Proceedings of the 62nd Annual Conference of the South African Statistical Association for 2021

1 – 3 December 2021
Stellenbosch

Proceedings of the 62nd Annual Conference of the South African Statistical Association for 2021 (SASA 2021)

ISBN 978-0-86886-877-6

Editor

Sheetal Silal University of Cape Town

Assistant Editors

Allan Clark University of Cape Town
Justin Harvey Stellenbosch University
Andréhette Verster University of the Free State

Managing Editor

Charl Pretorius University of the Free State

Review Process

Eight (8) manuscripts were submitted for possible inclusion in the Proceedings of the 62nd Annual Conference of the South African Statistical Association, 2021. All submitted papers were assessed by the Editorial team for suitability, after which 7 papers were sent to be reviewed by at least two independent reviewers each. Papers were reviewed according to the following criteria: relevance to conference themes, relevance to audience, standard of writing, originality and critical analysis. Of the 8 submitted manuscripts, 3 were ultimately (after consideration and incorporation of reviewer comments) judged to be suitable for inclusion in the proceedings of the conference.

Reviewers

The editorial team would like to thank the following reviewers:

Renette Blignaut	University of the Western Cape
Tertius de Wet	University of Stellenbosch
Freedom Gumedze	University of Cape Town
Linda Haines	University of Cape Town
Francesca Little	University of Cape Town
Retha Luus	University of the Western Cape
Sibusisiwe Makhanya	IBM
Chioneso Marange	University of Fort Hare
Johané Nienkemper-Swanepoel	Stellenbosch University
Broderick Oluyede	Georgia Southern University
Sulaiman Salau	University of Cape Town
Leonard Santana	North-West University
Sean van der Merwe	University of the Free State
Michael von Maltitz	University of the Free State

Contact Information

Queries can be sent by email to the Managing Editor (managing.editor@sastat.org).

Table of Contents

On an omnibus test for the parametric proportional hazards model 1
J. S. Allison, E. Bothma, M. Smuts and I. J. H. Visagie

A note on the probabilistic determinant of independent generalised beta entries 9
J. Ferreira, A. Bekker, T. Botha and S. Makgai

A new fixed point characterisation based test for the Pareto distribution in the presence of random censoring 17
L. Ndwandwe, J. S. Allison and I. J. H. Visagie

Proceedings of the 62nd Annual Conference of SASA, pp. 1–8

ON AN OMNIBUS TEST FOR THE PARAMETRIC PROPORTIONAL HAZARDS MODEL

J. S. Allison, E. Bothma, M. Smuts and I. J. H. Visagie
Subject Group Statistics, North-West University, Potchefstroom, South Africa

We propose an omnibus test of fit for the parametric Cox proportional hazards model in the presence of random right censoring. The proposed test results from a modification of an existing test for the uniform distribution. This test is demonstrated to be able to detect deviations from the hypothesised model in two cases; first when the baseline distribution is misspecified and second when the regression component of the model is misspecified. Two modified classical tests are considered and a Monte Carlo study shows that the newly proposed test outperforms these tests for the majority of alternatives included. As a result of independent interest, we outline the procedure required to use the newly modified test in the framework of independent and identically distributed random variables.

Keywords: Kaplan-Meier estimate, Parametric bootstrap, Proportional hazard model, Uniformity.

1. Introduction

The parametric Cox proportional hazards (CPH) model is widely used in medical research as well as financial applications; see, for example, Bover et al. (1996), Klein and Moeschberger (2006) as well as Smuts and Allison (2020). In these fields it is often important to study the effect of explanatory variables on time-to-event outcomes. The parametric CPH model specifies the baseline survival function as well as the form of the regression function of the model. As a result the fitted model includes parameters which allow for a simple interpretation which is a desirable feature when using the model in practice. However, in order to ensure that the inference drawn from the fitted model is valid, one has to test the assumptions of the model. Testing the validity of the model requires simultaneously testing the assumption of the specified baseline distribution and the regression function included in the model. A testing procedure for this goodness-of-fit problem has been proposed for use with full samples in Cockeran et al. (2019).

Random right censoring frequently occurs in the fields mentioned above, meaning that the assumptions of the parametric CPH model should be tested based on censored data. Consider, for example, a medical study estimating the distribution of lifetimes of patients with a specific disease. For a given patient this distribution will likely depend on several covariates such as the patient's smoking habits. Some of the patients' lifetime's will be observed while other will not; for instance, one of the patients may relocate to another country and leave the study, meaning that this patient's lifetime is censored. In this case testing the goodness-of-fit of the parametric CPH model is complicated by the presence of the censoring mechanism.

Corresponding author: I. J. H. Visagie (jaco.visagie@nwu.ac.za)
MSC2020 subject classifications: 62F03, 62F40, 62N01

Lin and Spiekerman (1996) proposes a test for the baseline distribution of the parametric CPH model. However, there are situations in which one wishes to not only test the assumption of the baseline distribution, but the entire model simultaneously. In this case, it is possible to modify classical tests such as the Kolmogorov-Smirnov and Cramér-von Mises in order to test the fit of the entire model; this procedure is illustrated in this paper. In addition, we also propose a new test for the parametric CPH model based on a test for uniformity found in Meintanis (2009). The main idea of the test rests on fitting the CPH model and applying the inverse of the probability integral transform to obtain a censored sample of variables, the lifetime distribution of which is approximately uniform under the null hypothesis. These statements are made exact in Section 2. Furthermore, we modify the mentioned test to accommodate censoring when testing simple hypothesis or composite hypotheses in the classical goodness-of-fit framework of independent and identically distributed (i.i.d.) random variables; the required procedures are outlined later in the paper.

Some notation is introduced before proceeding. Let $Y_1, \ldots, Y_n$ be independent lifetime variables with continuous distribution and survival functions F and S respectively and let $C_1, \ldots, C_n$ be i.i.d. censoring variables with distribution function G, independent of $Y_1, \ldots, Y_n$. We assume non-informative censoring throughout. Below we denote the transpose of a vector v by $v^\top$. Let

$$T_j = \min(Y_j, C_j), \quad \delta_j = \begin{cases} 1, & \text{if } Y_j \leq C_j, \\ 0, & \text{if } Y_j > C_j, \end{cases} \quad \text{and} \quad \mathbf{x}_j = (x_{j,1}, \ldots, x_{j,m})^\top, j = 1, \ldots, n,$$

where $\mathbf{x}_j$ is a possible vector valued set of m realised covariates. Further let $\tilde{S}(t|\mathbf{x}_j)$ denote the conditional survival function given an observed set of covariates. Based on the observed triplets $(T_j, \delta_j, \mathbf{x}_j)$, $j = 1, \ldots, n$ we wish to test the composite hypothesis that the conditional survival function can be modelled by a parametric Cox model, i.e.

$$H_0 : \tilde{S}_{\theta,\beta}(t|\mathbf{x}_j) = S_\theta(t)^{e^{\beta^\top \mathbf{x}_j}}, \tag{1}$$

where S_θ is a specified parametric baseline survival function indexed by a parameter θ (possibly a vector) and $\beta = (\beta_1, \ldots, \beta_m)^\top$ denotes the vector of regression parameters associated with the covariates. Denote the order statistics of $Y_1, \ldots, Y_n$ and $T_1, \ldots, T_n$ by $Y_{(1)} < \cdots < Y_{(n)}$ and $T_{(1)} < \cdots < T_{(n)}$, respectively. Note that $\delta_{(j)}$ and $\mathbf{x}_{(j)}$ represent the indicator variable and set of covariates corresponding to $T_{(j)}$, respectively.

The remainder of the paper is structured as follows. In Section 2 we indicate how three existing tests are modified to accommodate random right censoring. Since the null distribution of each of the test statistics depends on the unknown censoring distribution, we propose a parametric bootstrap procedure in Section 2 in order to compute critical values for the tests under consideration. Section 3 contains the results of a Monte Carlo study where the empirical powers of the newly modified tests are studied and compared. This Monte Carlo study considers various deviations from the null hypothesis. Section 4 concludes and provides some directions for further research.

2. The proposed test

Under the null hypothesis in (1), $\tilde{S}_{\beta,\theta}(Y_i|\mathbf{x}_i)$ follows a standard uniform distribution (i.e. uniform on the interval (0,1)). This is a direct consequence of the probability integral transform. Let $\widehat{\beta}$ and

$\widehat{\theta}$ denote consistent estimators, under H_0, for β and θ, respectively (throughout the paper we make use of maximum likelihood estimation), whence $\tilde{S}_{\widehat{\beta},\widehat{\theta}}(Y_i|\mathbf{x}_i)$ should approximately follow a standard uniform distribution. Let

$$\widehat{U}_j = \tilde{S}_{\widehat{\beta},\widehat{\theta}}(T_j|\mathbf{x}_j) = S_{\widehat{\theta}}(T_j)^{e^{\widehat{\beta}^\top \mathbf{x}_j}}.$$

Under H_0, $(\widehat{U}_j, \delta_j)$ are censored observations for which the lifetime distribution is approximately standard uniform.

Above we have demonstrated that any test for censored uniformity on the basis of $(\widehat{U}_j, \delta_j)$ is in effect a test of the fit of the CPH model itself. Note that there are a number of misspecifications of the model in (1) that could lead to the rejection of the null hypothesis. These include an incorrect choice of the baseline distribution, a different form of the parametric regression function (e.g., $1 + \beta^\top \mathbf{x}$ instead of $e^{\beta^\top \mathbf{x}}$) and covariates dependent on time. Due to page restrictions, we investigate only the first two of these types of deviations in the Monte Carlo study in Section 3.

To test for uniformity we will use the Kolmogorov-Smirnov (KS_n) and Cramér-von Mises (CV_n) tests which are modified in Koziol and Green (1976) and Barr and Davidson (1973), respectively, for the random censoring case. In addition, we also modify a complete sample test for uniformity proposed in Meintanis (2009), where a test statistic based on the difference between the characteristic function (CF) of the standard uniform distribution and the empirical characteristic function is introduced. The proposed test statistic in the complete sample case is

$$R_{n,a} = n \int |\widehat{\varphi}_n(t) - \varphi_u(t)|^2 e^{-a|t|} dt, \tag{2}$$

where $\widehat{\varphi}_n(t) = \int_{-\infty}^{\infty} e^{itx} dF_n(x) = n^{-1} \sum_{j=1}^{n} e^{itY_j}$ is the empirical CF, $\varphi_u(t) = t^{-1}(\sin t + i(1 - \cos t))$ is the CF of the standard uniform distribution, $F_n(x)$ is the empirical distribution function, and $a > 0$ is a user defined tuning parameter. In the presence of random right censoring, we replace $\widehat{\varphi}_n(t)$ in (2) with

$$\widetilde{\varphi}_n(t) = \int_{-\infty}^{\infty} e^{itx} d\tilde{F}_n(x),$$

where $\tilde{F}_n(x)$ is the Kaplan-Meier estimate of the distribution F. After some straightforward algebra, the modified test statistic can be shown to have the calculable form

$$\tilde{R}_{n,a} = \sum_{j=1}^{n} \sum_{k=1}^{n} \Delta_j \Delta_k \frac{2a}{\left(\widehat{U}_j - \widehat{U}_k\right)^2 + a^2} + 4 \tan^{-1}\left(\frac{1}{a}\right) - 2a \log\left(1 + \frac{1}{a^2}\right) - 4 \sum_{j=1}^{n} \left[\tan^{-1}\left(\frac{\widehat{U}_j}{a}\right) + \tan^{-1}\left(\frac{1 - \widehat{U}_j}{a}\right)\right],$$

with

$$\Delta_j = \frac{\delta_{(j)}}{n} \prod_{k=1}^{j-1} \left[1 + \left(1 - \frac{\delta_{(k)}}{n-k}\right)\right], \; j = 1, \ldots, n-1, \quad \text{and} \quad \Delta_n = \prod_{j=1}^{n-1} \left(\frac{n-j}{n-j+1}\right),$$

where an empty product is understood to be 1. Note that Δ_j corresponds to the size of the jump of the Kaplan-Meier estimate of the distribution function at T_j. The test rejects for large values of $\tilde{R}_{n,a}$.

In the current context the null distributions of all test statistics depend on the unknown censoring distribution G. As a result, when approximating the null distribution of a test statistic, we should take care to estimate G and sample censoring times from this estimated distribution. As a result, we propose the bootstrap algorithm below which can be used to estimate the critical value of each of the tests. For ease of notation, let $Q := Q((\widehat{U}_1, \delta_1), \ldots, (\widehat{U}_n, \delta_n))$ be a generic test statistic.

1. Given $(T_j, \delta_j, \mathbf{x}_j)$, $j = 1, \ldots, n$, calculate $\widehat{\theta}$ and $\widehat{\beta}$ under H_0.
2. Calculate the Kaplan-Meier estimate, $\tilde{G}_n$, of the censored distribution function G.
3. Sample $C_1^*, \ldots, C_n^*$ from $\tilde{G}_n$.
4. Generate $Y_1^*, \ldots, Y_n^*$ from $\tilde{S}_{\widehat{\theta},\widehat{\beta}}$. This is accomplished by generating $U_1, \ldots, U_n$ from a standard uniform distribution and setting $Y_j^* = S^{-1}_{\widehat{\theta},\widehat{\beta}}(U_j^{\exp(-\widehat{\beta}^\top \mathbf{x}_j)})$, $j = 1, \ldots, n$.
5. Let $T_j^* = \min(Y_j^*, C_j^*)$ and $\delta_j^* = I(Y_j^* \leq C_j^*)$, $j = 1, \ldots, n$.
6. Calculate $\widehat{\theta}^*$ and $\widehat{\beta}^*$ based on $(T_j^*, \delta_j^*, \mathbf{x}_j^*)$, $j = 1, \ldots, n$, under H_0.
7. Let $\widehat{U}_j^* = \tilde{S}_{\widehat{\theta}^*,\widehat{\beta}^*}(T_j^* | \mathbf{x}_j)$, $j = 1, \ldots, n$.
8. Calculate the test statistic $Q^* = Q((\widehat{U}_1^*, \delta_1^*), \ldots, (\widehat{U}_n^*, \delta_n^*))$.
9. Repeat steps 3–8 B times to obtain $Q_1^*, \ldots, Q_B^*$. Let $Q_{(1)}^*, \ldots, Q_{(B)}^*$ denote the ordered values of the test statistics.
10. The estimated critical value is $\widehat{C}_{n,B} = Q^*_{(\lfloor B(1-\alpha) \rfloor)}$, with $\lfloor \cdot \rfloor$ denoting the floor function.

By simply removing all reference to the covariates $\mathbf{x}_j$, $j = 1, \ldots, n$, from the above algorithm, we obtain an algorithm which can be used to approximate critical values for the newly proposed test in the classical goodness-of-fit testing framework (in which observations are assumed to be realisations from i.i.d. random variables). In this framework, we can test a simple hypothesis by transforming the observed sample using the specified distribution function. In the case where the null hypothesis specifies a parametric family of distributions, we may estimate the necessary parameters and perform the requisite transformation using the estimated parametric distribution function (as is done above). In either case, if the null hypothesis is true, then we obtain censored realisations, the lifetime distribution of which should be approximately standard uniform and the testing procedure proposed above remains valid.

3. Monte Carlo simulation

In this section the finite sample performance of KS_n, CM_n and $\tilde{R}_{n,a}$ is analysed. The KS_n and CM_n tests are ubiquitous in the literature and we do not provide their details here, we merely note that these tests are obtained by replacing the empirical distribution function used in the classical goodness-of-fit testing framework by the Kaplan-Meier estimate.

The methodology described in the previous section is general and can be applied to test for any specified baseline distribution. However, due to the page limitations imposed on this publication, the

Table 1. Density functions of the baseline distributions.

Baseline distribution	**Density**	**Notation**
Weibull	$\theta\lambda(x\lambda)^{\theta-1}\exp(-(\lambda x)^{\theta})$	$W(\lambda,\theta)$
Gamma	$(\Gamma(\theta))^{-\theta}\lambda^{\theta}x^{\theta-1}\exp(-\lambda x)$	$\Gamma(\lambda,\theta)$
Lognormal	$(\theta x\sqrt{2\pi})^{-1}\exp(-\log^2(x-\lambda)(2\theta^2)^{-1})$	$LN(\mu,\sigma^2)$
Chi square	$(2^{\nu/2}\Gamma(\nu/2))^{-1}x^{\nu/2-1}\exp(-x/2)$	$\chi^2(\nu)$

results shown below are limited to the the case where baseline distribution is Weibull. This baseline is chosen due to its flexibility and its popularity in practical applications.

We present empirical sizes and powers of the tests for two scenarios below.

Scenario 1. The null hypothesis H_0 corresponds to the CPH model with a certain baseline distribution against an alternative H_A with a different baseline distribution, i.e., we consider the powers achieved by the various tests when the baseline distribution is misspecified while the regression component of the model is specified correctly.

Scenario 2. Under H_0 and H_A the model has the same baseline distribution, but the regression function under H_A is either $1+\beta^{\top}\mathbf{x}_j$ or $\log(1+\beta^{\top}\mathbf{x}_j)$, i.e., we consider the powers achieved when the baseline is correctly specified but the regression component of the model is misspecified.

3.1 Setting

The nominal significance level is set to 10%. We employ the so-called "warp-speed" bootstrap discussed in Giacomini et al. (2013) to obtain the empirical powers reported in this section. This methodology is popular for the calculation of empirical powers; see, for example, Allison et al. (2019). All size and power estimates are calculated based on 50 000 independent Monte Carlo samples for sample sizes $n = 100$ and $n = 200$. The reported powers are calculated in the case of 10% and 20% censoring. As censoring distribution we use a uniform distribution on the interval $(0, b)$, where b is a suitably chosen constant so as to achieve the desired censoring proportion (obtained using a simple optimisation algorithm).

Throughout the study we use a single continuous covariate x_j, randomly generated from a standard uniform distribution. The alternative distributions used for the baseline survival function are displayed in Table 1.

All calculations and simulations are performed in R, see R Core Team (2019). All parameter estimation is performed using maximum likelihood estimation. For this purpose we use Brent optimisation, see Brent (1971), built into R's *optim* function.

For ease of interpretation, we print the highest power (including ties) against each deviation from the null hypothesis in bold in each of the power tables presented in this section.

Scenario 1 above, the misspecification of the baseline distribution, is considered in Tables 2 and 3. Table 2 shows empirical sizes and powers when the censoring proportion is set to 10%, while Table 3 contains sizes and powers in the presence of 20% censoring. When comparing the powers in these two tables a number of remarks are in order. First, when considering the size of the test shown in the first line of these tables, we see that the tests typically reject in fewer instances than the 10% specified by the significance level, especially in the case where the sample size is 100. Second, as

Table 2. Empirical powers against misspecified baseline distributions with 10% censoring.

Distribution	KS_{100}	CV_{100}	$\tilde{R}_{100,.2}$	$\tilde{R}_{100,.5}$	KS_{200}	CV_{200}	$\tilde{R}_{200,.2}$	$\tilde{R}_{200,.5}$
$W(2,2)$	7	6	6	6	8	7	8	7
$\Gamma(1,3)$	**15**	13	**15**	14	**24**	22	**24**	**24**
$\chi^2(10)$	22	21	**23**	**23**	36	37	**39**	**39**
$LN(0.5,4)$	33	27	4	**36**	56	49	**73**	62

Table 3. Empirical powers against misspecified baseline distributions with 20% censoring.

Distribution	KS_{100}	CV_{100}	$\tilde{R}_{100,.2}$	$\tilde{R}_{100,.5}$	KS_{200}	CV_{200}	$\tilde{R}_{200,.2}$	$\tilde{R}_{200,.5}$
$W(2,2)$	8	7	8	6	10	9	10	8
$\Gamma(1,3)$	**11**	**11**	**11**	**11**	**17**	16	**17**	16
$\chi^2(10)$	15	**16**	**16**	15	24	26	**27**	25
$LN(0.5,4)$	16	8	**17**	13	28	13	**31**	24

Table 4. Empirical powers against misspecified regression function with 10% censoring.

Regression function	KS_{100}	CV_{100}	$\tilde{R}_{100,.2}$	$\tilde{R}_{100,.5}$	KS_{200}	CV_{200}	$\tilde{R}_{200,.2}$	$\tilde{R}_{200,.5}$
$1+x$	**18**	**18**	**18**	16	27	**29**	**29**	24
$1+3x$	18	**19**	**19**	17	28	**30**	**30**	25
$\log(1+x)$	12	**16**	13	13	15	**23**	17	18
$\log(1+3x)$	11	**15**	12	12	13	**21**	16	16

Table 5. Empirical powers against misspecified regression function with 20% censoring.

Regression function	KS_{100}	CV_{100}	$\tilde{R}_{100,.2}$	$\tilde{R}_{100,.5}$	KS_{200}	CV_{200}	$\tilde{R}_{200,.2}$	$\tilde{R}_{200,.5}$
$1+x$	26	23	**28**	19	45	40	**48**	32
$1+3x$	26	23	**29**	19	44	40	**48**	32
$\log(1+x)$	9	**11**	**11**	8	14	**19**	**19**	12
$\log(1+3x)$	9	11	**12**	8	15	18	**19**	12

expected, the powers of the tests generally increase with sample size and decrease with the censoring proportion. Third, although none of the tests outperform the others uniformly, $\tilde{R}_{n,0.2}$ and CV_n are clearly quite powerful when compared to their competitors.

Next we turn our attention to *Scenario 2* above, the misspecification of the regression component. This scenario is considered in Tables 4 and 5. Again, the results associated with 10% and 20% censoring are shown in two separate tables. As before, the powers of the tests generally increase with sample size and decrease with the censoring proportion. Again, $\tilde{R}_{n,0.2}$ and CV_n prove to be the most powerful tests. In addition to comparing the powers for different specifications of the regression component of the model, Tables 4 and 5 show powers associated with different choices for the value of β; specifically, this parameter is set either to 1 or 3. We note that the powers achieved against the different values of β are very similar for the cases considered.

4. Conclusion

The Monte Carlo study presented above shows that the CV_n and $\tilde{R}_{n,0.2}$ tests are especially powerful. When considering the empirical powers associated with a sample size of 100, CV_n outperforms the competing tests, while the $\tilde{R}_{n,0.2}$ is more powerful when the sample size is increased to 200. The powers associated with the smaller of these sample sizes do not differ much between the mentioned tests, while there is a pronounced difference in the case of the larger sample size. As a result, we recommend using $\tilde{R}_{n,a}$ for practical purposes.

For practical implementation of the new test one needs a so-called compromise choice of the tuning parameter a. This is a choice of the tuning parameter that provides reasonably high power for a wide range of alternatives. Based on the limited Monte Carlo study presented we suggest using $a = 0.2$. However, a more comprehensive power study is needed including different baseline distributions as well as other forms of misspecification in order to fully investigate the role of the parameter a.

The asymptotic properties of the new tests are unknown and should be derived. Specifically one needs to show that the test is consistent against a wide class of alternatives (or misspecifications). Furthermore, deriving the asymptotic null distributions of the test statistics for given censoring distributions are unknown at present. The above are matters of ongoing research.

References

Allison, J. S., Betsch, S., Ebner, B., and Visagie, I. J. H. (2019). New weighted L^2-type tests for the inverse Gaussian distribution. *arXiv:1910.14119*.

Barr, D. R. and Davidson, T. (1973). A Kolmogorov-Smirnov test for censored samples. *Technometrics*, **15**, 739–757.

Bover, O., Arellano, M., and Bentolila, S. (1996). Unemployment duration, benefit duration, and the business cycle. *In Estudios Económicos*, volume 57. Banco de España – Servicio de Estudios, Madrid.

Brent, R. P. (1971). An algorithm with guaranteed convergence for finding a zero of a function. *The Computer Journal*, **14**, 422–425.

Cockeran, M., Meintanis, S. G., and Allison, J. S. (2019). Goodness-of-fit tests in the Cox proportional hazards model. *Communications in Statistics – Simulation and Computation*, 1–12.

Giacomini, R., Politis, D. N., and White, H. (2013). A warp-speed method for conducting Monte Carlo experiments involving bootstrap estimators. *Econometric Theory*, **29**, 567–589.

Klein, J. P. and Moeschberger, M. L. (2006). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York.

Koziol, J. A. and Green, S. B. (1976). A Cramér-von Mises statistic for randomly censored data. *Biometrika*, **63**, 465–474.

Lin, D. and Spiekerman, C. (1996). Model checking techniques for parametric regression with censored data. *Scandinavian Journal of Statistics*, **23**, 157–177.

Meintanis, S. G. (2009). Goodness-of-fit tests and minimum distance estimation via optimal transformation to uniformity. *Journal of Statistical Planning and Inference*, **139**, 100–108.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Smuts, M. and Allison, J. (2020). An overview of survival analysis with an application in the credit risk environment. *ORiON*, **36**.

Proceedings of the 62nd Annual Conference of SASA, pp. 9–15

A NOTE ON THE PROBABILISTIC DETERMINANT OF INDEPENDENT GENERALISED BETA ENTRIES

Johan Ferreira, Andriëtte Bekker, Tanita Botha and Seite Makgai

Department of Statistics, University of Pretoria, Pretoria, South Africa
Centre of Excellence in Mathematical and Statistical Sciences, South Africa

Random determinants play an essential role within multivariate analysis, but their distributions often present theoretical and computational challenges. To circumvent these challenges, this note investigates two lower bounds for the probabilistic analysis of the determinant emanating from a matrix consisting of independent but not necessarily identically distributed generalised beta entries. The 2×2 and 3×3 cases receive particular attention, and a brief simulation study discusses the results.

Keywords: Chebyshev, Determinant, Generalised beta, Inference, Moments, Vysochanksij-Petunin.

1. Introduction

The behaviour of a random determinant has been studied to some extent to date. Early theoretical literature with this specific focus include Alagar (1978), Williamson and Downs (1988), Dembo (1989), and Wise and Hall (1991), with Saha and Chakraborty (2021) as a more recent contribution. This continued open problem in distribution theory was already under consideration as far back as the 1950's with the work of Nyquist et al. (1954) and Nicholson (1958). In particular, the investigation of the behaviour of random determinants is essential in geometric probability as well as wireless communication systems (see Alagar, 1978). In these cases, specific choices for the underlying distribution were used to statistically infer on the probabilistic behaviour of the determinant when a matrix with independent and identically distributed (i.i.d.) entries of these specific choices was considered. Considered models include the normal, exponential, gamma, and Weibull candidates (see Alagar, 1978, and Saha and Chakraborty, 2021). Furthermore, these and similar theoretical considerations have enjoyed applications in chemometrics and wireless communications, as illustrated by Aris and Mah (1963), Kim et al. (2010), Li et al. (2016), and Jabloun (2017) more recently.

For matrices of arbitrary dimension, the analytic consideration of the determinant is challenging and cumbersome. In this preliminary work, we report on the 2×2 and 3×3 cases when the matrix entries are independent generalised beta variates. This generalised beta distribution, proposed by McDonald and Xu (1995), has been considered extensively in literature in a variety of applications due to its attractive analytical and computational features, as well as its analytical form which includes a wealth of often considered models in the statistical arena. These include the generalised gamma, Pareto, usual beta, inverse beta, log-normal, and the usual normal model. The theoretical

Corresponding author: Johan Ferreira (johan.ferreira@up.ac.za)
MSC2020 subject classifications: 62E15, 62P12, 60E15

consideration of this generalised beta as an underlying choice when one is interested in the inferential aspects of the determinant from a matrix with these elements, is thus meaningful and unifying in the study of random determinants with independent entries within the current published literature.

Our direct interest lies in investigating aspects of the determinant $D = g(a_{ij})$, when the a_{ij} are independent but not necessarily identically distributed generalised beta variates, and $g(\cdot)$ simply denotes the analytical form of the expansion for the determinant in terms of the a_{ij}. In particular, we determine bounds for D and subsequently determine lower bounds for the probability of D being observed within these bounds using Chebyshev's inequality as well as the Vysochanksij-Petunin inequality. This focus using probabilistic inequalities circumvents the need for exact distributional derivations, many of which often result in inelegant theoretical representations (see Alagar, 1978) or approximations to the exact distributions of products of independent variates and their linear combinations (see Marques et al., 2021). Furthermore, the purposeful consideration of these two inequalities aim to illustrate the known "loose" inequality that Chebyshev offers, and to observe a tighter bound which may be more meaningful than Chebyshev in practice, that of Vysochanksij-Petunin. Williamson and Downs (1987) specifically mentions the challenge in computing explicit analytical forms of functions of random variables, and the meaningful consideration of lower order moments for probabilistic analysis. Furthermore, the consideration of the generalised beta positions the user uniquely to consider any of the submodels that this distribution includes; and so, this note generalises the work of Saha and Chakraborty (2021) in addition to considering non-identically distributed variates.

In Section 2, relevant properties of the generalised beta distribution are reviewed together with Chebyshev's and Vysochanksij-Petunin's inequalities. Section 3 contains a numerical illustration to discuss the theoretical arguments, and Section 4 wraps up with a brief discussion and final thoughts.

2. Generalised beta as candidate and the determinant

Let $a_{ij} = y$. The density of the generalised beta distribution is given by (see McDonald and Xu, 1995)

$$f(y) = \frac{|z|y^{zp-1}(1-(1-v)(\frac{y}{m})^z)^{q-1}}{m^{zp}B(p,q)(1+v(\frac{y}{m})^z)^{p+q}}, \tag{1}$$

for $0 < y^z < m^z/(1-v)$ and 0 otherwise, $0 \leq v \leq 1$, and z, m, p, q positive real numbers. Here, $B(\cdot,\cdot)$ denotes the usual beta function. The moments of this distribution are given by

$$E(Y^h) = \frac{m^h B(p+\frac{h}{z},q)}{B(p,q)} {}_2F_1\left(p+\frac{h}{z},\frac{h}{z};p+q+\frac{h}{z};v\right), \tag{2}$$

where ${}_2F_1(\cdot)$ denotes the Gauss hypergeometric function. The first and second moments are of immediate interest:

$$E(Y) = \frac{mB(p+\frac{1}{z},q)}{B(p,q)} {}_2F_1\left(p+\frac{1}{z},\frac{1}{z};p+q+\frac{1}{z};v\right) = \mu_1 \tag{3}$$

and

$$E(Y^2) = \frac{m^2 B(p+\frac{2}{z},q)}{B(p,q)} {}_2F_1\left(p+\frac{2}{z},\frac{2}{z};p+q+\frac{2}{z};v\right) = \mu_2. \tag{4}$$

Chebyshev's inequality provides a lower bound to investigate the probabilistic behaviour of a random variable, in this case D (Bain and Engelhardt, 1987):

$$P(|D - E(D)| \leq rSD(D)) = P(E(D) - rSD(D) < D < E(D) + rSD(D)) \geq 1 - \frac{1}{r^2}, \quad (5)$$

where (2) is used directly to determine $E(D)$ and $SD(D)$, which denote the mean and standard deviation of D, respectively. In (5), r denotes the rth standard deviation from the mean. The Vysochanksij-Petunin lower bound (see Klyushin et al., 2002) is given by

$$P(|D - E(D)| \leq \lambda SD(D)) = P(E(D) - \lambda SD(D) < D < E(D) + \lambda SD(D)) \geq 1 - \frac{4}{9\lambda^2} \quad (6)$$

and serves as a refinement of (5) which potentially addresses the lack of tightness of (5), via the inclusion of a scaling factor $\frac{4}{9}$ in the bound directly. In this case, λ plays a similar role as r in (5), and note that $\lambda \geq \sqrt{8/3}$. Subsequently, $var(D)$ denotes the variance operator.

2.1 For the 2×2 case

The determinant for the 2×2 case can be written as

$$D = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc \quad (7)$$

and thus, when a, b, c, d are independent random generalised beta variates:

$$\begin{aligned} E(D) &= E(ad - bc) \\ &= E(a)E(d) - E(b)E(c) \\ &= \mu_{1,a}\mu_{1,d} - \mu_{1,b}\mu_{1,c}\,, \end{aligned}$$

where $\mu_{1,a}$ denotes the first moment (3) for variable a and similarly for the others, and

$$\begin{aligned} var(D) &= var(ad - bc) \\ &= E(ad - bc)^2 - (E(a)E(d) - E(b)E(c))^2 \\ &= E(a^2)E(d^2) + E(b^2)E(c^2) - 2E(a)E(d)E(b)E(c) - (E(a)E(d) - E(b)E(c))^2 \\ &= \mu_{2,a}\mu_{2,d} + \mu_{2,b}\mu_{2,c} - 2\mu_{1,a}\mu_{1,d}\mu_{1,b}\mu_{1,c} - \mu_{1,a}^2\mu_{1,d}^2 + 2\mu_{1,a}\mu_{1,d}\mu_{1,b}\mu_{1,c} - \mu_{1,b}^2\mu_{1,c}^2 \\ &= \mu_{2,a}\mu_{2,d} + \mu_{2,b}\mu_{2,c} - \mu_{1,a}^2\mu_{1,d}^2 - \mu_{1,b}^2\mu_{1,c}^2\,, \end{aligned} \quad (8)$$

where $\mu_{2,a}$ denotes the second moment (4) for variable a and similarly for the others. It follows from (8) that

$$SD(D) = \sqrt{\mu_{2,a}\mu_{2,d} + \mu_{2,b}\mu_{2,c} - \mu_{1,a}^2\mu_{1,d}^2 - \mu_{1,b}^2\mu_{1,c}^2}.$$

In the case where the variables are identically distributed, this simplifies to

$$SD(D) = \sqrt{2\mu_2^2 - 2\mu_1^4}.$$

These expressions are used in (5) for an inferential investigation into the behaviour of D for specific choices of the parameters of (1) for the independent distributed entries of D.

2.2 For the 3×3 case

The determinant for the 3×3 case can be written as

$$D = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei - afh - bdi + bfg + cdh - ceg \tag{9}$$

and thus

$$E(D) = \mu_{1,a}\mu_{1,e}\mu_{1,i} - \mu_{1,a}\mu_{1,f}\mu_{1,h} - \mu_{1,b}\mu_{1,d}\mu_{1,i} + \mu_{1,b}\mu_{1,f}\mu_{1,g} + \mu_{1,c}\mu_{1,d}\mu_{1,h} - \mu_{1,c}\mu_{1,e}\mu_{1,g}$$

as before, and

$$\begin{aligned}
&var(D) \\
&= var(aei - afh - bdi + bfg + cdh - ceg) \\
&= E(aei - afh - bdi + bfg + cdh - ceg)^2 - (E(D))^2 \\
&= \mu_{2,a}\mu_{2,e}\mu_{2,i} + \mu_{2,a}\mu_{2,f}\mu_{2,h} + \mu_{2,b}\mu_{2,d}\mu_{2,i} + \mu_{2,b}\mu_{2,f}\mu_{2,g} + \mu_{2,c}\mu_{2,d}\mu_{2,h} + \mu_{2,c}\mu_{2,e}\mu_{2,g} \\
&\quad - \mu_{2,a}\mu_{1,e}\mu_{1,f}\mu_{1,h}\mu_{1,i} - \mu_{2,i}\mu_{1,a}\mu_{1,b}\mu_{1,d}\mu_{1,e} - \mu_{2,e}\mu_{1,a}\mu_{1,c}\mu_{1,g}\mu_{1,i} - \mu_{2,a}\mu_{1,e}\mu_{1,f}\mu_{1,h}\mu_{1,i} \\
&\quad - \mu_{2,f}\mu_{1,a}\mu_{1,b}\mu_{1,g}\mu_{1,h} - \mu_{2,h}\mu_{1,a}\mu_{1,c}\mu_{1,d}\mu_{1,f} - \mu_{2,i}\mu_{1,a}\mu_{1,b}\mu_{1,d}\mu_{1,e} - \mu_{2,b}\mu_{1,d}\mu_{1,f}\mu_{1,g}\mu_{1,i} \\
&\quad - \mu_{2,d}\mu_{1,b}\mu_{1,c}\mu_{1,h}\mu_{1,i} - \mu_{2,f}\mu_{1,a}\mu_{1,b}\mu_{1,g}\mu_{1,h} - \mu_{2,b}\mu_{1,d}\mu_{1,f}\mu_{1,g}\mu_{1,i} - \mu_{2,g}\mu_{1,b}\mu_{1,c}\mu_{1,e}\mu_{1,f} \\
&\quad - \mu_{2,h}\mu_{1,a}\mu_{1,c}\mu_{1,d}\mu_{1,f} - \mu_{2,d}\mu_{1,b}\mu_{1,c}\mu_{1,h}\mu_{1,i} - \mu_{2,c}\mu_{1,d}\mu_{1,e}\mu_{1,g}\mu_{1,h} - \mu_{2,e}\mu_{1,a}\mu_{1,c}\mu_{1,g}\mu_{1,i} \\
&\quad - \mu_{2,g}\mu_{1,b}\mu_{1,c}\mu_{1,e}\mu_{1,f} - \mu_{2,c}\mu_{1,d}\mu_{1,e}\mu_{1,g}\mu_{1,h} \\
&\quad + \mu_{1,a}\mu_{1,b}\mu_{1,e}\mu_{1,f}\mu_{1,g}\mu_{1,i} + \mu_{1,a}\mu_{1,c}\mu_{1,d}\mu_{1,e}\mu_{1,h}\mu_{1,i} + \mu_{1,a}\mu_{1,b}\mu_{1,d}\mu_{1,f}\mu_{1,h}\mu_{1,i} \\
&\quad + \mu_{1,a}\mu_{1,c}\mu_{1,e}\mu_{1,f}\mu_{1,g}\mu_{1,h} + \mu_{1,a}\mu_{1,b}\mu_{1,d}\mu_{1,f}\mu_{1,h}\mu_{1,i} + \mu_{1,b}\mu_{1,c}\mu_{1,d}\mu_{1,e}\mu_{1,g}\mu_{1,i} \\
&\quad + \mu_{1,a}\mu_{1,b}\mu_{1,e}\mu_{1,f}\mu_{1,g}\mu_{1,i} + \mu_{1,b}\mu_{1,c}\mu_{1,d}\mu_{1,f}\mu_{1,g}\mu_{1,h} + \mu_{1,a}\mu_{1,c}\mu_{1,d}\mu_{1,e}\mu_{1,h}\mu_{1,i}. \\
&\quad + \mu_{1,b}\mu_{1,c}\mu_{1,d}\mu_{1,f}\mu_{1,g}\mu_{1,h} + \mu_{1,a}\mu_{1,c}\mu_{1,e}\mu_{1,f}\mu_{1,g}\mu_{1,h} + \mu_{1,b}\mu_{1,c}\mu_{1,d}\mu_{1,f}\mu_{1,g}\mu_{1,i} \\
&\quad - (\mu_{1,a}\mu_{1,e}\mu_{1,i} - \mu_{1,a}\mu_{1,f}\mu_{1,h} - \mu_{1,b}\mu_{1,d}\mu_{1,i} + \mu_{1,b}\mu_{1,f}\mu_{1,g} + \mu_{1,c}\mu_{1,d}\mu_{1,h} - \mu_{1,c}\mu_{1,e}\mu_{1,g})^2.
\end{aligned} \tag{10}$$

We thus have that $SD(D)$ from (10) can be written as

$$SD(D) = \sqrt{var(D)},$$

In the case where the variables are identically distributed, note that

$$SD(D) = \sqrt{6\mu_2^3 - 18\mu_2\mu_1^4 + 12\mu_1^6}.$$

These expressions are used in (5) for an inferential investigation into the behaviour of D for specific choices of the parameters of (1) for the independent distributed entries of D.

3. A numerical illustration

For the 2×2 case, we follow the algorithm below to numerically investigate the behaviour of D by using (5) under the assumption of identically distributed elements:

1. Initialise parameters p, q, z, m, v for particular choices.
2. Simulate the entries for the 2×2 case from the distribution with density (1).
3. Calculate the determinant of each simulated matrix.
4. Calculate the theoretical bounds for (5) and (6) for the particular parameter choices in Step 1.
5. Determine whether the (empirical) determinant in Step 3 lies within the (theoretical) bounds of Step 4.
6. Repeat this process n times.
7. Determine the number of times that D was found to be within the bounds of (5) and (6) (i.e. empirical probability of inclusion) and the corresponding lower bound(s) on the probability.

In this way, the empirical behaviour of D can be juxtaposed and judged according to the theoretical bounds under consideration in this note. We investigate certain models contained in (1) (see McDonald and Xu, 1995) for illustrative purposes, namely the generalised beta of the first kind ($p = 3, q = 1, z = 2, m = 2.5, v = 0$ – Model 1), the power ($p = 3, q = 1, z = 1, m = 2.5, v = 0$ – Model 2), and the generalised gamma ($p = 3, q \to \infty, z = 2, m = 2.5, v = 0$ – Model 3), to observe that the bound does indeed hold from Chebyshev's- and Vysochanksij-Petunin's inequality. In the case where 1.00 is observed, this indicates that across the entire simulated sample all values of D were indeed found to be between the bounds of (5) for the corresponding value of r. Here, the bound is given by $1 - r^{-2}$ (*Bound 1*) and $1 - \frac{4}{9}\lambda^{-2}$ (*Bound 2*) respectively for different values of r and λ, and for different samples sizes.

This numerical illustration is done using the R software, v. 4.1.0. In Table 1 for each model, the bound holds across different values of r and λ, and also for the considered range of sample sizes. Bain and Engelhardt (1987) mention that Chebyshev's bound (see (5)) is not necessarily tight, and it is evident that the bound of (6) may provide the practitioner with more detailed insight into the probabilistic analysis of D with a tighter lower bound. It is a direct consequence of the construction of these bounds that Vysochanksij-Petunin's bound will always be tigher than that of Chebyshev's. However, the sustained relevance of Chebyshev's bound (5) is highlighted by the fact that (6) retains a restriction on choice of λ, whereas there is no distinct restriction on r – as is illustrated for the case when $r, \lambda = 1.5$.

Table 1. Empirical probabilities of (5) and (6) for different considered sub-models of (1), different values of r and λ, and different sample sizes.

	$n = 50$			$n = 500$			$n = 5000$				
r, λ	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Bound 1	Bound 2
1.5	0.92	0.84	0.86	0.880	0.880	0.894	0.8619	0.8586	0.8664	0.5556	N/A
1.75	1.00	0.98	1.00	0.992	0.986	0.994	0.9884	0.9892	0.9866	0.6913	0.8549
2	1.00	1.00	1.00	1.000	1.000	1.000	0.9994	0.9998	0.9990	0.7500	0.8889
2.2	1.00	1.00	1.00	1.000	1.000	1.000	1.0000	1.0000	1.0000	0.7934	0.9122

4. Discussion and future directions

The importance of analysing a determinant probabilistically has been illustrated within multivariate analysis. In this note, the generalised beta distribution of McDonald and Xu (1995) is considered, and expressions derived under the assumption of independence but allowing for elements to not be identically distributed for both the 2×2 and the 3×3 cases. A simulation study indicates that both lower bounds do indeed hold for the considered underlying model (1), and particular consideration of Vysochanksij-Petunin's inequality for this probabilistic analysis of the determinant illustrated due to its tighter nature, compared to the often considered Chebyshev's inequality. Specific interest in the implementation of such tighter bounds would be of meaning in the calculation of very tight capacity bounds for multiple-input-multiple-output systems in communications systems, which 1) relies on the determinant of random matrices for inferential purposes, 2) requires a tight bound for implementation such that complex expressions of the capacity (which is a determinant) may be approximated or even substituted with the lower bounds for practical reasons and execution, and 3) regularly relies on the description and analysis of 2×2 and 3×3 matrices as well as the practical assumption of independence (Zhang et al., 2005). Subsequent work includes expanding this computational aspect to matrices of higher dimension, and to consider dependence between elements.

Acknowledgements. The authors acknowledge the support of the Department of Statistics at the University of Pretoria, Pretoria, South Africa. This work enjoys support from the following grants: RDP296/2021 at the University of Pretoria; NRF ref. SRUG190308422768 nr. 120839; the DST-NRF South African Research Chair Initiative in Biostatistics UID: 71199; as well as the Centre of Excellence in Mathematical and Statistical Sciences at the University of the Witwatersrand. Finally, we acknowledge two anonymous reviewers for their constructive and critical feedback on this manuscript.

References

Alagar, V. (1978). The distribution of random determinants. *Canadian Journal of Statistics*, **6**, 1–9.

Aris, R. and Mah, R. (1963). Independence of chemical reactions. *Industrial & Engineering Chemistry Fundamentals*, **2**, 90–94.

Bain, L. J. and Engelhardt, M. (1987). *Introduction to Probability and Mathematical Statistics*. Brooks Cole.

Dembo, A. (1989). On random determinants. *Quarterly of Applied Mathematics*, **47**, 185–195.

Jabloun, M. (2017). A new generalization of the discrete Teager-Kaiser energy operator-application to biomedical signals. *In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4153–4157.

Kim, S. T., Choi, J., Beck, S., Song, T., Lim, K., and Laskar, J. (2010). Subthreshold current mode matrix determinant computation for analog signal processing. *In Proceedings of 2010 IEEE International Symposium on Circuits and Systems*. IEEE, 1260–1263.

Klyushin, D. A., Petunin, Y. I., Andrushkiw, R. I., Boroday, N. V., and Karelia, P. G. (2002). Cancer diagnostic method based on pattern recognition of DNA changes in buccal epithelium in

the pathology of the thyroid and mammary glands. *Center for Applied Mathematics and Statistics*. URL: *https://m.njit.edu/CAMS/Technical_Reports/CAMS02_03/report4.pdf*

Li, Y., Ding, S., Tan, B., Li, Z., and Zhao, H. (2016). Nonnegative sparse representation based on the determinant measure. *In 2016 IEEE International Conference on Digital Signal Processing (DSP)*. IEEE, 599–603.

Marques, F., Ghosh, I., Ferreira, J., and Bekker, A. (2021). A note on the product of independent beta random variables. *Contributions of Barry C. Arnold to Statistical Science – Theory and Applications*, 69–83.

McDonald, J. B. and Xu, Y. J. (1995). A generalization of the beta distribution with applications. *Journal of Econometrics*, **66**, 133–152.

Nicholson, W. L. (1958). On the distribution of 2×2 random normal determinants. *Annals of Mathematical Statistics*, **29**, 575–580.

Nyquist, H., Rice, S., and Riordan, J. (1954). The distribution of random determinants. *Quarterly of Applied Mathematics*, **12**, 97–104.

Saha, N. and Chakraborty, S. (2021). Using Chebyshev's inequality to predict a random determinant for iid gamma and Weibull elements. *Journal of Statistics and Management Systems*, **24**, 613–623.

Williamson, R. C. and Downs, T. (1987). Probabilistic arithmetic and the distribution of functions of random variables. *In 1st IASTED International Symposium on Signal Processing and its Applications*. Institution of Engineers, Australia, p. 112.

Williamson, R. C. and Downs, T. (1988). The inverse and determinant of a 2×2 uniformly distributed random matrix. *Statistics & Probability Letters*, **7**, 167–170.

Wise, G. L. and Hall, E. B. (1991). A note on the distribution of the determinant of a random matrix. *Statistics & Probability Letters*, **11**, 147–148.

Zhang, Q., Cui, X., and Li, X. (2005). Very tight capacity bounds for MIMO-correlated Rayleigh-fading channels. *IEEE Transactions on Wireless Communications*, **4**, 681–688.

Proceedings of the 62nd Annual Conference of SASA, pp. 17–23

A NEW FIXED POINT CHARACTERISATION BASED TEST FOR THE PARETO DISTRIBUTION IN THE PRESENCE OF RANDOM CENSORING

L. Ndwandwe, J. S. Allison and I. J. H. Visagie
School of Mathematical and Statistical Sciences, North-West University, South Africa

We propose a new goodness-of-fit test for the Pareto Type I lifetime distribution in the presence of random right censoring. The test is based on a fixed point characterisation, which is a generalisation of the well-known Stein method for the approximation of distributions. The empirical power performance of the new test is compared to the modified Cramér-von Mises and Kolmogorov-Smirnov tests for two different censoring proportions and two alternative lifetime distributions by means of a limited Monte Carlo study.

Keywords: Fixed point characterisation, Goodness-of-fit testing, Pareto distribution, Random censoring.

1. Introduction

The Pareto distribution, nowadays commonly known as the Pareto Type I distribution, was first introduced by Pareto (1897). It has became a popular model to use in economics, finance and actuarial science, especially where phenomena characterised by heavy tails are studied (see, Nofal and El Gebaly, 2017, Ismaïl, 2004, Dyer, 1981, Malik, 1970 and Fisk, 1961). Due to the popularity of this distribution, goodness-of-fit tests have been developed in order to test the hypothesis that an observed dataset is realised from the Pareto distribution. For a recent overview and discussion of some of these tests, see Chu et al. (2019) and the references therein.

The Pareto distribution is also used to model lifetimes in survival analysis and reliability theory; see, e.g., Amin (2007), Ouyang and Wu (1994) and Davis and Feldstein (1979). In many of these type of applications random right censoring is present and one would like to test the hypothesis that the lifetime distribution follows the Pareto Type I distribution. Apart from the traditional Kolmogorov-Smirnov and Cramér-von Mises tests, no other tests are available in the statistical literature to test the goodness-of-fit of the Pareto distribution when random censoring is present. In this paper we propose a new fixed point characterisation based test for the Pareto distribution in the presence of random right censoring.

Before proceeding, we introduce some notation. Let $X, X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) positive random variables with unknown continuous distribution function G. Let $C_1, \ldots, C_n$ be i.i.d. censoring variables with unknown continuous distribution function H. We assume non-informative censoring throughout. The observed values are the pairs (T_j, δ_j), $j = 1, \ldots, n$,

Corresponding author: I. J. H. Visagie (jaco.visagie@nwu.ac.za)
MSC2020 subject classifications: 62G09, 62G10, 62N03

where $T_j = \min(X_j, C_j)$ and $\delta_j = I(X_j < C_j)$, with $I(\cdot)$ denoting the indicator function. The order statistics of $T_1, \ldots, T_n$ are denoted by $T_{(1)} < \cdots < T_{(n)}$ and $\delta_{(j)}$ represents the indicator variable corresponding to $T_{(j)}$. Denote the Pareto Type I distribution with shape parameter $\beta > 0$ by

$$F_\beta(x) = 1 - x^{-\beta}, \quad x \geq 1.$$

Formally, we are interested in testing the composite hypothesis

$$H_0 : X \sim F_\beta, \tag{1}$$

for some $\beta > 0$ against general alternatives.

Based on the data (T_j, δ_j) we may estimate the distribution function underlying the lifetime using the Kaplan-Meier estimator, $\tilde{G}_n$;

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

The remainder of the article is organised as follows. In Section 2 the new test statistic is introduced and discussed. The results of a Monte Carlo study, where the size and power performance of the newly proposed test are compared to those of the two classical distribution function based tests (modified to account for censoring), are given in Section 3. The paper concludes in Section 4 with some suggestions for future research.

2. Test statistic

Throughout the paper, β is estimated using maximum likelihood. Simple calculations show that the maximum likelihood estimator of β based on the observed data is

$$\hat{\beta}(T_1, \ldots, T_n) = \frac{\sum_{j=1}^n \delta_j}{\sum_{j=1}^n \log T_j}.$$

Note that $X \sim F_\beta \iff X^\beta \sim F_1$. As a result, the tests used below are based on the transformed variables $Y_j = T_j^{\hat{\beta}(T_1,\ldots,T_n)}$, $j = 1, \ldots, n$. This is justified by the fact that

$$\hat{\beta}(Y_1, \ldots, Y_n) = \frac{\sum_{j=1}^n \delta_j}{\sum_{j=1}^n \log Y_j} = \frac{\sum_{j=1}^n \delta_j}{\hat{\beta}(T_1, \ldots, T_n) \sum_{j=1}^n \log T_j} = \frac{\hat{\beta}(T_1, \ldots, T_n)}{\hat{\beta}(T_1, \ldots, T_n)} = 1.$$

We propose a test statistic based on a fixed point characterisation of the Pareto distribution. Betsch and Ebner (2018) provide these characterisations (which are generalisations of the well-known Stein's method) for a large class of distributions. The fixed point characterisation of F_1 is given in Theorem 1.

Theorem 1. *Let Y be a continuous random variable with support $[1, \infty)$, distribution function G and $E[Y^{-1}] < \infty$. $Y \sim F_1$ if, and only if,*

$$E\left[\frac{2}{Y}(\min(Y, t) - 1)\right] = F_1(t), \quad \forall t \geq 1.$$

Theorem 1 implies that $Y \sim F_1$ if, and only if,

$$\varphi(t) := V_Y(t) - F_1(t) = 0, \tag{2}$$

for all $t > 1$, where $V_Y(t) := E\left[\frac{2}{Y}(\min(Y,t) - 1)\right]$.

Using the Kaplan-Meier estimator, $V_Y(t)$ can be estimated as

$$\hat{V}_Y(t) = \int_1^\infty \frac{2}{y}(\min(y,t) - 1)d\tilde{G}_n(y) = 2\sum_{j=1}^{n} \frac{\Lambda_j}{Y_j}(\min(Y_j,t) - 1), \tag{3}$$

where $\Lambda_j, j = 1, \ldots, n$ is the jump size in $\tilde{G}_n(Y_{(j)})$, given by

$$\Lambda_1 = \frac{\delta_{(1)}}{n},\ \Lambda_n = \prod_{j=1}^{n-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{(j)}} \text{ and}$$

$$\Lambda_j = \frac{\delta_{(j)}}{n-j+1}\prod_{l=1}^{j-1}\left(\frac{n-l}{n-l+1}\right)^{\delta_{(l)}},\ j = 2, \ldots, n-1.$$

Under the hypothesis stated in (1), the distribution function underlying $Y_1, \ldots, Y_n$ is approximately F_1, at least for large n. As a result,

$$\hat{\varphi}_n(t) := \hat{V}_Y(t) - F_1(t)$$

should be close to 0 for all $t > 1$. Therefore, we propose the test statistic

$$S_{n,a} = \int_1^\infty \hat{\varphi}_n^2(t)t^{-a}d\tilde{G}_n(t) = \sum_{k=1}^{n}\Lambda_k\left(2\sum_{j=1}^{n}\frac{\Lambda_j}{Y_j}[\min(Y_j,Y_k) - 1] + Y_k^{-1} - 1\right)^2 Y_k^{-a},$$

where t^{-a} is a weight function ensuring the existence of the integral and $a > 2$ is a tuning parameter. The test rejects the null hypothesis for large values of $S_{n,a}$. Since the null distribution of the test statistic is a function of the unknown censoring distribution H, we propose the following bootstrap algorithm to estimate the critical value.

1. Based on the pairs (T_j, δ_j),$j = 1, \ldots, n$, estimate β by $\hat{\beta} = \sum_{j=1}^{n}\delta_j/(\sum_{j=1}^{n}\log T_j)$.
2. Obtain a parametric bootstrap sample $X_1^*, \ldots, X_n^*$ by sampling from $F_{\hat{\beta}}$.
3. Obtain a non-parametric bootstrap sample $C_1^*, \ldots, C_n^*$ by sampling from the Kaplan-Meier estimate of the distribution of the censoring times.
4. Let
$$T_j^* = \min(X_j^*, C_j^*) \text{ and } \delta_j^* = \begin{cases} 1, & \text{if } X_j^* \le C_j^* \\ 0, & \text{if } X_j^* > C_j^*. \end{cases}$$
5. Calculate $\hat{\beta}^* = \sum_{j=1}^{n}\delta_j^*/(\sum_{j=1}^{n}\log T_j^*)$ and obtain the transformed bootstrap values $Y_j^* = T_j^{*\hat{\beta}^*}$.
6. Calculate the test statistic, say $S^* = S((Y_1^*, \delta_1^*), (Y_2^*, \delta_2^*), \ldots, (Y_n^*, \delta_n^*))$, based on the data (Y_j^*, δ_j^*), $j = 1, \ldots, n$.
7. Repeat steps 2–6 B times to obtain $S_1^*, \ldots, S_B^*$. Use the $(1-\alpha)$th quantile of $S_1^*, \ldots, S_B^*$ as the estimated critical value for the test.

3. Simulation study

In this section the empirical power performance of the newly proposed test $S_{n,a}$ is compared to that of the modified Kolmogorov-Smirnov (KS_n) and Cramér-von-Mises (CV_n) tests. For a discussion on these two modified tests, see D'Agostino and Stephens (1986) as well as Koziol and Green (1976). A significance level of $\alpha = 0.05$ is used throughout and critical values of all the tests are obtained using the bootstrap algorithm from Section 2. Estimated powers are shown for samples of size $n = 50$ and $n = 100$. The reported empirical powers are calculated in the case of 10% and 20% censoring for various alternative lifetime distributions. The alternative distributions used are the gamma and lognormal distributions, both shifted by 1 to ensure that these distributions have the same support as that of the Pareto distribution. The gamma distribution is denoted $\Gamma(\theta)$ and has density

$$h(x) = \frac{1}{\Gamma(\theta)}(x-1)^{\theta-1}e^{-(x-1)}, \quad x \geq 1.$$

The lognormal distribution is denoted $LN(\theta)$ and has density

$$h(x) = \exp\left(-\frac{1}{2}\left(\log(x-1)/\theta\right)^2\right)\left(\theta(x-1)\sqrt{2\pi}\right)^{-1}, \quad x \geq 1.$$

Empirical sizes are presented for the Pareto distributions with parameters 2 and 3, denoted by F_2 and F_3 in Tables 1 to 4. The censoring distribution used is the uniform distribution on the interval $(1, c)$, where $c > 1$ is chosen to produce the desired censoring proportion. For computational efficiency, power calculations are done using the warp-speed bootstrap proposed by Giacomini et al. (2013). For another example of the warp-speed bootstrap methodology used to calculate empirical powers, see Allison et al. (2019). All calculations are performed in R (R Core Team, 2020).

Tables 1 to 4 contain the percentage (rounded to the nearest integer) of 10 000 independent Monte Carlo samples for which the hypothesis stated in (1) was rejected. Each of these tables show the empirical powers associated with a given combination of sample size and censoring proportion. We display the highest power against each alternative in bold.

The power tables indicate that each of the tests maintain the specified significance level of 5% closely. The powers achieved by the tests increase with an increase in sample size and decrease as the censoring proportion increases. When considering the powers achieved against the various alternatives we see that the CV_n test is less powerful than the other tests considered for the alternatives above. However, this may be a result of the specific censoring distribution used. A more extensive Monte Carlo study will be required if more general conclusions are to be drawn regarding the power of the CV_n test.

Based on Tables 1 to 4, it seems that the KS_n test is quite powerful against gamma alternatives while $S_{n,a}$ achieves the highest powers against lognormal alternatives. Furthermore, it seems that smaller values of the tuning parameter in $S_{n,a}$ generally lead to higher powers.

4. Concluding remarks

To use our newly proposed test in a real world setting, a choice of the tuning parameter is necessary. Based on the Monte Carlo study above we suggest choosing $a = 2.1$ as this choice generally produces high powers for the alternatives considered. Tenreiro (2019) and Allison and Santana (2015) proposed

Table 1. Empirical powers for $n = 50$ with 10% censoring.

Distribution	KS_n	CV_n	$S_{n,2.1}$	$S_{n,2.2}$	$S_{n,2.5}$
F_2	5	4	5	5	5
F_3	5	4	5	5	5
$\Gamma(0.9)$	**29**	13	14	12	8
$\Gamma(1)$	**47**	15	28	26	17
$LN(1.2)$	51	4	**66**	64	60
$LN(1.5)$	8	4	**12**	**12**	11

Table 2. Empirical powers for $n = 100$ with 10% censoring.

Distribution	KS_n	CV_n	$S_{n,2.1}$	$S_{n,2.2}$	$S_{n,2.5}$
F_2	5	4	5	5	5
F_3	5	5	5	5	5
$\Gamma(0.9)$	**54**	16	23	20	12
$\Gamma(1)$	**78**	23	49	44	29
$LN(1.2)$	86	12	**93**	**93**	90
$LN(1.5)$	12	5	**21**	20	19

Table 3. Empirical powers for $n = 50$ with 20% censoring.

Distribution	KS_n	CV_n	$S_{n,2.1}$	$S_{n,2.2}$	$S_{n,2.5}$
F_2	5	4	5	5	4
F_3	4	4	5	5	4
$\Gamma(0.9)$	**22**	10	14	13	10
$\Gamma(1)$	**37**	10	27	25	21
$LN(1.2)$	42	8	**58**	**58**	**58**
$LN(1.5)$	7	5	**10**	**10**	**10**

Table 4. Empirical powers for $n = 100$ with 20% censoring.

Distribution	KS_n	CV_n	$S_{n,2.1}$	$S_{n,2.2}$	$S_{n,2.5}$
F_2	5	5	5	5	5
F_3	5	5	5	5	5
$\Gamma(0.9)$	**41**	10	25	23	17
$\Gamma(1)$	**68**	11	52	49	39
$LN(1.2)$	77	14	**92**	**92**	91
$LN(1.5)$	10	5	19	19	**20**

methods to choose tuning parameters in certain settings data-dependently. Unfortunately, neither of these methods are applicable in our situation as our critical values must be obtained using the bootstrap. A data-dependent choice for this type of scenario is still an open problem in the goodness-of-fit literature.

Another challenging possibility for future research is to develop the asymptotic theory relating to the newly proposed test. The dependence introduced by the presence of censoring complicates the derivation of the asymptotic results. Recently Fernández and Rivera (2020) studied Kaplan-Meier U- and V-statistics in the presence of random censoring. The results found in the mentioned paper may prove useful as tools to derive the asymptotic null distribution of the newly proposed test statistic.

References

Allison, J. and 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**, 3276–3288.

Allison, J. S., Betsch, S., Ebner, B., and Visagie, I. J. H. (2019). New weighted L^2-type tests for the inverse Gaussian distribution. *arXiv:1910.14119*.

Amin, Z. H. (2007). Tests for the validity of the assumption that the underlying distribution of life is Pareto. *Journal of Applied Statistics*, **34**, 195–201.

Betsch, S. and Ebner, B. (2018). Characterizations of continuous univariate probability distributions with applications to goodness-of-fit testing. *arXiv:1810.06226*.

Chu, J., Dickin, O., and Nadarajah, S. (2019). A review of goodness of fit tests for Pareto distributions. *Journal of Computational and Applied Mathematics*, **361**, 13–41.

D'Agostino, R. B. and Stephens, M. A. (1986). *Goodness-of-fit Techniques*. Marcel Dekker, New York.

Davis, H. T. and Feldstein, M. L. (1979). The generalized Pareto law as a model for progressively censored survival data. *Biometrika*, **66**, 299–306.

Dyer, D. (1981). Structural probability bounds for the strong Pareto law. *Canadian Journal of Statistics*, **9**, 71–77.

Fernández, T. and Rivera, N. (2020). Kaplan-Meier V- and U-statistics. *Electronic Journal of Statistics*, **14**, 1872–1916.

Fisk, P. R. (1961). The graduation of income distributions. *Econometrica*, 171–185.

Giacomini, R., Politis, D. N., and White, H. (2013). A warp-speed method for conducting Monte Carlo experiments involving bootstrap estimators. *Econometric Theory*, **29**, 567–589.

Ismaïl, S. (2004). A simple estimator for the shape parameter of the Pareto distribution with economics and medical applications. *Journal of Applied Statistics*, **31**, 3–13.

Koziol, J. A. and Green, S. B. (1976). A Cramér-von Mises statistic for randomly censored data. *Biometrika*, **63**, 465–474.

Malik, H. J. (1970). A characterization of the Pareto distribution. *Scandinavian Actuarial Journal*, **1970**, 115–117.

Nofal, Z. M. and El Gebaly, Y. M. (2017). New characterizations of the Pareto distribution. *Pakistan Journal of Statistics and Operation Research*, **13**, 63–74.

Ouyang, L. Y. and Wu, S. J. (1994). Prediction intervals for an ordered observation from a Pareto distribution. *IEEE Transactions on Reliability*, **43**, 264–269.

Pareto, V. (1897). *Cours d'Économie Politique*. F. Rouge, Lausanne.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Tenreiro, C. (2019). On the automatic selection of the tuning parameter appearing in certain families of goodness-of-fit tests. *Journal of Statistical Computation and Simulation*, **89**, 1780–1797.