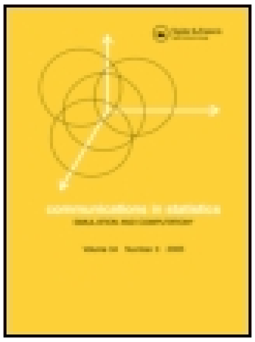

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: http://www.tandfonline.com/loi/lssp20

Model choice in separate families: A comparison between the FBST and the Cox test

Cachimo Combo Assane, Basilio de Bragança Pereira & Carlos Alberto de Bragança Pereira

To cite this article: Cachimo Combo Assane, Basilio de Bragança Pereira & Carlos Alberto de Bragança Pereira (2018): Model choice in separate families: A comparison between the FBST and the Cox test, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2018.1458139

To link to this article: https://doi.org/10.1080/03610918.2018.1458139

COMMUNICATIONS IN STATISTICS—SIMULATION AND COMPUTATION®
2018, VOL. 0, NO. 0, 1–14
https://doi.org/10.1080/03610918.2018.1458139

Model choice in separate families: A comparison between the FBST and the Cox test

Cachimo Combo Assane[a], Basilio de Bragança Pereira[b],
and Carlos Alberto de Bragança Pereira[c]

[a]Universidade Eduardo Mondlane (UEM), Maputo, Mozambique; [b]Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil; [c]Universidade de São Paulo (USP), São Paulo, Brazil

ABSTRACT
The important problem of discriminating between separate families of distributions is the theme of this work. The Bayesian significance test, FBST, is compared with the celebrated Cox test. The three families most used in survival analysis, lognormal, gamma and Weibull, are considered for the discrimination. A convex combination—with unknown weights—of the three densities is used for this discrimination. After these weights have been estimated, the one with the highest value indicates the best statistical model among the three. Another important feature considered is the parameterization used. All the three densities are written as a function of the common population mean and variance. Including the weights, the number of parameters is reduced from eight (two of each density and two of the convex combination) to four (two from the common mean and variance plus two of the weights). Some numerical results from simulations are given. In these simulations, the results of FBST are compared with those obtained with the Cox test. Two real examples properly illustrate the procedures.

ARTICLE HISTORY
Received 23 January 2018
Accepted 27 February 2018

KEYWORDS
Cox Test; FBST; Mixture model; Model choice; Separate Models; Significance test

MATHEMATICS SUBJECT CLASSIFICATION
62F15; 62F03

1. Introduction

Looking for a statistical model is the first step of a statistical analysis. Thus, choosing a model from a group of alternatives is one of the most important statistical problems. The Neyman-Pearson test is the statistical technique most used for this purpose; however, it is only applicable in cases where all alternative models belong to the same parametric family of distributions. In such cases, the choice of parametric values among alternatives ones is, in fact, the problem to be solved. When the statistician first has to choose a family of distributions in a set of separate (non-nested) families, other techniques are needed. Two families are said to be separate or non-nested if an arbitrary member of a family cannot be obtained by a limit of any sequence of members of the other family. The set of separate families treated in the present work is restricted to three of the most important families of distributions: Lognormal, Gamma, and Weibull. These distributions play a very significant role in reliability and survival analysis; see, for instance, Pereira (1981) and Lawless (2002).

The literature is full of research on separate families since Cox (1961, 1962), the first works that dealt with the problem. Pereira (2005), Araujo et al. (2005), Araujo and Pereira (2007)

CONTACT Cachimo Combo Assane cachimo.assane@gmail.com Universidade Eduardo Mondlane (UEM), Maputo, Mozambique.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

and Pereira and Pereira (2017) are few of the reviews of Cox‘s work. The present paper is the natural companion of Assane, Pereira, and Pereira (2017).

Cox (1977) and Kempthorne (1976) considered a test of significance as a procedure to measure the consistency of the data with a null hypothesis through the value *p*. Berger and Delampady (1987) presented Bayesian alternatives by considering Bayes Factors and posterior probabilities of the null hypothesis as measures of significance. Using another point of view, Pereira and Stern (1999) defined another measure that is based on the posterior probability of a specific event in the parameter space, the tangent set. The associated test is known as the Fully Bayesian Significance Test (FBST) and its measure of evidence is called e-value. Recall that the p-values are tails of the null distribution calculated from the observed data and the e-values are tails of the posterior distribution calculated from the null hypothesis. Araujo et al. (2005), Araujo and Pereira (2007) and Kamary et al. (2014) have documented some difficulties using traditional Bayesian model choices when Bayes factors or posterior probabilities of null sets are considered as measures of significance. These difficulties occur, for example, when the prior information is weak or when an improper prior is used. To overcome these difficulties, due to inadequate priors, modified Bayes factors were proposed; see, for example, Araujo and Pereira (2007) and Pereira and Pereira (2017).

The e-value of the FBST is an alternative to both the Bayes factor and classical p-values for the case of precise hypotheses. Precise hypotheses are hypotheses defined in sets with a smaller dimension*red*ality than *red*that of the *red*full parameter space. The basis for the FBST is the e-value (e standing for evidence), which measures the inconsistency of the hypothesis with the observed data. For further references on FBST, see Pereira, Stern, and Wechsler (2008) and Stern and Pereira (2014). For a possible relation between p-values, based on likelihood ratio tests, and e-values, see Diniz et al. (2012). Three interesting applications illustrating the use of FBST with its e-values can be found in Irony et al. (2002), Lauretto et al. (2003) and Chakrabarty (2016).

In order to discriminate between the three separate models, lognormal, gamma and Weibull distributions, we consider first a convex linear combination of their three densities and we use the FBST not only to estimate the parameters of the distributions, but also to estimate the weights of the convex combination and for testing the hypotheses defined on the weights space. The chosen prior distribution for the vector of the weights is the Dirichlet distribution of order three (equivalently of dimension two). This combination is not a novelty but a suggestion presented by Cox (1961) and applied by Lauretto et al. (2007) and Lauretto, Pereira, and Stern (2003). The recent novelty could be the parameterization using the mean and the variance in the place of the original parameters. These methods were used by Kamary et al. (2014), Assane, Pereira, and Pereira (2017), and Pereira and Pereira (2017). Note that population means and variances are independent of the distribution one use and also all three distributions considered here may be written with these most common population characteristics. Instead of the six original parameters, we have now two parameters, common to the three densities. With the weights of the convex (linear) combination, we have four parameters to be estimated. Also, anyone that assesses a prior distribution to population means and variances do not need to know which statistical model the statistician will choose.

To compare the method described above with the Cox test, numerical results based on simulations is presented together with a discussion of the empirical results obtained by both methods. Also, the lognormal-gamma-Weibull mixture model was applied to the simulated data in order to evaluate the performance of the FBST for choosing the right density. Two examples of real observed uncensored datasets are presented.

Section 2 is a brief review of Cox test. The basic FBST concepts are presented in Section 3. Section 4 discusses the FBST formulation for discriminating between separate models in the context of the convex combination of densities. Section 5 presents the simulation results of both the FBST and the Cox test. Section 6 focuses on two real datasets, and illustrates the performance of the procedures. Final remarks are presented in Section 7.

2. The Cox test

Let $y = (y_1, \dots, y_n)$ be independent and identically distributed observations from some unknown distribution F. Suppose that there is a null hypothesis, $H_f : F \in \mathfrak{F}_f$, where $\mathfrak{F}_f$ is a family of probability distributions with density $f(y|\alpha)$, and an alternative hypothesis, $H_g : F \in \mathfrak{F}_g$, where $\mathfrak{F}_g$ is another family of probability distributions with density $g(y|\beta)$. Hence α and β are unknown parameter vectors and it is assumed further that the families of f and g are separate in the sense defined above. Formal definitions of separate hypotheses are given in Pereira and Pereira (2017).

The asymptotic test developed by Cox (1961, 1962) is based on a modification of the Neyman-Pearson maximum likelihood ratio. The test statistic for H_f against H_g is

$$T_{fg} = \ell_f(\hat{\alpha}) - \ell_g(\hat{\beta}) - n\left[\operatorname*{plim}_{n\to\infty} \frac{\ell_f(\hat{\alpha}) - \ell_g(\hat{\beta})}{n}\right]_{\alpha=\hat{\alpha}},$$

where $\ell_f(\hat{\alpha})$ and $\ell_g(\hat{\beta})$ are the maximized log-likelihoods under H_f and H_g, respectively; $\hat{\alpha}$ and $\hat{\beta}$ denote the maximum likelihood estimates; plim represents convergence in probability; and the subscript α indicates that the means are calculated under H_f.

Cox showed that, asymptotically, under the alternative hypothesis, T_{fg} has a negative mean, whereas, under the null hypothesis, it is normally distributed with mean zero and variance

$$V_\alpha(T_{fg}) = V_\alpha\{\ell_f(\alpha) - \ell_g(\beta_\alpha)\} - C'_\alpha I_\alpha^{-1} C_\alpha,$$

where β_α is the probability limit of $\hat{\beta}$ under H_f, as $n \longrightarrow \infty$, $C_\alpha \equiv n\frac{\partial}{\partial\alpha}[\operatorname{plim}_{n\to\infty} \frac{\ell_f(\hat{\alpha})-\ell_g(\hat{\beta})}{n}]$, and I_α the information matrix of α. When H_g is the null hypothesis and H_f is the alternative hypothesis, analogous results are obtained for a statistic T_{gf}. Therefore, $T^*_{fg} = T_{fg}\{V(T_{fg})\}^{-1/2}$ and $T^*_{gf} = T_{gf}\{V(T_{gf})\}^{-1/2}$ under H_f and H_g, respectively, are approximately standard normal variables, and two-tailed tests can be performed. For example, if T^*_{fg} is significantly negative, there is evidence of a departure from H_f in the direction of H_g. If T^*_{fg} is significantly positive, there is evidence of a departure form H_f in the opposite direction to H_g. The possible results of application of both tests (T_{fg} and T_{gf}) are presented in Pereira and Pereira (2017).

As an illustration of the calculations for the Cox test statistics, following Pereira (1978), suppose that H_f specifies that the distribution is lognormal and H_g specifies that it is Weibull; that is

$$H_f : f(y|\alpha) = \frac{1}{y\sqrt{2\pi\alpha_2}} \exp\left\{-\frac{(\log y - \alpha_1)^2}{2\alpha_2}\right\}, \quad \alpha = (\alpha_1, \alpha_2),$$

$$H_g : g(y|\beta) = \frac{\beta_2}{\beta_1^{\beta_2}} y^{\beta_2 - 1} \exp\left\{-\left(\frac{y}{\beta_1}\right)^{\beta_2}\right\}, \quad \beta = (\beta_1, \beta_2).$$

We then have

$$T_{fg} = n\big\{\hat{\beta}_2 \ln \hat{\beta}_1 - \beta_{2\hat{\alpha}} \ln \beta_{1\hat{\alpha}} - \ln \hat{\beta}_2 + \ln \beta_{2\hat{\alpha}} - \hat{\alpha}_1(\hat{\beta}_2 - \beta_{2\hat{\alpha}})\big\}$$

and

$$V_\alpha(T_{fg}) = 0.2183n,$$

where $\beta_{1\hat{\alpha}} = \exp\{\hat{\alpha}_1 + \sqrt{\hat{\alpha}_2}/2\}$ and $\beta_{2\hat{\alpha}} = \hat{\alpha}_2^{-1/2}$ are the estimated values of $\beta_{1\alpha}$ and $\beta_{2\alpha}$ that are the probability limits of $\hat{\beta}_1$ and $\hat{\beta}_2$ under H_f, respectively.

Also,

$$T_{gf} = n\left\{\hat{\beta}_2(\hat{\alpha}_1 - \alpha_{1\hat{\beta}}) + \frac{1}{2}\ln\frac{\hat{\alpha}_2}{\alpha_{2\hat{\beta}}}\right\}$$

and

$$V_\beta(T_{gf}) = 0.2834n,$$

where $\alpha_{1\hat{\beta}} = -0.5772/\hat{\beta}_2 + \ln\hat{\beta}_1$ and $\alpha_{2\hat{\beta}} = 1.6449/\hat{\beta}_2^2$ are the estimated values of $\alpha_{1\beta}$ and $\alpha_{2\beta}$, which are the probability limits of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ under H_g, respectively.

The Cox test statistics for discriminating between exponential vs. lognormal, lognormal vs. gamma, and gamma vs. Weibull distributions can be found in Pereira and Pereira (2017).

3. Fully Bayesian Significance Test (FBST)

The FBST of Pereira and Stern (1999), which is reviewed in Pereira, Stern, and Wechsler (2008), is a Bayesian version of significance testing, as considered by Cox (1977) and Kempthorne (1976), for precise (or sharp) hypotheses.

First, let us consider a real parameter θ, a point in the parameter space $\Theta \subset \Re$, and an observation y of the random variable Y. A frequentist looks for the set $I \in \Re$ of sample points that are at least as inconsistent with the hypothesis as y is. A Bayesian looks for the tangential set $T(y) \subset \Theta$ (Pereira, Stern, and Wechsler 2008), which is a set of parameter points that are more consistent with the observed y than the hypothesis is. An example of a sharp hypothesis in a parameter space of the real line is of the type $H: \theta = \theta_0$. The evidence value in favor of H for a frequentist is the usual p-value, $P(Y \in I|\theta_0)$, whereas for a Bayesian, the evidence in favor of H is the e-value, $ev = 1 - \Pr(\theta \in T(y)|y)$.

In the general case of multiple parameters, $\Theta \subset \Re^k$, let the posterior distribution for θ given y be denoted by $q(\theta|y) \propto \pi(\theta)L(y, \theta)$, where $\pi(\theta)$ is the prior probability density of θ and $L(y, \theta)$ is the likelihood function. In this case, a sharp hypothesis is of the type $H: \theta \in \Theta_H \subset \Theta$, where Θ_H is a submanifold of smaller dimension than Θ. Letting $\sup_H$ denote the supremum of Θ_H, we define the e-value evidence measure and the tangential set, $T(y)$, as follows:

$$q^* = \sup_H q(\theta|y) \text{ and } T(y) = \{\theta : q(\theta|y) > q^*\}. \tag{3.1}$$

The Bayesian evidence value against H is the posterior probability of $T(y)$,

$$\overline{ev} = \Pr(\theta \in T(y)|y) = \int_{T(y)} q(\theta|y)d\theta; \text{ consequently, } ev = 1 - \overline{ev}. \tag{3.2}$$

It is important to note that evidence that favors H is not evidence against the alternative, $\overline{H} = \Theta \setminus H$, because it is not a sharp hypothesis. This interpretation also holds for p-values in the frequentist paradigm. As in Pereira, Stern, and Wechsler (2008), we would like to point out that this Bayesian significance index uses only the posterior distribution, with no need for additional artifacts such as the inclusion of positive prior probabilities for the hypotheses or the elimination of nuisance parameters. The computation of e-values does not require asymptotic methods, and the only technical tools needed are numerical optimization and integration methods.

Let us consider the distribution function of the evidence value against the hypothesis, $\overline{V}(c) = \Pr(\overline{\text{ev}} \leq c)$, given θ^0, the true value of the parameter. Under appropriate regularity conditions, for increasing sample size, $n \to \infty$, we can state the following:

- If H is false, $\theta^0 \notin H$, then $\overline{ev}$ converges (in probability) to 1, that is, $\overline{V}(0 < c < 1) \to 0$.
- if H is true, $\theta^0 \in H$, then $\overline{V}(c)$, the confidence level, is approximated by the function

  $$Q(t, h, c) = F_{t-h}\left[F_t^{-1}(c)\right],$$

  where $t = \dim(\Theta)$, $h = \dim(H)$, $F_g(x)$ is the cumulative density function of chi-square distribution with g degree of freedom.

Hence, for large n, to reject H with level of significance γ, we set c such that $Q(t, h, c) = 1 - \gamma$, i.e., $c = F_t[F_{t-h}^{-1}(1 - \gamma)]$. Therefore, the FBST procedure rejects H if $\overline{\text{ev}}(H) > c$.

Diniz et al. (2012) have shown that, asymptotically, there is a relationship between $\overline{\text{ev}}(H)$ and the p-value based on the Likelihood ratio test. Thus, from the asymptotic normality property, $\overline{\text{ev}}(H) \approx F_t[F_{t-h}^{-1}(1 - \text{p-value})]$. We then have

$$\text{p-value} = 1 - F_{t-h}\left[F_t^{-1}(\overline{\text{ev}}(H))\right]. \tag{3.3}$$

4. Mixture of separate models

Let us consider a dataset $y = \{y_1, \ldots, y_n\}$ and m alternative probability distributions with densities $f_1(y|\psi_1), f_2(y|\psi_2), \ldots, f_m(y|\psi_m)$. Here, ψ_k, $k = 1, \ldots, m$, are unknown (vector) parameters and the families of distributions are separate. The problem of interest is to measure the evidence in favor of each model for fitting the dataset. As suggested by Cox (1961), we can consider a general model including all candidate distributions where the choice of a specific distribution is a special case. In this work, we formulate the FBST for the convex linear combination of separate models as a selection procedure. Denoting $\boldsymbol{\theta} = (\psi_1, \ldots, \psi_m, \boldsymbol{p})$, the density function for the $m-$component mixture model is

$$f(y_j|\boldsymbol{\theta}) = p_1 f_1(y_j|\psi_1) + \cdots + p_m f_m(y_j|\psi_m) \;\; p_k \geq 0, \quad \sum_{k=1}^{m} p_k = 1. \tag{4.1}$$

where $\boldsymbol{p} = (p_1, \ldots, p_m)$ is the vector of the mixture weights.

In this paper, the density functions of the mixture components in (4.1) are reparameterized in terms of the common population mean μ and variance σ^2. These methods were used by Kamary et al. (2014), Assane, Pereira, and Pereira (2017), and Pereira and Pereira (2017). Therefore, we have $\boldsymbol{\theta} = (\mu, \sigma^2, \boldsymbol{p})$ denoting all parameters of the mixture model, where μ and σ^2 are the parameters connecting the different survival distributions, with $\boldsymbol{p}$ corresponding to the vector of the mixture weights.

Assuming that the y_i are conditionally (on the parameter) independent and identically distributed, then, the likelihood function is

$$L(y, \boldsymbol{\theta}) = \prod_{j=1}^{n} \sum_{k=1}^{m} p_k f_k(y_j|\mu, \sigma). \tag{4.2}$$

The families of distributions considered in this work include the lognormal, gamma and Weibull models. The relationship between the parameters of these models through the μ and σ^2 is described as follows.

(i) Let y be a lognormal(α_1, α_2), $\alpha_1 \in \mathbb{R}$ and $\alpha_2 > 0$, with probability density function

$$f_L(y|\alpha_1, \alpha_2) = \frac{1}{y\sqrt{2\pi\alpha_2}} \exp\left\{-\frac{(\log y - \alpha_1)^2}{2\alpha_2}\right\}.$$

We then have

$$\begin{cases} \mu = E(y|\alpha_1, \alpha_2) = \mathrm{e}^{\alpha_1 + \alpha_2/2} \\ \sigma^2 = \mathrm{Var}(y|\alpha_1, \alpha_2) = (\mathrm{e}^{\alpha_2} - 1)\mathrm{e}^{2\alpha_1 + \alpha_2} \end{cases} \Rightarrow \begin{cases} \alpha_1 = \log \frac{\mu^2}{\sqrt{\mu^2 + \sigma^2}} \\ \alpha_2 = \sqrt{\log \frac{\mu^2 + \sigma^2}{\mu^2}}. \end{cases} \tag{4.3}$$

(ii) Let y be a gamma(γ_1, γ_2), $\gamma_1 > 0$ and $\gamma_2 > 0$, with probability density function

$$f_G(y|\gamma_1, \gamma_2) = \frac{1}{\Gamma(\gamma_2)\gamma_1^{\gamma_2}} y^{\gamma_2 - 1} \exp\left\{-\frac{y}{\gamma_1}\right\}.$$

Therefore

$$\begin{cases} \mu = E(y|\gamma_1, \gamma_2) = \gamma_1\gamma_2 \\ \sigma^2 = \mathrm{Var}(y|\gamma_1, \gamma_2) = \gamma_2\gamma_1^2 \end{cases} \Rightarrow \begin{cases} \gamma_1 = \frac{\sigma^2}{\mu} \\ \gamma_2 = \frac{\mu^2}{\sigma^2}. \end{cases} \tag{4.4}$$

(iii) When $y \sim$ Weibull(β_1, β_2), $\beta_1 > 0$ and $\beta_2 > 0$, with probability density function

$$f_W(y|\beta_1, \beta_2) = \frac{\beta_2}{\beta_1^{\beta_2}} y^{\beta_2 - 1} \exp\left\{-\left(\frac{y}{\beta_1}\right)^{\beta_2}\right\},$$

then

$$\begin{cases} \mu = E(y|\beta_1, \beta_2) = \beta_1\Gamma(1 + 1/\beta_2) \\ \sigma^2 = \mathrm{Var}(y|\beta_1\beta_2) = \beta_1^2\Gamma(1 + 2/\beta_2) - \beta_1^2\Gamma^2(1 + 1/\beta_2) \end{cases}$$
$$\Rightarrow \begin{cases} \beta_1 = \frac{\mu}{\Gamma(1 + 1/\beta_2)} \\ 2\log\Gamma(1 + 1/\beta_2) - \log\Gamma(1 + 2/\beta_2) + \log\frac{\mu^2 + \sigma^2}{\mu^2} = 0. \end{cases} \tag{4.5}$$

In order to find β_2, the Newton-Rapson method can be used to solve the nonlinear equation. Here, we use the nleqslv” function in the R” package of the same name.

Assuming independence, the joint prior density function of $\boldsymbol{\theta} = (\mu, \sigma^2, \boldsymbol{p})$ is given by $\pi(\boldsymbol{\theta}) = \pi_1(\boldsymbol{p})\pi_2(\mu)\pi_3(\sigma^2)$. Therefore, according to the Bayesian paradigm, the posterior density of $\boldsymbol{\theta}$ is

$$f(\boldsymbol{\theta}|y) \propto L(y, \boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{4.6}$$

In this paper, the prior distributions for the connecting parameters, μ and σ^2, are assumed to be independent gamma distributions, both with a mean of one and a variance of 100, that is, $\mu, \sigma^2 \sim$ gamma$(0.01, 100)$ (Pereira and Pereira 2017). For the mixture weights, we use a Dirichlet prior, $\boldsymbol{p} \sim$ Dir$(1, 1, 1)$ when all families of models are considered ($m = 3$) or a Beta prior with parameters (1,1) (uniform$(0, 1)$) for any combination of $m = 2$ (Pereira and Stern 2008).

In order to measure the evidence in favour of each model, hypotheses on the mixture weights are tested (Kamary et al. 2014; Assane, Pereira, and Pereira 2017; Pereira and Pereira 2017).

The hypothesis specifying that y has the density function $f_k(y|\psi_k)$ is equivalent to

$$H_k : p_k = 1 \wedge p_i = 0, \quad i \neq k. \tag{4.7}$$

On the other hand, the hypothesis that y does not have the density $f_k(y|\psi_k)$ is equivalent to

$$H : p_k = 0 \wedge \sum_{i \neq k} p_i = 1. \tag{4.8}$$

The alternative hypotheses to (4.7) and (4.8) are $A_k : p_k < 1$ and $A_k : p_k > 0$, respectively, which are not sharp anyway.

The FBST procedure is used to test H_k, $k = 1, \ldots, m$, according to expressions (3.1) and (3.2). For the optimization step, we used the conjugate gradient method (Fletcher and Reeves 1964). In order to perform the integration over the posterior measure, we used an Adaptive Metropolis Markov chain Monte Carlo algorithm, MCMC, of Haario, Saksman, and Tamminen (2001).

In this paper, the implementation of the computations in the Bayesian models is carried out using the LaplacesDemon" R" package. LaplacesDemon" is an open-source package that provides a complete environment for simulation in Bayesian inference (Statisticat, LCC 2016).

5. Simulations

In this section we present some numerical results based on simulated sample points in order to evaluate the performance of the FBST for discriminating between separate families of distributions. Our main interest is to measure the convergence rate of correct decisions concerning the acceptance/rejection of the true/false distribution of the generated sample, when using the FBST on the mixture model. In this paper, the simulation study is carried out in two parts. First, we compare the empirical results of the FBST and Cox test on discriminating between two separate models. Second, we apply the lognormal-gamma-Weibull mixture model (LGW) to the simulated data in order to evaluate the performance of the FBST on identifying the true distribution used to generate the sample.

The simulations of this paper were performed on a computer with an Intel(R) Core(TM) i7-5500U CPU@ 2.40 GHz.

5.1. Discriminating between two separate models

Simulation scheme of sample points

In this paper, we illustrate simulations of the lognormal and Weibull distributions. Let H_L and H_W be the hypotheses specifying the probability density functions of the lognormal and Weibull models, respectively, as defined in Section 2. For each hypothesis, we generate 500 samples of sizes $n = 20, 40, 60, 80, 100, 150$ and 200 from the distribution and, for every sample data, n, we compute the evidence in favor of the hypothesis using the FBST procedure and Cox test. Due to the invariance of the e-value (Madruga, Pereira, and Stern 2003) and of the maximum likelihood ratio (Pereira 1978), this case did not require changes in parameter values for the simulations. Therefore, the various sample sizes n from a lognormal were obtained with $\alpha_1 = 0$ and $\alpha_2 = 1$ ($LN(0, 1)$) and the samples from a Weibull were generated with $\beta_1 = 1$ and $\beta_2 = 1$ ($W(1, 1)$).

As an acceptance/rejection threshold, we adopted the critical level c according to the criterion presented in Section 3, with a significance level of 5%. We chose this asymptotic criterion because of our benchmark (the Cox test), which is an asymptotic procedure as well. Since the mixture model and the restricted model have three and two degree of freedom, respectively,

Table 1. Acceptance rates of true null hypothesis H_L. Data from $LN(0, 1)$.

n	20	40	60	80	100	150	200
FBST	0.992	0.978	0.984	0.980	0.984	0.982	0.994
Cox test	0.988	0.980	0.968	0.972	0.964	0.974	0.974

Table 2. Rejection rates of false null hypothesis H_W. Data from $LN(0, 1)$.

n	20	40	60	80	100	150	200
FBST	0.316	0.608	0.704	0.930	0.998	1.000	1.000
Cox test	0.160	0.384	0.670	0.822	0.938	1.000	1.000

Table 3. Acceptance rates of true null hypothesis H_W. Data from $W(1, 1)$.

n	20	40	60	80	100	150	200
FBST	0.998	0.966	0.972	0.982	0.976	0.982	0.984
Cox test	0.994	0.972	0.964	0.978	0.950	0.944	0.936

we have $c = F_3[F_2^{-1}(0.95)] = 0.72$. Therefore, we reject H if $\overline{\text{ev}}(H) > 0.72$ or, equivalently, if $\text{ev}(H) < 0.28$.

For the Cox test, adopting a significance level of 5%, we define the rejection region as follows: $R = \{y : |T^*| > 1.96\}$, where $T^* \sim N(0, 1)$. The expressions for the computations of the Cox test statistics are given in Section 2.

Simulation results

The simulation results are summarized in the tables shown below. As expected, both the FBST and Cox test have achieved high acceptance rates of the null hypotheses that specify the true distributions used to generate the samples (see Tables 1 and 3). The type-I error rates (rejection rates of the true model) obtained by FBST are always below the predefined significance level (5%). The significance levels attained from the Cox test are, in general, very close to 5%. This is what would be expected in a specific application (Pereira 1978).

Regarding the rejection of the hypotheses that specify the false models, it is clear from the Tables 2 and 4 that, as the sample size increases, the rejection rate converges to 1. The rejection rates obtained from the FBST are higher than those of the Cox test mainly when sample sizes are small. This means that the FBST presents higher discrimination power compared to the Cox test. Note that Cox's asymptotic tests are developed under the assumption that a higher power for the alternative hypothesis is required (Cox 1961).

5.2. *Discriminating based on the LGW mixture model*

Let H_L, H_G and H_W be the hypotheses specifying the probability density functions of the log-normal, gamma and Weibull distributions, respectively. From each distribution, we generate 200 samples of sizes $n = 25$, 50, 100, and 200 and, for every sample, we use the FBST on the

Table 4. Rejection rates of false null hypothesis H_L. Data from $W(1, 1)$.

n	20	40	60	80	100	150	200
FBST	0.410	0.784	0.860	0.928	0.956	0.990	1.000
Cox test	0.304	0.580	0.774	0.896	0.942	0.994	1.000

Table 5. Mean of estimates for LGW model parameters and percentages of correct decisions made by FBST in selecting the true distribution of the generated samples.

		μ	σ^2	p_L	p_G	p_W	
Model	n	20	50	—	—	—	% of Cd*
Lognormal	25	19.93	51.74	0.39	0.35	0.26	55
	50	19.89	49.81	0.44	0.35	0.21	67
	100	19.98	48.52	0.49	0.35	0.16	68
	200	20.02	48.79	0.58	0.31	0.11	80
Gamma	25	20.22	58.47	0.34	0.34	0.32	25
	50	19.96	53.14	0.36	0.35	0.29	32
	100	20.02	51.80	0.37	0.38	0.25	46
	200	20.02	51.23	0.37	0.41	0.22	51
Weibull	25	20.26	59.95	0.28	0.32	0.40	71
	50	20.09	54.19	0.24	0.30	0.46	80
	100	20.07	52.07	0.18	0.25	0.57	90
	200	20.05	50.89	0.13	0.20	0.67	96

*percentage of correct decision.

LGW mixture model in order to compute the evidence measures in favor of the models specified in the hypotheses.

Criteria for evaluating the performance of the FBST

In order to evaluate the performance of the FBST on selecting the true distribution used to generate the sample, we have compared the measures of evidence in favor of the hypotheses $H : p_k = 0$ and $H : p_k = 1$, $k = L, G, W$, where p_k are, respectively, the mixture weights associated with the lognormal, gamma and Weibull components in the LGW mixture model.

For instance, suppose that the sample has a lognormal distribution. We consider that the FBST has made a correct choice on the LGW model, if the evidence in favor of $H : p_L = 0$ is less than that in favor of $H : p_G = 0$ and $H : p_W = 0$, and the evidence in favor of $H : p_L = 1$ is greater than that in favor of $H : p_G = 1$ e $H : p_W = 1$. The calculation of the proportions of correct decisions made by FBST is based on 200 replicates. An analogous procedure is employed when the samples are generated from gamma or Weibull distributions.

In these simulations, we have assigned $\mu = 20$ and $\sigma^2 = 50$.

Simulation results

Table 5 presents the mean of the estimates for the LGW mixture model parameters and the percentages of correct decisions made by FBST in selecting the true distribution used to generate the samples. It is observed that, regardless of the distribution used for generating the data and the sample sizes, the estimates for the mean μ are very close to each other and to the true value of the parameter. For the estimates of the variance σ^2, we observe a variation between them but, in general, they approach the true value of the parameter as the sample size increases.

We also observe that the FBST presents good performance in identifying the Weibull distribution as the true data-generation process and low performance in identifying the gamma distribution. This happens because, for the parameters chosen for these simulations, the gamma and lognormal densities are very similar.

6. Applications

In this section we analyze two uncensored datasets and use the FBST and the Cox test to discriminate between lognormal, gamma and Weibull distributions.

Table 6. Measures of evidence provided by yarn data.

Comparison	Null hypothesis	e-value (FBST)	Standard normal deviate, T_{fg}^*	p-value (Cox test)
		Evidence in favor of null hypothesis		
$H_L \times H_W$	H_L	0.000	− 3.048	0.002
	H_W	0.997	− 0.549	0.583
$H_L \times H_G$	H_L	0.000	− 3.033	0.002
	H_G	0.871	0.773	0.439
$H_G \times H_W$	H_G	0.697	1.016	0.309
	H_W	0.725	0.967	0.333

Let us consider again the probability densities specified in the hypotheses H_L, H_W and H_G. Here, the goal is to decide which of these alternative models best fits the datasets.

Example 1. Quesenberry and Kent (1982) present a method for selecting the member of a collection of families of distribution that best fits a set of observations. A selection statistic is proposed that is essentially the value of the density function of a scale transformation maximal invariant. The dataset observations consist of experiments for testing the tensile fatigue characteristics of polyester/viscose yarn to study the problem of warp breakage during weaving. The experiment consisted of placing 100 samples of yarn into a 10-station testing apparatus that subject the yarn to 80 cycles per minute of a given strain level. The cycle at which the yarn failed (cycles-to-failure) was recorded. The FBST and the Cox test are used to compare the distributions for the data from the experiment at the 2.3 percent strain level.

Table 6 presents the Bayesian and classical measures of evidence provided by the yarn data in favor of null hypothesis on the comparisons between pairs of the distributions. For selecting between the lognormal and the Weibull distributions, we have the following results: the e-values $ev(H_L) = 0.000$ and $ev(H_W) = 0.871$, and the values of standard normal deviate for Cox's test statistics $T_{LW}^* = -3.048$ and $T_{WL}^* = -0.549$ with the corresponding p-values of 0.002 and 0.583, respectively. These results indicate rejecting the lognormal distribution and choosing the Weibull distribution, which provides the best fit to the dataset. In Quesenberry and Kent (1982), the Weibull distribution is also preferred over the lognormal distribution. Araujo and Pereira (2007) used intrinsic and fractional Bayes factors to discriminate between these distributions and also obtained very strong evidence against the lognormal distribution.

Since the comparison between the lognormal and gamma distributions suggests rejecting the lognormal model, gamma versus Weibull distributions were tested. The results of the tests indicate that both the distributions provide good fit to the dataset. Again we agree with the findings of Quesenberry and Kent (1982) and Araujo and Pereira (2007), which observed that it would be difficult to distinguish between those two models because both families of distributions fit these data equally well.

In order to simultaneously test the three hypotheses, we have applied the LGW mixture model,

$$f(y|p, \mu, \sigma) = p_1 f_L(y|\mu, \sigma) + p_2 f_G(y|\mu, \sigma) + p_3 f_W(y|\mu, \sigma), \qquad (6.1)$$

to the yarn data from the experiment at the 2.3 percent strain level.

Table 7 presents the estimates for the parameters of the model (6.1). Here, SD, 2.5% and 97.5% denote the standard deviation, the 2.5th and the 97.5th percentiles of the posterior distribution of the LGW parameters, respectively. Table 8 gives the results of hypothesis testing on the mixture weights. The p-values are calculated according to Diniz et al. (2012), as described in Section 3. The results of the tests are similar to the previous comparisons between

Table 7. Summary of the posterior distribution of the LGW parameters.

Parameter	Mean	SD	2.5%	Median	97.5%
p_1-lognormal	0.170	0.127	0.007	0.143	0.469
p_2-gama	0.381	0.249	0.018	0.355	0.869
p_3-Weibull	0.449	0.237	0.032	0.461	0.869
μ	220.423	14.239	193.679	219.966	249.759
σ^2	20665.944	4001.274	14369.620	20248.042	30166.612

Table 8. Hypothesis testing on the mixture weights of the LGW model.

Hypothesis	e-value	p-value*
$p_1 = 0$	0.652	0.116
$p_2 = 0$	0.206	0.015
$p_3 = 0$	0.073	0.003

*p-value calculated according to Diniz et al. (2012).

Table 9. Summary of the posterior distribution of Weibull parameters.

Parameter	Mean	SD	2.5%	Median	97.5%
μ	220.409	13.675	194.595	219.938	248.975
σ^2	19862.278	3170.874	14708.324	19523.516	27038.00

pairs of the distributions. Both the classical and the Bayesian measures of evidence indicate that, among the three models, the lognormal model is the one that should not be considered because the null hypothesis $H : p_1 = 0$ is not rejected.

Figure 1 displays the survival curves calculated using Bayesian estimates of the Weibull model (Table 9), the LGW mixture model (Table 7) and a procedure called the piecewise exponential estimator (PEXE), introduced by Kim and Proschan (1976), representing the observed data. Unlike the well-known Kaplan-Meier estimator, the PEXE is smooth and continuous estimator of the survival function. It appears that the Weibull model by itself produces a good estimate of survival function.

The results from Tables 6 and 8 show that the preference for the Weibull model is quite clear in evaluating the three-component mixture model more than in the two-component model

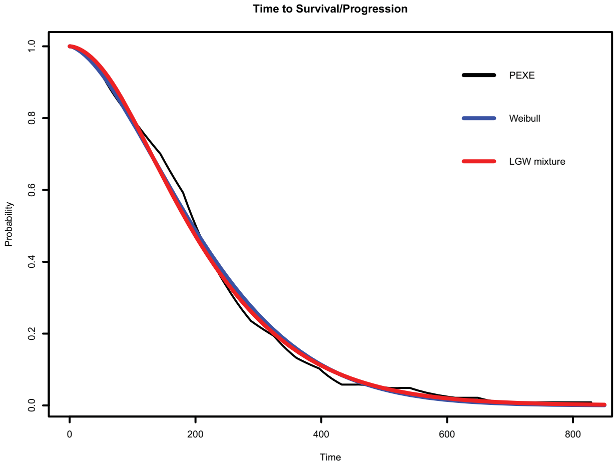

Figure 1. Survival curves based on estimates of the Weibull model, the LGW model and PEXE for yarn data.

Table 10. Hypothesis testing on the mixture weights of the LGW model.

Hypothesis	e-value	p-value*
$p_1 = 0$	0.834	0.227
$p_2 = 0$	0.856	0.249
$p_3 = 0$	0.078	0.004

*p-value calculated according to Diniz et al. (2012).

(comparison $H_G \times H_W$), where the evidence measures in favor of both models are very close. This means that the discrimination power provided by the LGW model is much higher than the power of the pairwise comparisons. This finding is in agreement with the discussion of Sawyer (1984).

Example 2 (Lagakos, Barraj, and De Gruttola 1988). This dataset contains the induction times of AIDS in patients infected by contaminated blood transfusions. The times are for 258 adults and 37 children (less than 5 years old), infected by June 30, 1985, given by US Center for Disease Control.

Pereira (1997) analyzed the data of the adult population ($n = 258$) and used Cox tests to discriminate between $H_G \times H_W$, with the result indicating that the Weibull distribution is preferable. Araujo and Pereira (2007) used the intrinsic and fractional Bayes factors to discriminate between these distributions and also obtained positive evidence against the gamma distribution.

Here, the LGW model is applied to the data of induction times for 258 adults and the FBST is used to discriminate between the distributions by testing hypotheses on the mixture weights. The results, given in Table 10, indicate that neither the lognormal nor gamma models should be considered, because the null hypotheses $H_L : p_1 = 0$ and $H_G : p_2 = 0$ are not rejected. Consequently, among the three models, the Weibull model should be chosen for further analyses of the data. From Figure 2, it seems reasonable to disregard both the lognormal and gamma models, since the Weibull model by itself produces a good estimate of the survival function.

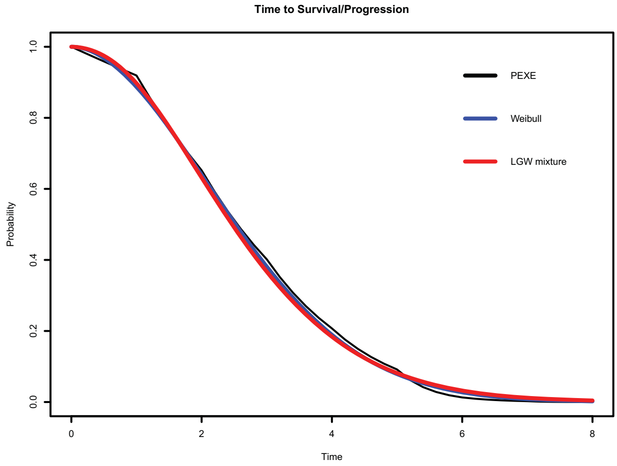

Figure 2. Survival curves based on estimates of the Weibull, the LGW model and the PEXE for induction times of AIDS.

7. Final remarks

In this paper we considered the problem of discriminating between separate families of distributions. We analyzed this problem in the context of a convex linear combination of the candidate models. The families of distributions considered include the lognormal, gamma and Weibull models. In order to discriminate between the three separate models, we used the FBST for testing the hypotheses defined on the weights space.

The simulation results indicate that both the FBST and Cox test have similar behavior in discriminating between separate models. Nevertheless, the discrimination power of the FBST is slightly higher than those of the Cox test, mainly for small sample sizes. For selecting based on the lognormal-gamma-Weibull mixture model, the FBST achieved good performance in identifying the true distribution used to generate the data. In the examples with real datasets, the FBST reached the same conclusion as the other selection procedures used by Quesenberry and Kent (1982), Araujo and Pereira (2007) and Pereira (1997). Therefore, our proposed selection procedure can be used effectively for discriminating between separate models even when the sample size is small.

When using the FBST for discriminating between separate models, it is recommended to apply a mixture model including all candidate models in order to avoid the problems that arise when pairwise comparisons are performed (Sawyer 1984). Whenever possible, we also recommend reparametrizing the models in terms of the common parameters.

It would be interesting to compare the proposed procedure with other selection procedures that allow the use of data with censoring mechanisms.

Acknowledgements

The authors are grateful to Mark Gannon for his support in technical and language editing, and proofreading of this manuscript, to referees and the editor for many perceptive comments and suggestions that improved the article. C.C. Assane and C. A. B. Pereira thank research fellowships from CNPq - Brazilian Research Funding Agency - processes 141349/2013-1 and 302767/2017-7, respectively.

References

Araujo, M. I., B. B. Pereira, R. Cleroux, M. Fernandes, and A. Lazraq. 2005. Separate families of models: Sir David Cox contributions and recent developments. *Student* 5:251–8.

Araujo, M. I., and B. B. Pereira. 2007. A comparison of Bayes factors for separated models: Some simulation results. *Communications in Statistics–Simulation and Computation* 36:297–309.

Assane, C. C., B. B. Pereira, and C. A. B. Pereira. 2017. Bayesian significance test for discriminating between survival distributions. *Communications in Statistics–Theory and Methods*, doi: 10.1080/03610926.2017.1406117.

Berger, J. O., and M. Delampady. 1987. Testing precise hypothesis. *Statistical Science* 2:317–52.

Chakrabarty, D. 2016. A new Bayesian test to test for the intractability-countering hypothesis. *Journal of the American Statistical Association* 36:561–77. doi:10.1080/01621459.2016.1240684

Cox, D. R. 1961. Tests of separate families of hypotheses. *Proceedings 4th Berkeley Symposium in Mathematical Statistics and Probability* 1:105–23.

Cox, D. R. 1962. Further results on test of separate families of hypotheses. *Journal of the Royal Statistical Society* B:406–24.

Cox, D. R. 1977. The role of significance tests. *Scand. J. Statist* 4:49–70.

Diniz, M., C. A. B. Pereira, A. Polpo, J. M. Stern, and S. Wechsler. 2012. Relationship between Bayesian and frequentist significance indices. *International Journal for Uncertainty Quantification* 2:161–72.

Fletcher, R., and C. M. Reeves. 1964. Function minimization by conjugate gradients. *Computer Journal* 7:148–54.

Haario, H., E. Saksman, and J. Tamminen. 2001. An adaptive Metropolis algorithm. *Bernoulli* 7:223–42.

Irony, T. Z., M. Lauretto, C. A. B. Pereira, and J. M. Stern. 2002. A Weibull wearout test: Full Bayesian approach. In *Systems and Bayesian reliability*, ed. Y. Hayakawa, T. Z. Irony, and M. Xie, 287–300. Singapore: World Scientific.

Kamary, K., K. Mengersen, C. P. Robert, and J. Rousseau. 2014. Testing hypotheses via a mixture estimation model. *arXiv:1412.2044v2*.

Kempthorne, O. 1976. Of what use are tests of significance and tests of hypothesis. *Communications in Statistics -Theory and Methods* 8:763–77.

Kim, J. S., and F. Proschan. 1991. Piecewise exponential estimator of the survivor function. *IEEE Transactions on Reliability* 40:134–9.

Lagakos, S. W., L. M. Barraj, and V. De Gruttola. 1988. Nonparametrics analysis of truncated survival data, with application to AIDS. *Biometrika* 75:515–23.

Lauretto, M., C. A. B. Pereira, and J. M. Stern. 2008. The FBST for mixture models: Results in gene expression clustering. *Genetics & Molecular Research* 7:883–97.

Lauretto, M., C. A. B. Pereira, J. M. Stern, and S. Zacks. 2003. Comparing parameters of two bivariate normal distributions using the invariant full Bayesian significance test. *Brazilian Journal of Probability and Statistics* 17:147–68.

Lauretto, M. S., S. R. Faria Jr, B. B. Pereira, C. A. B. Pereira, and J. M. Stern. 2007. The problem of separate hypotheses via mixtures models. In *Bayesian inference and maximum entropy methods in science and engineering*ed. K. H. Knuth, A. Caticha, J. L. Center-Jr, A. Giffin, and C. C. Rodriguez, 268–75. Melville: American Institute of Physics Press.

Lawless, J. F. 2002. *Statistical models and methods for lifetime data*, 2nd ed. New York: John Wiley and Sons.

Madruga, M., C. A. B. Pereira, and J. M. Stern. 2003. Bayesian evidence test for precise hypotheses. *Journal of Statistical Planning and Inference* 117:185–98.

Pereira, B. B. 1978. Empirical comparisons of some tests of separate families of hypotheses. *Metrika* 25:219–34.

Pereira, B. B. 1981. Choice of a survival model for patients with a brain tumour. *Metrika* 28:53–61.

Pereira, B. B. 1997. *On the distribution of the incubation period for AIDS patient*, Techinic report, 101, LES, IM-UFRJ.

Pereira, B. B. 2005. Separate families of hypotheses. *Encyclopedia of Biostatistics* 7:4881–6.

Pereira, B. B., and C. A. B. Pereira. 2017. *Model choice in nonnested families*, 1st ed. Berlin: Springer.

Pereira, C. A. B., and J. Stern. 1999. Evidence and credibility: Full Bayesian significance test for precise hypotheses. *Entropy* 1:69–80.

Pereira, C. A. B., and J. M. Stern. 2008. Special characterizations of standard discrete models. *BRevStat - Statistical Journal* 6:199–230.

Pereira, C. A. B., J. Stern, and S. Wechsler. 2008. Can a significance test be genuinely Bayesian. *Bayesian Analysis* 3:79–100.

Quesenberry, C. P., and J. Kent. 1982. Selecting among probability distributions used in reliability. *Technometrics* 24:59–65.

Sawyer, K. R. 1984. Multiple hypotheses testing. *Journal of teh Royal Statistical, Society-B* 46:419–24.

Statisticat, LCC 2016. *LaplacesDemon: A complete environment for Bayesian inference within R*. R Package version 17.07.2016. https://cran.r-project.org/web/packages/LaplacesDemon/LaplacesDemon.pdf.

Stern, J., and C. A. B. Pereira. 2014. Bayesian epistemic values: Focus on surprise, measure probability. *Logic Journal of The IGPL* 22:236–54.