$ is some unspecified cumulative hazard function and $q(x_j, \beta)$ is a regression function that depends on x_j , the covariate of the j th individual, with β the unknown regression parameter. On the basis of the observations $(T_j, x_j), j \in \{k : T_{(k)} \leq T_{(r)}\}$, we wish to test the null hypothesis

$$\mathcal{H}_0 : H(t) = H_0(t; a, b), \quad (9)$$

n	p	KS	CM	AD	CO	$\widetilde{EP}$	$\widetilde{BH}_{0.5}$	$\widetilde{L}_{0.05}$
100	0.1	42	58	55	6	89	89	61
	0.2	6	27	19	3	80	80	46
	0.3	2	5	4	2	63	67	31
	0.4	1	1	1	6	0	52	22
200	0.1	72	90	91	7	100	100	96
	0.2	12	65	61	4	97	97	87
	0.3	2	19	20	3	91	92	72
	0.4	2	3	6	4	0	83	58

Note that the highest estimated power for each alternative has been highlighted using bold font.

TABLE 5 Percentage of 10,000 independent Monte Carlo samples (rounded to the nearest integer) that resulted in the rejection of H_0 at 5% significance level, where H_0 is Weibull and H_1 is lognormal

where $H_0(t; a, b) = \left(\frac{t}{b}\right)^a$ is the Weibull cumulative baseline hazard function with unknown parameters a and b . We can now, under H_0 , estimate the parameters β , a , and b by their maximum likelihood estimates $\hat{\beta}$, $\hat{a}$, and $\hat{b}$. On the basis of these estimators, we obtain the Cox–Snell residuals, defined as

$$\hat{\varepsilon}_j = e^{\hat{\beta}x_j} H_0(t_j; \hat{a}, \hat{b}), \quad j = 1, \dots, r.$$

If H_0 is true, then the Cox–Snell residuals should (approximately) follow a standard exponential distribution. Hence, any test for exponentiality on the basis of $\hat{\varepsilon}_j$, $j = 1, 2, \dots, r$ constitutes, in effect, a test for the proportional hazard model itself. In our limited Monte Carlo study, we use a single covariate, x_j , randomly generated from a uniform (0,1) distribution, with the regression function $q(x_j, \beta) = e^{\beta x_j}$, with $\beta = 0.1$. To simulate the lifetimes from a PH model, we made use of two baseline CHRs. The first was the Weibull CHR with $a = 1.1$ and $b = 0.8$ which corresponds to H_0 . The second was the lognormal CHR with associated density $f(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\{-(\log(x) - \mu)^2/2\sigma^2\}$ and with $\mu = 1.3$ and $\sigma = 1.5$, which constitutes an alternative to the Weibull baseline. Due to the fact that the limiting distribution of each of the test statistics is very complicated and depends on unknown parameters, we made use of the following parametric bootstrap algorithm to calculate the critical value of a test statistic, say $K := K(\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_r)$. We now illustrate the application of the bootstrap procedure in order to calculate the critical values of K under the PH model.

1. Generate a bootstrap sample of failure times $Y_1^*, Y_2^*, \dots, Y_n^*$, where

$$Y_j^* = H_0^{-1} \left[-\log(U_j) e^{-\hat{\beta}x_j}; \hat{a}, \hat{b} \right]$$

and where U_j , $j = 1, \dots, n$ are independent random variates from a uniform distribution on (0, 1). Denote by $Y_{(1)}^* < Y_{(2)}^* < \dots < Y_{(r)}^*$, the first r ordered values of $Y_1^*, Y_2^*, \dots, Y_n^*$.

2. Now, based on a Type II right censored sample of size r from these failure times, fit a PH model under H_0 using the observations (T_j^*, x_j) , $j \in \{k : T_k^* \leq T_{(r)}^*\}$, where $T_{(j)}^* = Y_{(j)}^*$, $j = 1, 2, \dots, r$.
3. Calculate the Cox–Snell residuals for the PH model fitted in step 2,

$$\hat{\varepsilon}_j^* = e^{\hat{\beta}^* x_j} H_0(T_j^*; \hat{a}^*, \hat{b}^*), \quad j = 1, 2, \dots, r,$$

where $\hat{a}^*$ and $\hat{b}^*$ are the estimators of the model parameters based on the censored bootstrap sample in Step 2.

4. Calculate the test statistic, say $K^* := K(\hat{\varepsilon}_1^*, \dots, \hat{\varepsilon}_r^*)$.
5. Repeat Steps 1 to 4 a number of times, say B , to obtain B copies of K^* denoted $K_1^*, K_2^*, \dots, K_B^*$.
6. Obtain the order statistics $K_{(1)}^* \leq K_{(2)}^* \leq \dots \leq K_{(B)}^*$.
7. The estimated critical value at an $100\alpha\%$ significance level is then given by $\hat{c}_n(\alpha) = K_{([B(1-\alpha)])}^*$.

Given that the above algorithm might be computationally demanding, we employ the warp-speed method proposed by Giacomini, Politis, and White (2013). With the warp-speed method, only one resample is generated for each Monte Carlo sample, and the bootstrap results are aggregated over the Monte Carlo replications.

We now turn your attention to the estimated sizes reported in Table 4, which were obtained from data generated from a PH model with a Weibull baseline distribution, and the estimated powers in Table 5, which employed data generated from a PH model with a lognormal baseline. Considering the results in Table 4, we note that the CO, $\widetilde{EP}$, $\widetilde{BH}$, and $\widetilde{L}$ tests produce sizes close to the nominal specified level of 5%. The EDF-based tests, on the other hand, all tend to underestimate the size, especially for large censoring proportions. From Table 5, we note that the $\widetilde{EP}$ and $\widetilde{BH}$ tests are the best performing tests, whereas CO is the worst performer. However, in the case of the $\widetilde{EP}$ test, it only produces reasonable results for lower censoring proportions, with a dramatic drop in performance for censoring proportions of 40% (or larger). One should note that, based on the results in this study, the EDF-based tests also do not perform well for large censoring proportions.

5 | CONCLUSION

In this study, we modified three tests for exponentiality on the basis of either the characteristic function or the Laplace transform to accommodate Type II right censored data and compared the power performance of our three newly modified test statistics with that of the existing Kolmogorov–Smirnov, Cramér–von Mises, Anderson–Darling, and Cox and Oakes test in an i.i.d. and model-based setup. Based on the results of the Monte Carlo study conducted in this paper, it was found that the three newly modified test statistics outperformed the other tests under all alternatives. The $\widetilde{BH}$ test based on the empirical Laplace transform is consistently a good performer. Although we only considered these three tests for illustrative purposes, methods discussed in this paper can easily be extended to the tests by, for example, Henze (1992), Henze and Meintanis (2002), Henze and Meintanis (2002), and Meintanis et al. (2014).

6 | DATA AVAILABILITY STATEMENT

The R code for data generation for this study is available from the corresponding author upon reasonable request.

ORCID

Marike Cockeran  <https://orcid.org/0000-0002-3990-8345>

REFERENCES

- Ahmad, I. A., & Alwasel, I. A. (1999). A goodness-of-fit test for exponentiality based on the memoryless property. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 61(3), 681–689. <https://doi.org/10.1111/1467-9868.00200>
- Allison, J. S., Santana, L., Smit, N., & Visagie, I. J. (2017). An 'apples to apples' comparison of various tests for exponentiality. *Computational Statistics*, 32(4), 1241–1283. <https://doi.org/10.1007/s00180-017-0733-3>
- Ascher, S. (1990). A survey of tests for exponentiality. *Communications in Statistics-Theory and Methods*, 19(5), 1811–1825. <https://doi.org/10.1080/03610929008830292>
- Balakrishnan, N., Chimitova, E., & Vedernikova, M. (2015). An empirical analysis of some nonparametric goodness-of-fit tests for censored data. *Communications in Statistics: Simulation and Computation*, 44(4), 1101–1115. <https://doi.org/10.1080/03610918.2013.796982>
- Baratpour, S., & Rad, A. H. (2012). Testing goodness-of-fit for exponential distribution based on cumulative residual entropy. *Communications in Statistics-Theory and Methods*, 41(8), 1387–1396. <https://doi.org/10.1080/03610926.2010.542857>
- Baringhaus, L., & Henze, N. (1991). A class of consistent tests for exponentiality based on the empirical Laplace transform. *Annals of the Institute of Statistical Mathematics*, 43(3), 551–564.
- Baringhaus, L., & Henze, N. (2000). Tests of fit for exponentiality based on a characterization via the mean residual life function. *Statistical Papers*, 41(2), 225–236. <https://doi.org/10.1007/BF02926105>
- Barr, D. R., & Davidson, T. (1973). A Kolmogorov–Smirnov test for censored samples. *Technometrics*, 15(4), 739–757.
- Breslow, N., & Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, 2(3), 437–453.
- Castro-Kuriss, C. (2011). On a goodness-of-fit test for censored data from a location-scale distribution with applications. *Chilean Journal of Statistics*, 2(1), 115–136.
- Castro-Kuriss, C., Kelmansky, D. M., Leiva, V., & Martinez, E. J. (2010). On a goodness-of-fit test for normality with unknown parameters and type-II censored data. *Journal of Applied Statistics*, 37(7), 1193–1211. <https://doi.org/10.1080/02664760902984626>
- Csorgo, S. (1984). Estimating characteristic functions under random censorship. *Theory of Probability and its Applications*, 28(3), 615–623. <https://doi.org/10.1137/1128059>
- D'Agostino, R. B., & Stephens, M. (1986). *Goodness-of-fit Techniques, Series Statistics: Textbooks and Monographs*, 68. New York: Marcel Dekker.
- Development Core Team, R (2016). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from <http://www.R-project.org>
- Dufour, R., & Maag, U. (1978). Distribution results for modified Kolmogorov–Smirnov statistics for truncated or censored samples. *Technometrics*, 20(1), 29–32.
- Efron, B. (1967). The two sample problem with censored data. In *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 4) (pp. 831–853). Berkeley, CA: University of California.
- Epps, T., & Pulley, L. (1986). A test of exponentiality vs. monotone-hazard alternatives derived from the empirical characteristic function. *Journal of the Royal Statistical Society: Series B: Methodological*, 48(2), 206–213. <https://doi.org/10.1111/j.2517-6161.1986.tb01403.x>
- Galambos, J., & Kotz, S. (1978). *Characterizations of Probability Distributions: A Unified Approach with an Emphasis on Exponential and Related Models*. Berlin: Springer-Verlag.
- Giacomini, R., Politis, D. N., & White, H. (2013). A warp-speed method for conducting Monte Carlo experiments involving bootstrap estimators. *Econometric Theory*, 29(3), 567–589. <https://doi.org/10.1017/S0266466612000655>
- Glen, A. G., & Foote, B. L. (2009). An inference methodology for life tests with complete samples or type-II right censoring. *IEEE Transactions on Reliability*, 58(4), 597–603. <https://doi.org/10.1109/TR.2009.2034287>
- Gnedenko, B. V., Belyayev, Y. K., & Solov'yev, A. D. (1969). *Mathematical Methods of Reliability Theory*. New York-London: Academic Press.
- Goldmann, C., Klar, B., & Meintanis, S. G. (2015). Data transformations and goodness-of-fit tests for type-II right censored samples. *Metrika*, 78(1), 59–83. <https://doi.org/10.1007/s00184-014-0490-z>

- Henze, N. (1992). A new flexible class of omnibus tests for exponentiality. *Communications in Statistics-Theory and Methods*, 22(1), 115–133.
- Henze, N., & Meintanis, S. G. (2002). Goodness-of-fit tests based on a new characterization of the exponential distribution. *Communications in Statistics-Theory and Methods*, 31(9), 1479–1497. <https://doi.org/10.1081/STA-120013007>
- Henze, N., & Meintanis, S. G. (2005). Recent and classical tests for exponentiality: a partial review with comparisons. *Metrika*, 61(1), 29–45. <https://doi.org/10.1007/S001840400322>
- Jammalamadaka, R. S., & Taufer, E. (2006). Use of mean residual life in testing departures from exponentiality. *Nonparametric Statistics*, 18(3), 277–292. <https://doi.org/10.1080/10485250600759454>
- Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457–481. <https://doi.org/10.1080/01621459.1958.10501452>
- Kim, N. (2012). Testing exponentiality based on EDF statistics for randomly censored data when the scale parameter is unknown. *Korean Journal of Applied Statistics*, 25(2), 311–319.
- Kotz, S., Balakrishnan, N., & Johnson, N. L. (2004). *Continuous Multivariate Distributions, Volume 1: Models and Applications*. New York: John Wiley & Sons.
- Koziol, J. A., & Green, S. B. (1976). A Cramer-von Mises statistic for randomly censored data. *Biometrika*, 63(3), 465–474.
- Meintanis, S. G., Swanepoel, J., & Allison, J. (2014). The probability weighted characteristic function and goodness-of-fit testing. *Journal of Statistical Planning and Inference*, 146(1), 122–132. <https://doi.org/10.1016/j.jspi.2013.09.011>
- Pettitt, A. N. (1976). Cramer-von Mises statistics for testing normality with censored samples. *Biometrika*, 63(3), 475–481.
- Pettitt, A. N. (1977). Tests for the exponential distribution with censored data using Cramer-von Mises statistics. *Biometrika*, 64(3), 629–632.
- Pettitt, A. N., & Stephens, M. A. (1976). Modified Cramer-von Mises statistics for censored data. *Biometrika*, 63(2), 291–298.
- Torabi, H., Narges, N. G., & Grane, A. (2018). A wide review on exponentiality tests and two competitive proposals with application on reliability. *Journal of Statistical Computation and Simulation*, 88(1), 108–139. <https://doi.org/10.1080/00949655.2017.1379522>
- Volkova, K. Y., & Nikitin, Y. Y. (2013). Exponentiality tests based on Ahsanullah's characterization and their efficiencies. *Journal of Mathematical Sciences*, 204(1), 42–54. <https://doi.org/10.1007/s10958-014-2185-4>
- Zardasht, V., Parsi, S., & Mousazadeh, M. (2015). On empirical cumulative residual entropy and a goodness-of-fit test for exponentiality. *Statistical Papers*, 56(3), 677–688. <https://doi.org/10.1007/s00362-014-0603-9>

How to cite this article: Cockeran M, Santana L, Allison J. Modified tests for exponentiality in the presence of Type II right censored data. *Stat.* 2019;8:e262. <https://doi.org/10.1002/sta4.262>

Chapter 5

Conclusion

5.1 Introduction

In this thesis we set out to conduct Monte Carlo simulation studies to evaluate a variety of tests for testing the goodness-of-fit (GOF) of parametric survival models (in particular the proportional hazard (PH) and accelerated failure time (AFT) models) and compare their performance to one another. Aspects of the models under consideration include the shape of the survival function and time-invariance of the covariates. In addition, interest lies in testing GOF of the proposed model against a more general model. These aspects are investigated by first evaluating the GOF of survival models in the complete sample case by considering various tests for exponentiality and then considering Type-II right censored data. The main objectives of the thesis can be summarised as follows:

- Develop new parametric bootstrap algorithms for approximating the null distribution of the various test statistics used to test the adequacy of the survival models.
- Evaluate the performance of various GOF tests in order to determine the adequacy of the PH model with respect to the three scenarios stated above in the complete samples case.
- Derive some theory relating to the consistency of one of the test statistics under consideration.
- Compare two different strategies that are often employed to conduct GOF testing in the presence of Type-II right censoring.
- Evaluate the performance of various GOF tests in order to determine the adequacy of PH and AFT models in the presence of Type-II right censoring.
- Suitably modify test statistics based on either the characteristic function or the Laplace transform to accommodate Type-II right censored data.

- Evaluate the effectiveness of the newly modified test statistics against existing modified tests.
- Make recommendations of which tests to use in which circumstances when testing the adequacy of a specific survival model.

5.2 Results

The results and conclusions of the three manuscripts prepared for this thesis are now presented:

Goodness-of-fit tests in the Cox proportional hazards model

In the first article we compared the performance of a variety of tests for exponentiality when testing GOF for the PH model in the complete sample case. The testing procedures considered included tests based on the empirical distribution function (EDF), tests based on empirical transforms, as well as a criterion based on a moment-type characterization of the exponential distribution. The performance of these test statistics when testing exponentiality was evaluated in a comprehensive simulation study in three distinct scenarios: (A) being able to distinguish between a number of alternative baseline hazards when using the PH model, (B) being able to distinguish between the PH model with a time-independent covariate against the same PH model, i.e. with the same baseline distribution, but with time-varying covariate, and (C) being able to distinguish between the PH and Hsieh models with the same baseline hazard. Based on the results of the Monte Carlo study, it was found that the test of Epps & Pulley (1986) outperformed the other tests for most alternatives. Although the Baringhaus & Henze (1991) test, based on the empirical Laplace transform, and the Henze & Meintanis (2005) test using the empirical characteristic function, did not achieve the highest powers, they were found to be consistently good performers. An analogous Monte Carlo study was carried out for AFT models and, since the performance of the tests under the AFT model was similar to the performance observed for the PH models, it was not reported in the article.

Goodness-of-fit testing of survival model in the presence of Type-II right censoring

The second article extended the study done in article one to consider GOF tests for parametric PH and AFT models in the presence of Type-II right censoring. We investigated the behavior of a number of exponentiality tests in a variety of settings. In particular, interest lay in the comparison of the two different strategies that are often employed to conduct these tests, namely (i) first transforming the censored sample to a complete sample and then using full sample versions of the test statistics, and (ii) the alternative where the statistics themselves are modified to accommodate Type-II right censoring. The test statistics considered under strategy (i) included tests based on the EDF, score function, characteristic function, Laplace transform, a criterion based on a moment-type characterization of the exponential distribution, as well as a test based on the mean

residual life. The modified statistics (strategy (ii)) included the Anderson-Darling and Cramér-von Mises EDF-based tests, as well as the corresponding modified test based on a score function. The performance of these test statistics when testing exponentiality was evaluated in a comprehensive simulation study in three distinct scenarios: (A) being able to distinguish between a number of alternative baseline hazards when using the PH model, (B) being able to distinguish between the PH and Hsieh models with the same baseline hazard, and (C) the ability to distinguish between the AFT model with a given baseline hazard against the same model with a different hazard rate. Based on the results of the Monte Carlo study, it was found that no single test dominated all other tests for all scenarios considered, but that the Epps & Pulley (1986) test based on the characteristic function, and the test of Henze & Meintanis (2005) based on the Laplace transform generally outperformed the other tests for *many* of the settings considered, while the test of Baringhaus & Henze (2000) based on the mean residual life is also consistently a good performer. Two transformations to complete samples were considered in this paper, namely those suggested by O'Reilly & Stephens (1988) and by Fischer & Kamps (2011). In the finite sample study it was found that the former transformation consistently produced higher powers than the latter for all scenarios investigated. Hence it seems reasonable to recommend the use of the O'Reilly & Stephens (1988) transformation if one opts for strategy (i) above. Based on the performance of strategies (i) and (ii) we found that with lower censoring proportions the difference in powers between the tests that make use of statistics modified to accommodate Type-II right censoring and the tests that transform the censored sample to a complete sample is relatively small. However, this difference becomes much more pronounced with higher censoring proportions. Specifically, the empirical powers for tests employing modified statistics display a rapid decline as the censoring proportions increase and consequently these tests have much lower powers for higher censoring proportions than the tests using transformations to complete samples. Therefore either of the two approaches is useful for conducting goodness-of-fit testing for exponentiality if the censoring proportion in a data set is lower than 30%, but we would recommend using the approach that first transforms the sample to a complete sample for higher censoring proportions.

Modified tests for exponentiality in the presence of Type-II right censoring

The third and final article presented a new way of modifying three well-known tests for exponentiality based on either the characteristic function or the Laplace transform to accommodate Type-II right censored data. These three tests are the tests of Epps & Pulley (1986), Baringhaus & Henze (1991), and Henze & Meintanis (2005). We compared the power of the three newly altered tests to that of the existing Kolmogorov-Smirnov, Cramér-von Mises, Anderson-Darling and Cox and Oakes test in an i.i.d. and model-based setup. Based on the results of the Monte Carlo study it was found that the three modified test statistics outperformed the other tests under all alternatives. The Baringhaus & Henze (1991) test based on the empirical Laplace transform is consistently one of the best performing tests.

5.3 Concluding remarks and future research

The overall aim of this thesis was to conduct Monte Carlo simulation studies to evaluate the performances of various GOF tests in order to determine the adequacy of the PH and AFT models in terms of (i) the parameterization of the baseline distribution, (ii) time-dependency of the covariates, (iii) as well as testing the GOF of the PH model against the more general Hsieh model. It was found that tests based on the characteristic function or the Laplace transform generally outperformed the other tests for *many* of the settings considered. In case of type-II right censoring we would recommend using one of the newly modified test statistics presented in the third article. We now conclude by presenting some avenues for future research. In all the tests based on either the characteristic function or the Laplace transform, a so-called tuning parameter appears. Allison & Santana (2015) proposed a data-dependent method for choosing the tuning parameter, which is only applicable to a class of distributions for which the null distribution of the test statistics is independent of unknown parameters. Obviously this is not the case in our model-based setup, and future research will entail to develop a data-dependent method for choosing this tuning parameter in this setup, such that the tests maintain the nominal significance level and produce reasonable high powers. In addition to the above-mentioned research problem, another challenging problem would be to develop methods to modify existing tests for exponentiality, based on the characteristic function or the Laplace transform, to accommodate random censoring, and to derive the necessary asymptotic results.

References

- Allison, J. & Santana, L. (2015). On a data-dependent choice of the tuning parameter appearing in certain goodness-of-fit tests, *Journal of Statistical Computation and Simulation* **85**(16): 3276–3288.
- Baringhaus, L. & Henze, N. (1991). A class of consistent tests for exponentiality based on the empirical Laplace transform, *Annals of the Institute of Statistical Mathematics* **43**(3): 551–564.
- Baringhaus, L. & Henze, N. (2000). Tests of fit for exponentiality based on a characterization via the mean residual life function, *Statistical Papers* **41**(2): 225–236.
- Epps, T. & Pulley, L. (1986). A test of exponentiality vs. monotone-hazard alternatives derived from the empirical characteristic function, *Journal of the Royal Statistical Society. Series B (Methodological)* **48**(2): 206–213.
- Fischer, T. & Kamps, U. (2011). On the existence of transformations preserving the structure of order statistics in lower dimensions, *Journal of Statistical Planning and Inference* **141**(1): 536–548.
- Henze, N. & Meintanis, S. G. (2005). Recent and classical tests for exponentiality: A partial review with comparisons, *Metrika* **61**(1): 29–45.
- O'Reilly, F. & Stephens, M. (1988). Transforming censored samples for testing fit, *Technometrics* **30**(1): 79–86.
- Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*, Wiley & Sons.