A pretest for using logrank or Wilcoxon in the two-sample problem

Annie Tordilla Darilay *, Joshua D. Naranjo
Department of Statistics, Western Michigan University, Kalamazoo, MI 49009, USA

**ABSTRACT**

In a two-sample location-scale model with censored data, the logrank test is asymptotically efficient when the error distribution is extreme minimum value. On the other hand, the Wilcoxon test is asymptotically efficient when the error distribution is logistic. We propose a pretest for choosing between logrank and Wilcoxon by determining if the error distribution is closer to extreme minimum value or logistic. This adaptive test is compared with the logrank and Wilcoxon tests through simulation.

1. Introduction

In survival analysis, treatment efficacy is often analyzed by comparing the survival rates of two treatment groups. Two commonly used tests for the comparison of survival distributions are the generalized Wilcoxon procedure (Gehan, 1965; Peto and Peto, 1972) and the logrank test (Mantel, 1966; Cox, 1972; Peto and Peto, 1972). Both tests are based on the ranks of the observations, and have several versions in the literature. In this paper, the Wilcoxon and logrank tests will refer to Peto and Peto’s (1972) version of the statistics proposed by Gehan (1965) and Mantel (1966), respectively. Leton and Zuluaga (2001, 2005) present a comprehensive summary of the different names, versions and representations of the generalized Wilcoxon and logrank tests.

The finite sample performance of these tests have been compared in several simulation studies. Lee et al. (1975) compared size and power of the tests using small samples from exponential and Weibull survival distributions with and without censoring. Latta (1981) extended the simulations to include the lognormal survival distribution, allow for unequal sample sizes, and allow for censoring in only one sample. Beltangady and Frankowski (1989) focussed on the effect of unequal censoring, using various combinations of censoring proportions. More recently, Leton and Zuluaga (2001, 2005) compared the performance of various versions of the generalized Wilcoxon and logrank tests under scenarios of early and late hazard differences.

In general, the logrank test is most powerful when the ratio of the two hazard functions is constant. When hazard ratio is nonconstant, the generalized Wilcoxon test may perform better than the logrank test especially when differences occur early in time (Lee et al., 1975; Leton and Zuluaga, 2001, 2005). In applications, the logrank test is often used after checking for validity of the proportional hazards (PH) assumption, with Wilcoxon being the fallback method when the PH assumption fails. There have been several graphical methods suggested for assessing the proportional hazards assumption (Hess, 1995). One commonly used graphical method that is available on many statistical software (i.e. SAS, Stata and R) is the plotting of the log of the cumulative hazard function against log time and checking for parallelism.
The properties of the logrank and Wilcoxon tests are discussed extensively in the literature (Breslow, 1970; Cox, 1972; Peto, 1972; Peto and Peto, 1972; Kalbfleisch and Prentice, 1980; Andersen et al., 1993; Klein and Moeschberger, 2003). Under a location-scale model, \( y = \mu + \beta z + \sigma \epsilon \), where \( \epsilon \) is right censored, the logrank test is known to be fully efficient for the extreme minimum value error distribution \( f(\epsilon) = \exp(\epsilon - \epsilon^*) \), while Peto and Peto’s Wilcoxon test is fully efficient for the logistic error distribution \( f(\epsilon) = \epsilon^*/(1 + \epsilon^*)^2 \) (Prentice, 1978; Gill, 1980). Since the efficiency of rank tests is invariant under monotone increasing data transformations, this also means that the logrank test is fully efficient when the failure time follows a Weibull distribution and Peto and Peto’s Wilcoxon test is fully efficient when the failure time follows a loglogistic distribution.

Prentice (1975) discussed a general family of distributions that contain both the extreme minimum value and logistic distributions. Maximum likelihood estimates of parameters can be used as a diagnostic for discriminating between the extreme minimum value and logistic in the context of one sample. In this paper, we extend his procedure to the two-sample problem and propose a statistic for detecting whether the common error distribution is closer to the extreme minimum value or logistic. Discrimination between Weibull and loglogistic distributions under the accelerated failure time model is equivalent to discriminative between their error distributions: extreme minimum value and logistic distributions. In this section, we will present Prentice’s (1975) discrimination procedure that embeds both extreme minimum value and logistic distributions in a larger parametric family of distributions.

### 2. Weibull versus loglogistic distribution

Discriminating between Weibull and loglogistic distributions under the accelerated failure time model is equivalent to discriminating between their error distributions: extreme minimum value and logistic distributions. In this section, we will present Prentice’s (1975) discrimination procedure that embeds both extreme minimum value and logistic distributions in a larger parametric family of distributions.

#### 2.1. Parametric family of distributions

Consider the location-scale model, \( y = \mu + \sigma \epsilon \), such that

\[
f(\epsilon) = \frac{1}{mB(1, m)} \epsilon^e \left( 1 + \frac{\epsilon^e}{m} \right)^{-(m+1)}
\]

where \( m > 0 \) and \( B \) is the beta function \( B(a, b) = \int_0^1 t^{a-1}(1 - t)^{b-1} dt \). The density of \( Y \) is

\[
f(y) = \frac{1}{\sigma mB(1, m)} e^{(y-\mu)/\sigma} \left( 1 + \frac{e^{(y-\mu)/\sigma}}{m} \right)^{-(m+1)}
\]

The distribution in (2) is a subset of a larger family of distributions discussed in Prentice (1975). Table 1 lists special cases of the distribution of \( Y \) and \( \exp(Y) \).

Since the logistic and extreme minimum value distributions correspond to \( m = 1 \) and \( m = \infty \), respectively, a diagnostic for choosing between these two models can be based on an estimate of \( m \) being closer to 1 or \( \infty \).

#### 2.2. Maximum likelihood estimation of \( m \)

Let \( y_1, y_2, \ldots, y_n \) be a random sample from the distribution (2). The likelihood of the sample is

\[
L(y; \mu, \sigma, m) = \left( \frac{1}{\sigma mB(1, m)} \right)^n \prod_{i=1}^n \left[ e^{(y_i-\mu)/\sigma} \left( 1 + \frac{e^{(y_i-\mu)/\sigma}}{m} \right)^{-(m+1)} \right].
\]
Taking partial derivatives of the log-likelihood and equating to 0, we get the following estimating equations

\[
\frac{\partial \ln L}{\partial \mu} = -\frac{n}{\sigma} + \frac{(1 + m)}{\sigma} \sum_{i=1}^{n} \frac{1}{1 + \frac{1}{m} e^{\epsilon_i}} \left( e^{\epsilon_i} \right) = 0, \tag{4}
\]

\[
\frac{\partial \ln L}{\partial \sigma} = -\sum_{i=1}^{n} \frac{\epsilon_i}{\sigma} + \frac{(1 + m)}{\sigma} \sum_{i=1}^{n} \frac{\epsilon_i}{1 + \frac{1}{m} e^{\epsilon_i}} \left( e^{\epsilon_i} / m \right) - \frac{n}{\sigma} = 0, \tag{5}
\]

\[
\frac{\partial \ln L}{\partial m} = -n + (1 + m) \sum_{i=1}^{n} \frac{e^{\epsilon_i}}{m^2 + me^{\epsilon_i}} - \sum_{i=1}^{n} \ln \left( 1 + \frac{1}{m} e^{\epsilon_i} \right) - n[\psi(m) - \psi(1 + m)] = 0 \tag{6}
\]

where \( \epsilon_i = (y_i - \mu) / \sigma \). To obtain (6), note that \( \partial B(a, b) / \partial b = B(a, b)[\psi(b) - \psi(a + b)] \), where \( \psi(k) = \partial \ln \Gamma(k)/\partial(k) = \int_{0}^{\infty} [e^{-t} - 1 - e^{-k} - e^{-k}(1 - e^{-t})] dt \) is the digamma function (Davis, 1965).

The MLE \( \hat{m} \) can be obtained simultaneously with \( \hat{\mu} \) and \( \hat{\sigma} \) using numerical methods. For our simulation, we will use R’s \textit{optim} function which outputs the three maximum likelihood estimates based on the inputted log-likelihood function and the estimating Eqs. (4)–(6).

To confirm that \( \hat{m} \) can indeed discriminate between the extreme minimum value and logistic distributions, we generated 50 observations from either distribution and estimated \( \hat{m} \). The simulation was replicated 5000 times for each of the following cases:

CaseA. \( \mu = 0 \), \( \sigma < 1 \) (pdf of both distributions is narrow and tall)

CaseB. \( \mu = 0 \), \( \sigma > 1 \) (pdf of both distributions is broad and shallow).

Observe from Table 2 that \( \hat{m} \) tends to be small under the logistic distribution (80% are smaller than 2.0), and tends to be quite large under the extreme minimum value distribution (95% are larger than 2.0).

The asymptotic distribution of \( \hat{m} \) can be derived using standard maximum likelihood theory. More precisely, we present the asymptotic distribution of \( 1/\hat{m} \), which has finite mean. If \( m = 1 \),

\[
\frac{1}{\hat{m}} \sim N \left( 1, \frac{4}{n} \left( \frac{\pi^2 + 3}{\pi^2 - 6} \right) \right), \tag{7}
\]

approximately when \( n \) is large. If \( m = \infty \),

\[
\frac{1}{\hat{m}} \sim N \left( 0, \frac{1}{n \left( 1 - \frac{6}{\pi^2} \right)} \right) \tag{8}
\]

approximately. See Prentice (1975) for details. The asymptotic variances above evaluate to 13.30/\( n \) and 2.55/\( n \), respectively.

Eqs. (7) and (8) and Table 2 suggest a pretest for logistic or extreme minimum value depending on whether \( 1/\hat{m} > 0.5 \), or equivalently, \( \hat{m} < 2.0 \).

In Sections 3 and 4, we investigate the properties of an adaptive test which uses Wilcoxon when \( \hat{m} < 2.0 \) and logrank otherwise.

3. Adaptive survival test

Let \( T_1 = (T_{11}, T_{12}, \ldots, T_{1m}) \) and \( T_2 = (T_{21}, T_{22}, \ldots, T_{2m}) \) be random samples of failure times whose log values are denoted by the following respective location-scale models:

### Table 2
Percentiles of \( \hat{m} \) at \( \mu = 0 \).

<table>
<thead>
<tr>
<th>Percentile</th>
<th>( \sigma = 0.5 )</th>
<th></th>
<th>( \sigma = 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Extreme minimum value (m = ( \infty ))</td>
<td>Logistic (m = 1)</td>
<td>Extreme minimum value (m = ( \infty ))</td>
</tr>
<tr>
<td>5th</td>
<td>2.2716</td>
<td>0.4310</td>
<td>2.1767</td>
</tr>
<tr>
<td>10th</td>
<td>3.1288</td>
<td>0.5283</td>
<td>3.2156</td>
</tr>
<tr>
<td>20th</td>
<td>4.8244</td>
<td>0.6590</td>
<td>5.0572</td>
</tr>
<tr>
<td>30th</td>
<td>13.7306</td>
<td>0.7733</td>
<td>13.5070</td>
</tr>
<tr>
<td>40th</td>
<td>37.6808</td>
<td>0.8871</td>
<td>35.2304</td>
</tr>
<tr>
<td>50th</td>
<td>18.1599</td>
<td>1.0165</td>
<td>40.5449</td>
</tr>
<tr>
<td>60th</td>
<td>37.6808</td>
<td>1.1816</td>
<td>61.7599</td>
</tr>
<tr>
<td>70th</td>
<td>30.6609</td>
<td>1.4183</td>
<td>75.8272</td>
</tr>
<tr>
<td>80th</td>
<td>36.0624</td>
<td>1.7928</td>
<td>89.9091</td>
</tr>
<tr>
<td>90th</td>
<td>42.4047</td>
<td>2.0774</td>
<td>106.8640</td>
</tr>
<tr>
<td>95th</td>
<td>47.3248</td>
<td>2.1804</td>
<td>132.9900</td>
</tr>
</tbody>
</table>
\[
\log T_1 = \mu + \sigma \epsilon \\
\log T_2 = \mu + \beta + \sigma \epsilon,
\]

where the distribution of \( \epsilon \) is unknown.

Based on (9) and (10), the relationship between \( \log T_1 \) and \( \log T_2 \) is characterized by the location-change model

\[ \log T_2 = \beta + \log T_1 \]

while the relationship between \( T_1 \) and \( T_2 \) is characterized by the scale-change model

\[ T_2 = e^{\beta} T_1. \tag{11} \]

In testing \( \beta = 0 \), it is known that the logrank test is asymptotically efficient for the extreme minimum value error distribution while Peto–Peto’s Wilcoxon test is asymptotically efficient for the logistic error distribution (Prentice, 1978).

Now we can use the diagnostic \( \hat{m} \) on the combined log failure times to decide whether the extreme minimum value or logistic distribution better fits the data. However, the two samples cannot be combined before they are standardized to the same scale.

From model (11), the scale-change parameter or acceleration factor is \( e^\beta = T_2/T_1 \), which can be intuitively estimated by the ratio of the two standard deviations, i.e. \( SD(T_2)/SD(T_1) \). Hence, multiplying \( T_1 \) by this ratio,

\[ T_1^* = \frac{SD(T_2)}{SD(T_1)} T_1, \]

results in a rescaled \( T_1^* \) which exhibits the same variability as \( T_2 \).

We now propose an adaptive test procedure for \( H_0 : \beta = 0 \). Let \( T_1 \) and \( C_1 \) be the failure and censored times of the first sample \( D_1 \). Let \( T_2 \) and \( C_2 \) be the failure and censored times of the second sample \( D_2 \).

a. Standardize the failure times from the two samples so they have the same variability:

\[ T_1^* = T_1 \times \frac{SD(T_2)}{SD(T_1)}. \]

b. Combine the standardized failure times from the two samples \( T_c = (T_1^* \text{ and } T_2) \).

c. Log-transform \( T_c \) and compute \( \hat{m} \).

d. If \( \hat{m} < 2.0 \), use the Wilcoxon test on \( D_1 \) and \( D_2 \). Otherwise, use the logrank test.

We chose the Peto–Peto version of the logrank and Wilcoxon tests. Unlike the Gehan version, Peto–Peto’s Wilcoxon test does not depend on the censoring rates and thus does not give misleading results when the censoring patterns are different (Prentice and Marek, 1979).

4. Simulation

In this section, we will examine the finite sample performance of the adaptive survival test for the two-sample accelerated failure time problem. We will assess its empirical validity, and compare its power against the dedicated Wilcoxon and logrank tests, respectively.

Failure times, \( T_1 \) and \( T_2 \), were generated from the following distributions:

a. Weibull

\[ f(t) = \frac{1}{\sigma} e^{-(\mu/\sigma)t} \left(\frac{1}{\sigma} \right)^{1/\sigma} \exp\left[-e^{-(\mu/\sigma)t} \left(\frac{1}{\sigma}\right)^{1/\sigma}\right], \quad \sigma > 0. \]

b. Loglogistic

\[ f(t) = \frac{1}{\sigma} e^{-(\mu/\sigma)t} \left(\frac{1}{\sigma} \right)^{1/\sigma} \left[1 + e^{-(\mu/\sigma)t} \left(\frac{1}{\sigma}\right)^{1/\sigma}\right]^{-2}, \quad \sigma > 0. \]

c. Lognormal

\[ f(t) = \frac{1}{t_\sigma \sqrt{2\pi}} e^{-0.5[\log(t) - \mu]^2/\sigma^2}, \quad \sigma > 0. \]

For each of these failure time distributions, the shape parameter \( \sigma \) is set to (i) greater than 1, (ii) equal to 1, and (iii) less than 1 to represent different degrees of skewness and tail weight of the survival function. The scale parameter for \( S_1 \) is \( \mu = 0 \) in all cases. In the size study, the scale parameter for \( S_2 \) is also \( \mu = 0 \) while in the power study, the scale parameter for \( S_2 \) is \( \mu > 0 \).

The simulations were also done using different levels of right censoring. The censoring distributions of the two groups were equal and uniform \( U(0, c) \). The values of \( c \) were used to control the desired censoring percentages considered in the simulation.
Tables 5–7

survival distribution while the Wilcoxon does better under the log-logistic distribution. When the distribution is lognormal, the percentage of censored observations. Between the logrank and Wilcoxon tests, the logrank test is better under the Weibull survival distribution while the Wilcoxon does better under the log-logistic distribution. When the distribution is lognormal, the logrank test is better under the Weibull distribution. The adaptive test is more powerful than logrank. Generally, as the censoring rate increases, the power of the logrank test increases.

The simulation results support the known result that Wilcoxon is efficient under the log-logistic survival distribution. The Wilcoxon test tends to be the most conservative in almost all cases.

4.1. Size study

Each test was conducted on 10,000 samples under each survival and censoring configuration. Results at the 5% significance level for sample sizes 20 and 50 are presented in Tables 3 and 4, respectively. All simulations were done in R.

At n = 20, all 3 tests seem to be a touch liberal, with the size of the adaptive test slightly larger than either non-adaptive test (though not by much). The Wilcoxon test tends to be the most conservative in almost all cases.

At n = 50, the observed significance level of all tests are closer to the nominal value of 0.05, with the Wilcoxon again being the most conservative.

4.2. Power study

The power of the tests at a 5% significance level was obtained from 10,000 samples for each survival and censoring configuration. The results for Weibull, log-logistic, and lognormal are displayed on Tables 5–7.

The logrank test is known to be efficient under the Weibull survival distribution. Our simulations confirm this, with logrank outperforming Wilcoxon in all cases. The difference in their power is minimal under heavily skewed Weibull distributions with censoring but increases up to 10% as the degree of skewness decreases. The adaptive test approximates the power of the logrank test, and is better than Wilcoxon in this situation.

The simulation results support the known result that Wilcoxon is efficient under the log-logistic survival distribution. Wilcoxon beats logrank by up to 10% in the uncensored cases and up to 4% in the censored cases. At n = 50, all tests perform well under lightly skewed survival distributions, with Wilcoxon beating logrank by 1% to 2% efficiency. The adaptive test approximates the power of the efficient Wilcoxon in the uncensored cases. Under censoring, the adaptive test is either better than or comparable to logrank.

Under the lognormal survival distribution, Wilcoxon beats logrank by 2% to 6% efficiency, depending on the degree of skewness of the survival distribution and the censoring rate. The adaptive test is more powerful than logrank. Generally, as the censoring rate decreases or as the survival distribution becomes less skewed, the gain in power of the adaptive test against logrank increases.

As expected, the power of all three tests decrease with increasing skewness of the survival distribution and increasing percentage of censored observations. Between the logrank and Wilcoxon tests, the logrank test is better under the Weibull survival distribution while the Wilcoxon does better under the log-logistic distribution. When the distribution is lognormal,
### Table 4
Size simulation results at significance level $\alpha = 0.05$, from 10,000 replications with equal censoring, $n_1 = n_2 = 50$.

<table>
<thead>
<tr>
<th>Relative skewness</th>
<th>Survival time</th>
<th>Average % censored</th>
<th>Adaptive</th>
<th>Logrank</th>
<th>Wilcoxon</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S(t) = \text{Weibull distribution}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heavy</td>
<td>$\mu_1 = 0, \sigma_1 = 2$</td>
<td>$C_1 = 0.0, C_2 = 0.0$</td>
<td>0.0545</td>
<td>0.0542</td>
<td>0.0485</td>
</tr>
<tr>
<td></td>
<td>$\mu_2 = 0, \sigma_2 = 2$</td>
<td>$C_1 = 29.8, C_2 = 29.7$</td>
<td>0.0552</td>
<td>0.0552</td>
<td>0.0556</td>
</tr>
<tr>
<td>Moderate</td>
<td>$\mu_1 = 0, \sigma_1 = 1$</td>
<td>$C_1 = 41.4, C_2 = 41.3$</td>
<td>0.0489</td>
<td>0.0489</td>
<td>0.0484</td>
</tr>
<tr>
<td></td>
<td>$\mu_2 = 0, \sigma_2 = 1$</td>
<td>$C_1 = 0.0, C_2 = 0.0$</td>
<td>0.0556</td>
<td>0.0522</td>
<td>0.0496</td>
</tr>
<tr>
<td>Light</td>
<td>$\mu_1 = 0, \sigma_1 = 0.75$</td>
<td>$C_1 = 16.6, C_2 = 16.4$</td>
<td>0.0540</td>
<td>0.0541</td>
<td>0.0510</td>
</tr>
<tr>
<td></td>
<td>$\mu_2 = 0, \sigma_2 = 0.75$</td>
<td>$C_1 = 43.3, C_2 = 43.3$</td>
<td>0.0502</td>
<td>0.0502</td>
<td>0.0488</td>
</tr>
<tr>
<td>$S(t) = \text{Logistic distribution}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heavy</td>
<td>$\mu_1 = 0, \sigma_1 = 1.5$</td>
<td>$C_1 = 30.5, C_2 = 30.5$</td>
<td>0.0513</td>
<td>0.0512</td>
<td>0.0476</td>
</tr>
<tr>
<td></td>
<td>$\mu_2 = 0, \sigma_2 = 1.5$</td>
<td>$C_1 = 43.4, C_2 = 43.5$</td>
<td>0.0519</td>
<td>0.0520</td>
<td>0.0491</td>
</tr>
<tr>
<td>Moderate</td>
<td>$\mu_1 = 0, \sigma_1 = 1$</td>
<td>$C_1 = 24.0, C_2 = 24.0$</td>
<td>0.0525</td>
<td>0.0526</td>
<td>0.0518</td>
</tr>
<tr>
<td></td>
<td>$\mu_2 = 0, \sigma_2 = 1$</td>
<td>$C_1 = 40.4, C_2 = 40.4$</td>
<td>0.0483</td>
<td>0.0485</td>
<td>0.0462</td>
</tr>
<tr>
<td>Light</td>
<td>$\mu_1 = 0, \sigma_1 = 0.5$</td>
<td>$C_1 = 14.7, C_2 = 14.8$</td>
<td>0.0555</td>
<td>0.0543</td>
<td>0.0514</td>
</tr>
<tr>
<td></td>
<td>$\mu_2 = 0, \sigma_2 = 0.5$</td>
<td>$C_1 = 33.2, C_2 = 33.1$</td>
<td>0.0526</td>
<td>0.0531</td>
<td>0.0510</td>
</tr>
<tr>
<td>$S(t) = \text{Lognormal distribution}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heavy</td>
<td>$\mu_1 = 0, \sigma_1 = 1.5$</td>
<td>$C_1 = 25.8, C_2 = 25.8$</td>
<td>0.0519</td>
<td>0.0516</td>
<td>0.0525</td>
</tr>
<tr>
<td></td>
<td>$\mu_2 = 0, \sigma_2 = 1.5$</td>
<td>$C_1 = 39.6, C_2 = 39.4$</td>
<td>0.0563</td>
<td>0.0559</td>
<td>0.0552</td>
</tr>
<tr>
<td>Moderate</td>
<td>$\mu_1 = 0, \sigma_1 = 1$</td>
<td>$C_1 = 19.5, C_2 = 19.6$</td>
<td>0.0541</td>
<td>0.0529</td>
<td>0.0491</td>
</tr>
<tr>
<td></td>
<td>$\mu_2 = 0, \sigma_2 = 1$</td>
<td>$C_1 = 35.1, C_2 = 35.0$</td>
<td>0.0583</td>
<td>0.0578</td>
<td>0.0546</td>
</tr>
<tr>
<td>Light</td>
<td>$\mu_1 = 0, \sigma_1 = 0.75$</td>
<td>$C_1 = 16.5, C_2 = 16.5$</td>
<td>0.0579</td>
<td>0.0524</td>
<td>0.0503</td>
</tr>
<tr>
<td></td>
<td>$\mu_2 = 0, \sigma_2 = 0.75$</td>
<td>$C_1 = 32.0, C_2 = 32.0$</td>
<td>0.0555</td>
<td>0.0535</td>
<td>0.0533</td>
</tr>
</tbody>
</table>

### Table 5
Power simulation results for Weibull distribution at significance level $\alpha = 0.05$, from 10,000 replications with equal censoring.

<table>
<thead>
<tr>
<th>Relative skewness</th>
<th>Survival time</th>
<th>Average % censored</th>
<th>Adaptive</th>
<th>Logrank</th>
<th>Wilcoxon</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1 = n_2 = 20$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heavy</td>
<td>$\mu_1 = 0, \sigma_1 = 2$</td>
<td>$C_1 = 23.5, C_2 = 23.5$</td>
<td>0.2158</td>
<td>0.2141</td>
<td>0.1736</td>
</tr>
<tr>
<td></td>
<td>$\mu_2 = 0.75, \sigma_2 = 2$</td>
<td>$C_1 = 29.6, C_2 = 22.2$</td>
<td>0.1655</td>
<td>0.1655</td>
<td>0.1485</td>
</tr>
<tr>
<td>Moderate</td>
<td>$\mu_1 = 0, \sigma_1 = 1$</td>
<td>$C_1 = 16.8, C_2 = 33.2$</td>
<td>0.5122</td>
<td>0.5137</td>
<td>0.4469</td>
</tr>
<tr>
<td></td>
<td>$\mu_2 = 0.75, \sigma_2 = 1$</td>
<td>$C_1 = 24.5, C_2 = 44.9$</td>
<td>0.4630</td>
<td>0.4644</td>
<td>0.4083</td>
</tr>
<tr>
<td>Light</td>
<td>$\mu_1 = 0, \sigma_1 = 0.75$</td>
<td>$C_1 = 15.4, C_2 = 32.1$</td>
<td>0.7535</td>
<td>0.7546</td>
<td>0.6688</td>
</tr>
<tr>
<td></td>
<td>$\mu_2 = 0.75, \sigma_2 = 0.75$</td>
<td>$C_1 = 22.9, C_2 = 45.9$</td>
<td>0.6990</td>
<td>0.7003</td>
<td>0.6294</td>
</tr>
<tr>
<td>$n_1 = n_2 = 50$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heavy</td>
<td>$\mu_1 = 0, \sigma_1 = 2$</td>
<td>$C_1 = 29.7, C_2 = 38.1$</td>
<td>0.2329</td>
<td>0.2324</td>
<td>0.1897</td>
</tr>
<tr>
<td></td>
<td>$\mu_2 = 0.5, \sigma_2 = 2$</td>
<td>$C_1 = 41.4, C_2 = 49.7$</td>
<td>0.1527</td>
<td>0.1527</td>
<td>0.1440</td>
</tr>
<tr>
<td>Moderate</td>
<td>$\mu_1 = 0, \sigma_1 = 1$</td>
<td>$C_1 = 24.6, C_2 = 37.7$</td>
<td>0.5344</td>
<td>0.5344</td>
<td>0.4698</td>
</tr>
<tr>
<td></td>
<td>$\mu_2 = 0.5, \sigma_2 = 1$</td>
<td>$C_1 = 43.4, C_2 = 57.9$</td>
<td>0.4027</td>
<td>0.4027</td>
<td>0.3840</td>
</tr>
<tr>
<td>Light</td>
<td>$\mu_1 = 0, \sigma_1 = 0.75$</td>
<td>$C_1 = 22.9, C_2 = 37.1$</td>
<td>0.7756</td>
<td>0.7758</td>
<td>0.7028</td>
</tr>
<tr>
<td></td>
<td>$\mu_2 = 0.5, \sigma_2 = 0.75$</td>
<td>$C_1 = 43.7, C_2 = 61.6$</td>
<td>0.6203</td>
<td>0.6208</td>
<td>0.5802</td>
</tr>
</tbody>
</table>
the Wilcoxon is better than the logrank. This is probably because the hazard function of the lognormal distribution is very similar to that of the loglogistic distribution (Klein and Moeschberger, 2003).

The adaptive test tends to fall somewhere in the middle under all distributions considered. The adaptive test is better than logrank under loglogistic and lognormal distributions, and better than Wilcoxon under the Weibull.

5. Application to catheter placement study

Let us consider the data on a clinical trial of the effectiveness of two catheterization procedures in kidney dialysis patients, taken from page 6 of Klein and Moeschberger (2003). The time to first exit-site infection (in months) was observed from 43
patients who utilized a surgically placed catheter (Group 1) and 76 patients who utilized a percutaneous placed catheter (Group 2).

We shall perform the adaptive survival test using the R code in the Appendix to test if there is a difference in the time to infection between the two groups. The calling program standardizes the two samples and calls the `getmle` function to obtain the $\hat{m}$ discrimination statistic. Based on $\hat{m}$, the `adaptivetest` function selects between logrank and Peto–Peto’s Wilcoxon test statistics, which are both derived from Harrington and Fleming’s $G^\rho$ family of tests (1982). The likelihood function is maximized at $\hat{m} = 50,359.12$, which suggests that Weibull distribution is a better fit to the data than the loglogistic distribution. Therefore, the logrank test ($z = 1.59, p = 0.112$) is chosen over the Wilcoxon test ($z = 1.18, p = 0.239$).

6. Conclusion

The simulation results have shown that the adaptive two-sample test is more robust to the underlying survival distribution, relative to the logrank and Wilcoxon tests. Under various distribution and censoring configurations, it performed better than the less efficient of the logrank and Wilcoxon tests, while comparing favorably with the better one. It seems to do this without substantial inflation of Type I error. We believe the discriminant diagnostic based on $\hat{m}$ can be a useful pretest for choosing between logrank and Wilcoxon tests, the same way a test for common variance is used to discriminate between the Welch two-sample $t$-test and the pooled-variance $t$-test.

The adaptive test can be extended to the $K$-sample problem by modifying the procedure’s step which standardizes the failure times from the two samples. Such $K$-sample adaptive test can then be further developed into a stratified test to account for covariates.

This paper only investigated the performance of the adaptive test on three well-known survival distributions: Weibull, loglogistic and lognormal. Their performance on other survival distributions is being evaluated in subsequent research.

Appendix. R code for adaptive test

See http://www.stat.wmich.edu/naranjo/adaptiveRcode/.

## Log likelihood function to be maximized

```r
fr <- function(theta,y){
  mean<-theta[1]
  sigma<-theta[2]
  m<-theta[3]
  v <- (log(y)-mean)/sigma
  sampsize <- length(y)
  logl<--sampsize*(log(m))+sum(v)-(1+m)*sum(log(1+(1/m)*exp(v)))-sampsize*log(sigma)-sampsize*log(beta(1,m))
  return(logl)
}
```

## Gradient function

```r
gr <- function(theta,y){
  mean<-theta[1]
  sigma<-theta[2]
  m<-theta[3]
  sampsize <- length(y)
  v <- (log(y)-mean)/sigma
  s <-log(m)+v
  gradient<--numeric(length(theta))
  gradient[1]<--(sampsize/sigma) + (((1+m)/sigma)*sum(exp(s)/(1+exp(s))))
  gradient[2]<--(1/sigma)*sum(v) + (((1+m)/sigma)*sum(exp(s)*v/(1+exp(s))))-
  (sampsize/sigma)
  gradient[3]<--(sampsize/m)-sampsize*(digamma(m)-digamma(1+m))+
  (((1+m)/m)*sum(exp(s)/(1+exp(s)))-sum(log(1+exp(s))))
  return(gradient)
}
```
## Function that computes for the MLE of m, mean and sigma

```r
def getmle <- function(time)
{
xbar <- mean(log(time))
sd <- sd(log(time))
maxiter <- 20000
maxxbar <- xbar+maxiter
maxsd <- sd+maxiter

for (initialmean in xbar:maxxbar)
{
  for (initialsigma in sd:maxsd)
  {
    for (initialm in 1e-10:maxiter)
    {
      mle<-try(optim(c(initialmean,initialsigma,initialm),
                     method="L-BFGS-B", fn=fr, gr=gr,
                     control=list(fnscale=-1, trace=1),
                     lower=c(-Inf,0,1e-10), upper=rep(Inf,3),
                     hessian=TRUE, y=time))

      if (class(mle) == "try-error") {next}
      else if (mle$convergence != 0) {next}
      else {break}
    }
  }
  if (class(mle) != "try-error") {
    if (mle$convergence != 0) {next}
    else {break}
  }
}

list(meanhat=mle$par[1], sigmahat=mle$par[2], mhat=mle$par[3],
estlogL = mle$value, convergence=mle$convergence)
}
```

## Function for the adaptive test

```r
def adaptivetest <- function(data,m)
{
LR <- survdiff(Surv(time,status==1)~group, data=data, rho=0)
WIL <- survdiff(Surv(time,status==1)~group, data=data, rho=1)
LRztest <- sqrt(LR$chisq)
Wilztest <- sqrt(WIL$chisq)
if (1/m > 0.5) newstat = Wilztest else newstat = LRztest
list(r=1/m, newstat=newstat, LRztest=LRztest, Wilztest=Wilztest)
}
```

### CALLING PROGRAM

Let "combdata" be a data frame which consists of at least three vectors: "time" (survival/failure time), "status" (0=censored, 1=uncensored) and "group" (1=Group 1, 2=Group 2).

```r
uncensoredT1 <- combdata$time[combdata$status==1 & combdata$group==1]
uncensoredT2 <- combdata$time[combdata$status==1 & combdata$group==2]
uncensoredT1_rescaled <- uncensoredT1*(sd(uncensoredT2)/sd(uncensoredT1))
mle <- getmle(time=c(uncensoredT1_rescaled, uncensoredT2))
output <- adaptivetest(data=combdata, m=mle$mhat)
```
References


