cpsurvsim: An R Package for Simulating Data from Change-Point Hazard Distributions

by Camille J. Hochheimer and Roy T. Sabo

Abstract Change-point hazard models have several practical applications, including modeling processes such as cancer mortality rates and disease progression. While the inverse cumulative distribution function (CDF) method is commonly used for simulating data, we demonstrate the shortcomings of this approach when simulating data from change-point hazard distributions with more than a scale parameter. We propose an alternative method of simulating this data that takes advantage of the memoryless property of survival data and introduce the R package **cpsurvsim** which implements both simulation methods. The functions of **cpsurvsim** are discussed, demonstrated, and compared.

1 Introduction: Simulating from time-to-event processes in R

When modeling time-to-event processes, especially over long periods of time, it is often unreasonable to assume a constant hazard rate. In these cases, change-point hazard models are applicable. The majority of research surrounding change-point hazard models focuses on the Cox proportional hazards and piecewise exponential models with one change-point (Yao, 1986; Gijbels and Gurler, 2003; Wu et al., 2003; Rojas et al., 2011; Dupuy, 2006), likely due to the straightforward extension for including fixed and time-varying covariates (Zhou, 2001; Hendry, 2014; Montez-Rath et al., 2017a; Wong et al., 2018). Research on hazard models with multiple change-points is also expanding as these models have a wide range of applications in fields such as medicine, public health, and economics (Liu et al., 2008; Goodman et al., 2011; He et al., 2013; Han et al., 2014; Qian and Zhang, 2014; Cai et al., 2017). In the interest of simulating time-to-event data featuring trends with multiple change-points, Walke (2010) presents an algorithm for simulating data from the piecewise exponential distribution with fixed type I censoring using the location of the change-points, the baseline hazard, and the relative hazard for each time interval in between change-points. As the research surrounding parametric change-point hazard models with multiple change-points continues to grow, likewise does the need to simulate data from these distributions. Simulation is also a powerful and popular tool for assessing the appropriateness of a model for one's data or conducting a power analysis.

Several R packages available from the Comprehensive R Archive Network (CRAN) provide functions for simulating time-to-event data in general, with a heavy focus on the Cox model. Some of the more popular packages are provided in Table 1, which expands on the METACRAN compilation (Allignol and Latouche, 2020). Although considerably smaller in scope, a few R packages provide functions for simulating data with change-points. **CPsurv** has functionality for simulating both nonparametric survival data and parametric survival data from the Weibull change-point distribution but requires existing data as an argument and only allows for one change-point (Krügel et al., 2017). **SimSCRPiecewise** simulates data using the piecewise exponential hazard model within the Bayesian framework, however, this method requires at least one covariate as an argument (Chapple, 2016).

Our package **cpsurvsim** allows users to simulate data from a both the exponential and Weibull hazard models with type I right censoring allowing for multiple change-points (Hochheimer, 2021). **cpsurvsim** provides two methods for simulating data, which are introduced in the following section. The first method draws on Walke (2010), using the inverse hazard function to simulate data. The second employs the memoryless simulation method, the details of which are also discussed in the next section. We then demonstrate how to simulate data using **cpsurvsim** and compare the performance of these methods through a simulation study with the motivation of enabling users to determine which method is best for their data.

Package Title	Brief Description
coxed	Simulates data for the Cox model using the flexible-hazard method and allows for the inclusion of time-varying covariates (Kropko and Harden, 2019)
CPsurv	Simulates one change-point for non-parametric survival analysis or parametric survival analysis using the Weibull distribution (Krügel et al., 2017)
cpsurvsim	Simulates data with multiple change-points from the exponential and Weibull distributions (Hochheimer, 2021)
discSurv	Simulates survival data from discrete competing risk models (Welchowski and Schmid, 2019)
gems	Simulates data from multistate models and allows for non-Markov models that account for previous events (Blaser et al., 2015)
genSurv	Gives users the option to generate data with a binary, time-dependent covariate (Araújo et al., 2015; Meira-Machado and Faria, 2014)
ipred	Provides a function for simulating survival data for tree-structured survival analysis (Peters and Hothorn, 2019)
MicSim	Performs continuous time miscrosimulations to simulate life courses (Zinn, 2018)
PermAlgo	Uses a permutational algorithm to generate time-to-event data allowing for the inclusion of several time-dependent covariates (Sylvestre et al., 2010)
prodlim	Has functions for simulating right censored non-parametric survival data with two covariates and with or without competing risks (Gerds, 2018)
simMSM	Uses inversion sampling to simulate data from multi-state models allowing for non-linear baseline hazards, time-varying covariates, and dependence on past events (Reulen, 2015)
simPH	Simulates data from Cox proportional hazards models (Gandrud, 2015)
simsurv	Simulates data from various parametric survival distributions, 2-component mixture distributions, and user-defined hazards (Brilleman, 2019)
SimSCRPiecewise	Uses Bayesian estimation to simulate data from the piecewise exponential hazard model allowing for the inclusion of covariates (Chapple, 2016)
SimulateCER	While not a formal R package, this package extends the methods found in **PermAlgo** and can be downloaded from GitHub (Montez-Rath et al., 2017b)
survsim	Allows users to simulate time-to-event, competing risks, multiple event, and recurrent event data (Moriña and Navarro, 2014)

Table 1: R packages for simulating time-to-event data

2 Simulating data from popular change-point hazard models

The piecewise exponential model with multiple change-points $(\tau_k, k = 1, \ldots, K)$ can be expressed as

$$f(t) = \begin{cases} \theta_1 \exp\{-\theta_1 t\} & 0 \le t < \tau_1 \\ \theta_2 \exp\{-\theta_1 \tau_1 - \theta_2 (t - \tau_1)\} & \tau_1 \le t < \tau_2 \\ \vdots & \\ \theta_{K+1} \exp\{-\theta_1 \tau_1 - \theta_2 (\tau_2 - \tau_1) - \ldots - \theta_{K+1}(t - \tau_K)\} & t \ge \tau_K \end{cases} \tag{1}$$

with corresponding hazard function

$$h(t) = \begin{cases} \theta_1 & 0 \le t < \tau_1 \\ \theta_2 & \tau_1 \le t < \tau_2 \\ \vdots & \vdots \\ \theta_{K+1} & t \ge \tau_K. \end{cases} \tag{2}$$

We draw on the work of Walke (2010) in that we use the inverse hazard function to simulate survival time t. Walke (2010) uses a baseline hazard and relative hazards to simulate each time interval between change-points, whereas our simulation is based on the value of the scale parameter $(\theta_i, i = 1, \ldots, K+1)$ corresponding to each interval as specified by the user. Starting with the relationship between the cumulative density function (CDF) and the cumulative hazard function $(F(t) = 1 - \exp(-H(t))$ where $H(t) = \int h(t)dt)$ and noting that $F(t) = U$ where U is a uniform random variable on (0,1), we derive $t = H^{-1}(-\log(1-U))$. Seeing as $x = -\log(1-U) \sim Exp(1)$, we can simulate random variables from the exponential distribution and plug them into the inverse hazard function to get simulated event time t. With this in mind, the inverse cumulative hazard function for the exponential change-point hazard model with four change-points is

$$H^{-1}(x) = \begin{cases} \frac{x}{\theta_1} & 0 \le x < A \\ \frac{x-A}{\theta_2} + \tau_1 & A \le x < A + B \\ \frac{x-A-B}{\theta_3} + \tau_2 & A + B \le x < A + B + C \\ \frac{x-A-B-C}{\theta_4} + \tau_3 & A + B + C \le x < A + B + C + D \\ \frac{x-A-B-C-D}{\theta_5} + \tau_4 & x \ge A + B + C + D \end{cases} \tag{3}$$

where $A = \theta_1 \tau_1$, $B = \theta_2(\tau_2 - \tau_1)$, $C = \theta_3(\tau_3 - \tau_2)$, and $D = \theta_4(\tau_4 - \tau_3)$. In **cpsurvsim**, this method of simulating time-to-event data is considered the CDF method. An end-of-study time horizon (or maximum measurement time) is specified by the user and all simulated event times with values greater than the end time are censored at that point (type I right censoring).

The Weibull distribution is another popular parametric model for survival data due to its flexibility to fit a variety of hazard shapes while still satisfying the proportional hazards assumption. Note that when $\gamma = 1$, it is identical to the exponential distribution. The Weibull change-point model has the probability density function

$$f(t) = \begin{cases} \theta_1 t^{\gamma-1} \exp\{-\frac{\theta_1}{\gamma} t^\gamma\} & 0 \le t < \tau_1 \\ \theta_2 t^{\gamma-1} \exp\{-\frac{\theta_2}{\gamma}(t^\gamma - \tau_1^\gamma) - \frac{\theta_1}{\gamma}\tau_1^\gamma\} & \tau_1 \le t < \tau_2 \\ \vdots & \\ \theta_{K+1} t^{\gamma-1} \exp\{-\frac{\theta_{K+1}}{\gamma}(t^\gamma - \tau_K^\gamma) - \frac{\theta_K}{\gamma}(\tau_K^\gamma - \tau_{K-1}^\gamma) - \ldots - \frac{\theta_1}{\gamma}\tau_1^\gamma\} & t \ge \tau_K \end{cases} \tag{4}$$

with corresponding hazard function

$$h(t) = \begin{cases} \theta_1 t^{\gamma-1} & 0 \le t < \tau_1 \\ \theta_2 t^{\gamma-1} & \tau_1 \le t < \tau_2 \\ \vdots & \vdots \\ \theta_{K+1} t^{\gamma-1} & t \ge \tau_K. \end{cases} \tag{5}$$

As with the exponential model, event times can be simulated using the inverse hazard function (shown

Function	Hazard model	Simulation method
`exp_cdfsim`	Piecewise constant	Inverse hazard function
`exp_memsim`	Piecewise constant	Memoryless
`weib_cdfsim`	Weibull change-point	Inverse hazard function
`weib_memsim`	Weibull change-point	Memoryless

Table 2: Summary of functions for simulating data using **cpsurvsim**

here with four change-points)

$$H^{-1}(x) = \begin{cases} (\frac{\gamma}{\theta_1}x)^{1/\gamma} & 0 \leq x < A \\ [\frac{\gamma}{\theta_2}(x-A) + \tau_1^\gamma]^{1/\gamma} & A \leq x < A+B \\ [\frac{\gamma}{\theta_3}(x-A-B) + \tau_2^\gamma]^{1/\gamma} & A+B \leq x < A+B+C \\ [\frac{\gamma}{\theta_4}(x-A-B-C) + \tau_3^\gamma]^{1/\gamma} & A+B+C \leq x < A+B+C+D \\ [\frac{\gamma}{\theta_5}(x-A-B-C-D) + \tau_4^\gamma]^{1/\gamma} & x \geq A+B+C+D \end{cases} \quad (6)$$

where $A = \frac{\theta_1}{\gamma}\tau_1^\gamma$, $B = \frac{\theta_2}{\gamma}(\tau_2^\gamma - \tau_1^\gamma)$, $C = \frac{\theta_3}{\gamma}(\tau_3^\gamma - \tau_2^\gamma)$, and $D = \frac{\theta_4}{\gamma}(\tau_4^\gamma - \tau_3^\gamma)$.

Zhou (2001) touches on the idea of the memoryless property as a means of interpreting the piecewise exponential model, however, we take this one step further by using this property to simulate data from change-point hazard models. In survival analysis, the memoryless property states that the probability of an individual experiencing an event at time t is independent of the probability of experiencing an event up to that point. Likewise, the probability of an event occurring after a change-point is independent of the probability that the event occurs before the change-point.

Our memoryless simulation method uses this extension of the memoryless property in that data between change-points are simulated from independent exponential or Weibull hazard distributions with scale parameters θ_i corresponding to each time interval. Participants with simulated survival times past the next change-point are considered surviving at least to that change-point and then an additional survival time is simulated for them in the next time interval. Total time to event is calculated as the sum of time in each interval between change-points, with those surviving past the study end time censored at that point. Survival times within each interval are calculated using the inverse hazard of the independent exponential or Weibull function representing that time period. In this way, the inverse hazard and memoryless methods are equivalent when there are no change-points.

3 The cpsurvsim package

The **cpsurvsim** package can be installed from CRAN. Functions for simulating data are summarized in Table 2

As an example of the functions `exp_cdfsim` and `weib_cdfsim`, which simulate data using the inverse hazard method from the exponential and Weibull distributions, respectively, consider the following:

```
library(cpsurvsim)
dta1 <- exp_cdfsim(n = 50, endtime = 100, theta = c(0.005, 0.01, 0.05),
+  tau = c(33, 66))
head(dta1)

       time censor
1 100.00000      0
2  85.99736      1
3  78.21772      1
4  71.03138      1
5 100.00000      0
6  82.71520      1

dta2 <- weib_cdfsim(n = 50, endtime = 100, gamma = 2,
+  theta = c(0.0001, 0.0002, 0.0001), tau = c(33, 66))
head(dta2)
```

```
       time censor
1  11.36844      1
2 100.00000      0
3  81.04904      1
4 100.00000      0
5  71.93590      1
6  56.40275      1
```

When simulating using the memoryless method, we use the following calls from **cpsurvsim**:

```
dta3 <- exp_memsim(n = 50, endtime = 100, theta = c(0.005, 0.01, 0.05),
+  tau = c(33, 66))
head(dta3)
```

```
      time censor
1 93.64262      1
2 63.47413      1
3 84.54253      1
4 89.01574      1
5 73.92685      1
6 23.67631      1
```

```
dta4 <- weib_memsim(n = 50, endtime = 100, gamma = 2,
+  theta = c(0.0001, 0.0002, 0.0001), tau = c(33, 66))
head(dta4)
```

```
       time censor
1  59.47848      1
2 100.00000      0
3  62.08739      1
4 100.00000      0
5 100.00000      0
6 100.00000      0
```

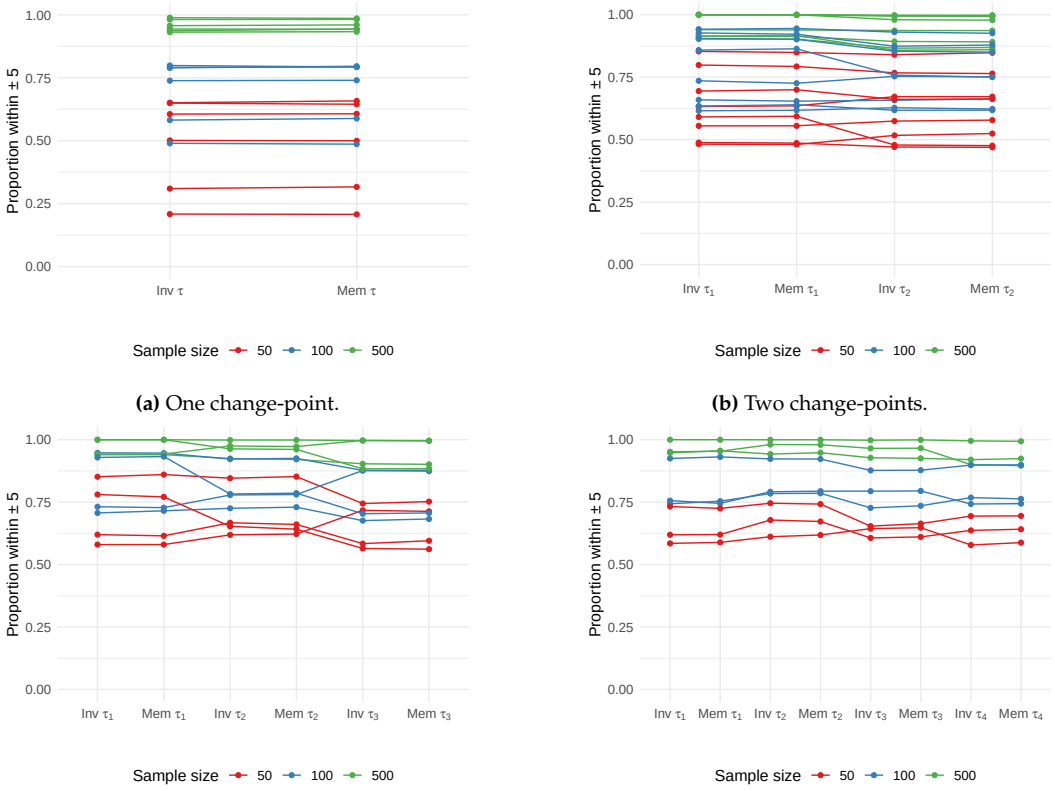
As seen in these examples, all four functions return a dataset with the survival times and a censoring indicator.

4 Comparison of simulation methods

To compare the performance of the inverse hazard method with the memoryless method under different settings, we conducted a simulation study using **cpsurvsim**. We simulated data with one, two, three, and four change-points using both the exponential and Weibull distributions. In our simulation, time t ranged from 0-100 and change-points occurred at various times within that range. Sample sizes of 50, 100, and 500 were tested and values of θ were chosen to demonstrate differences between the simulation methods when the hazard rate changes (e.g., smaller to larger hazard versus larger to smaller hazard). For the Weibull simulations, we set $\gamma = 2$. We conducted 10,000 simulations of each setting. We are primarily interested in comparing the ability of these two simulation methods to simulate data with the correct change-points τ_i. Therefore, we compared how often the estimated value ($\hat{\tau}_i$) was within a 10% range, in this case $[\tau_i - 5, \tau_i + 5]$ based on our time range. We also evaluated whether the known values of τ_i fell within the 95% confidence interval of the average simulated values for both methods as well as discuss bias in the model parameters. This simulation study was conducted in R 3.6.1.

When simulating from the exponential distribution, these two simulation methods had comparable accuracy in terms of the location of the change-points (see Figure 1). Sample size, however, had a large impact on the accuracy regardless of the simulation method. When estimating one change-point with a sample size of 50, there were a few simulation templates where less than a third of estimates $\hat{\tau}$ were within range of the known change-point (Figure 1a). In general, accuracy improved as the sample size increased. For every simulation scenario using the exponential distribution, the 95% confidence interval for the mean estimate of τ_i included the known value.

Simulations for the Weibull distribution, however, revealed important differences in accuracy between the two methods (see Figure 2). Although accuracy of the two methods was similar when simulating one change-point, there were many cases where accuracy was very low, even with a sample size of 500 (Figure 2a). In almost one quarter (4/18) of the simulation scenarios, the true value of τ was not within the 95% confidence interval of the average estimate for either method. In three of



(a) One change-point.

(b) Two change-points.

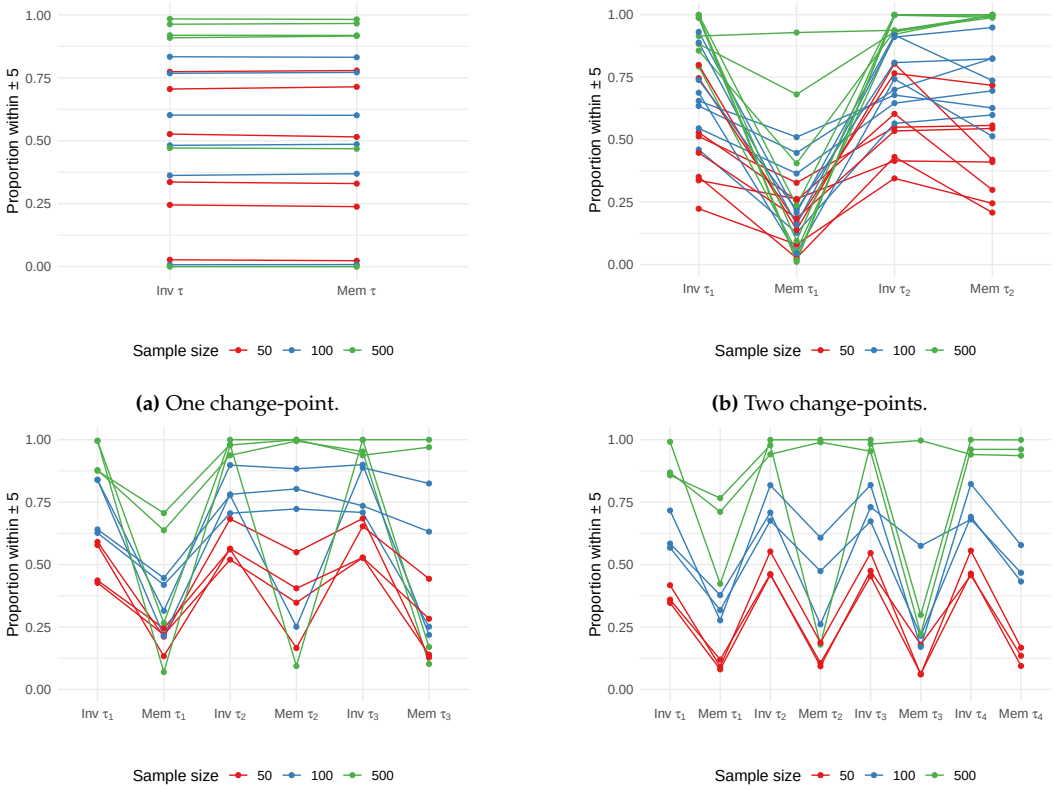
(c) Three change-points.

(d) Four change-points.

Figure 1: Accuracy of change-point simulations for the exponential distribution. The y-axis represents the proportion of change-point estimates ($\hat{\tau}$) within 10% of the known value. "Inv" refers to the inverse hazard method and "Mem" refers to the memoryless method. This figure demonstrates that accuracy is higher with a larger sample size and is similar for both simulation methods.

those scenarios, both simulation methods severely underestimated τ when the true value was 80. When simulating two change-points (Figure 2b), accuracy of the first change-point was often much lower when using the memoryless method, especially with a larger sample size. All except one of the simulation scenarios where the known value of τ_1 did not fall within the 95% confidence interval of the average estimate for the memoryless method had a sample size of 500. Accuracy was lower for all change-points in the three change-point simulations when using the memoryless method and the discrepancies between the two methods were larger for larger sample sizes (Figure 2c). In almost half of the simulation scenarios at least one value of τ_i was not within the 95% confidence interval of the mean estimate for the memoryless method and all except one of those scenarios had a sample size of 500. When simulating four change-points (Figure 2d), for most scenarios there was a large drop in accuracy at the first and third change-point for the memoryless method compared to the inverse hazard method. At the second and fourth change-points, however, there was a drop in accuracy for sample sizes of 50 and 100 whereas most simulation scenarios with a sample size of 500 had similar accuracy between the two methods. Every simulation scenario for four change-points with a sample size of 500 using the memoryless method had at least one change-point where the 95% confidence interval did not include the known value of τ_i. The known value of τ_4 was, however, included in the 95% confidence interval for every scenario with a sample size of 500.

We suspected that the inaccurate estimates using the Weibull distribution were due to inaccuracies in estimating the shape parameter γ, which is assumed constant across all time intervals. Indeed, γ was often under-estimated as seen in Figure 3. Values of γ, however, were similar for both methods except when there were three change-points (Figure 3c), in which case the estimates of γ using the inverse hazard method were closer to the known value of two. We were unable to estimate γ for the simulations using the memoryless method when there was a sample size of 50 (Figure 3d).



(a) One change-point.

(b) Two change-points.

(c) Three change-points.

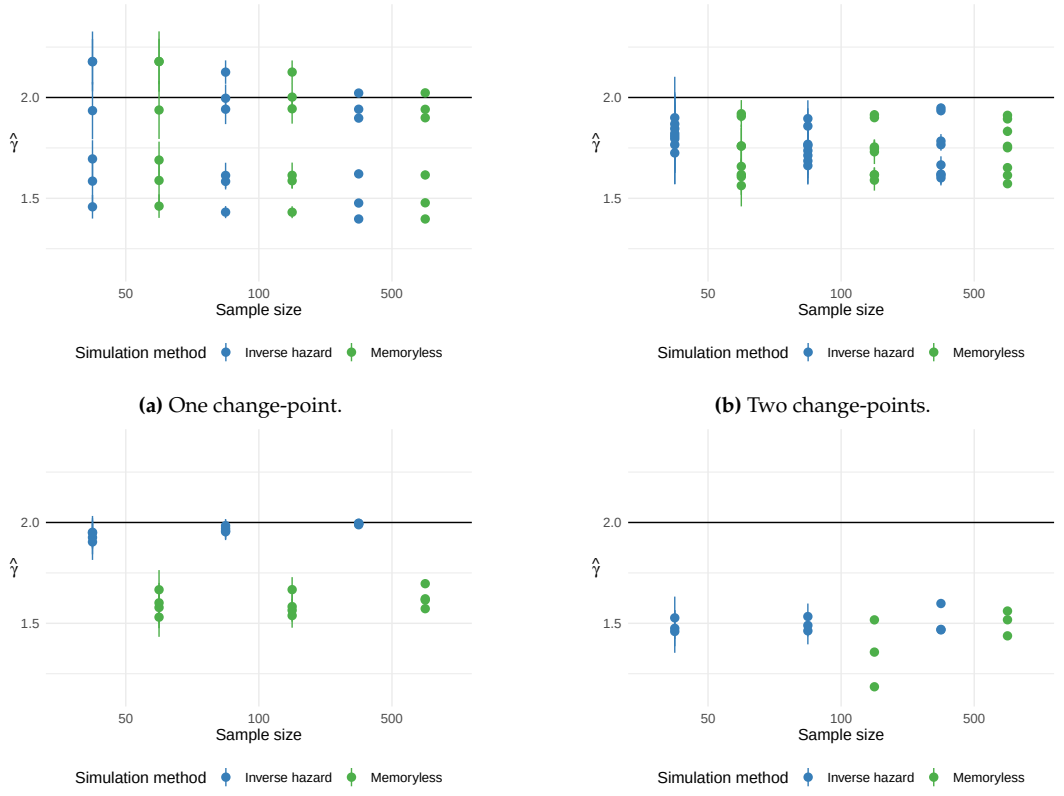
(d) Four change-points.

Figure 2: Accuracy of change-point simulations for the Weibull distribution. The y-axis represents the proportion of change-point estimates ($\hat{\tau}$) within 10% of the known value. "Inv" refers to the inverse hazard method and "Mem" refers to the memoryless method. This figure demonstrates that accuracy is generally higher with larger sample sizes and when using the inverse hazard method.

5 Summary

The R package **cpsurvsim** provides implementation of the standard method of simulating from a distribution, using the inverse CDF, and a new method that exploits the memoryless property of survival analysis. When simulating from the exponential distribution with multiple change-points, these methods have comparable performance. Simulating multiple change-points from the Weibull hazard, however, suggested that the inverse hazard method produces more accurate estimates of the change-points τ_i. The accuracy of the exponential simulations suffered when the sample size was less than 500 whereas in some cases, simulations of the Weibull distribution had worse accuracy with a sample size of 500. In practice, change-point hazard models are often applied to data from large longitudinal cohort studies where the sample size is very large (e.g., Goodman et al. (2011) and Williams and Kim (2013)). These results suggest that larger sample sizes are preferred when using an exponential model but to use caution even with a large sample when using the Weibull model. We hope that having an R package for simulating data from multiple change-point hazard distributions will aid in the development of extensions and alternatives to our research on tests for multiple change-points (Hochheimer and Sabo, 2021).

The inspiration to develop the memoryless simulation method and test it came from observing the shortcomings of the inverse hazard method in our research. The memoryless method performs better in some simulation scenarios, which led us to implement both methods in this R package. This simulation study, however, suggests that in the majority of cases the inverse hazard method simulates values of τ_i more accurately. Our simulation study also highlighted accuracy issues with both methods when simulating data from sample sizes of 50 or 100, which we suspect are due to using a relatively small amount of data to estimate several model parameters. One should consider exploring other methods to simulate a multiple change-point distribution with a small sample size. The acceptance-rejection method, for example, may produce more accurate parameter estimates at the cost of more computational time needed to reach the desired sample size, a cost that might be worthwhile if the sample size is smaller to begin with (Rizzo, 2007). Alternatively, one might run a simulation study to determine which of these two methods is best suited for their specific parameters.



(a) One change-point.

(b) Two change-points.

(c) Three change-points.

(d) Four change-points.

Figure 3: Estimated values of shape parameter γ for the Weibull distribution. Dots indicate the average estimated value for each simulation scenario with the vertical lines representing the 95% CI. The solid horizontal line represents the known value of γ. This plot demonstrates inaccurate estimates of γ for both simulation methods except when there are three change-points, in which case the estimates are more accurate using the inverse hazard method.

An important limitation of the `exp_cdfsim` and `weib_cdfsim` functions is that they only accommodate up to four change-points. While it's possible to have more than this many change-points in a dataset, it's also important to make sure that there is a meaningful interpretation for multiple change-points. Also, **cpsurvsim** only accommodates type I right censoring. For the Weibull distribution, γ is assumed fixed for every interval between change-points. In our simulation study, we only estimated an overall value of γ due to convergence issues when trying to estimate it within each interval between change-points. In an effort to be concise, the accuracy of the scale parameters θ_i are not discussed here, however, in some cases this parameter may be of more interest than the change-point τ. Thus, we briefly discuss these results in the appendix. Future versions of **cpsurvsim** could incorporate additional features such as accommodating informative censoring.

6 Acknowledgments

Thank you to Dr. Sarah Ratcliffe for her guidance and for assisting with running these simulations.

Bibliography

A. Allignol and A. Latouche. *Task View: Survival Analysis*, 2020. URL https://www.r-pkg.org/ctv/Survival. METACRAN. [p201]

A. Araújo, L. Meira-Machado, and S. Faria. *genSurv: Generating Multi-State Survival Data*, 2015. URL http://CRAN.R-project.org/package=genSurv. R package version 1.0.3. [p202]

N. Blaser, L. Salazar Vizcaya, J. Estill, C. Zahnd, B. Kalesan, M. Egger, O. Keiser, and T. Gsponer. gems: An r package for simulating from disease progression models. *Journal of Statistical Software*, 64(10): 1–22, 2015. doi: 10.18637/jss.v064.i10. [p202]

S. Brilleman. *simsurv: Simulate Survival Data*, 2019. URL https://CRAN.R-project.org/package=simsurv. R package version 0.2.3. [p202]

X. Cai, Y. Tian, and W. Ning. Modified information approach for detecting change points in piecewise linear failure rate function. *Statistics & Probability Letters*, 125:130–140, 2017. ISSN 0167-7152. doi: 10.1016/j.spl.2017.02.005. [p201]

A. G. Chapple. *SimSCRPiecewise: Simulates Univariate and Semi-Competing Risks Data Given Covariates and Piecewise Exponential Baseline Hazards'*, 2016. URL https://CRAN.R-project.org/package=SimSCRPiecewise. R package version 0.1.1. [p201, 202]

J.-F. Dupuy. Estimation in a change-point hazard regression model. *Statistics & Probability letters*, 76(2): 182–190, 2006. ISSN 0167-7152. doi: 10.1016/j.spl.2005.07.013. [p201]

C. Gandrud. simph: An r package for illustrating estimates from cox proportional hazard models including for interactive and nonlinear effects. *Journal of Statistical Software*, 65(3):1–20, 2015. doi: 10.18637/jss.v065.i03. [p202]

T. A. Gerds. *prodlim: Product-Limit Estimation for Censored Event History Analysis*, 2018. URL https://CRAN.R-project.org/package=prodlim. R package version 2018.04.18. [p202]

I. Gijbels and U. Gurler. Estimation of a change point in a hazard function based on censored data. *Lifetime Data Analysis*, 9(4):395–411, 2003. ISSN 1380-7870. doi: 10.1023/B:LIDA.0000012424.71723.9d. [p201]

M. S. Goodman, Y. Li, and R. C. Tiwari. Detecting multiple change points in piecewise constant hazard functions. *Journal of Applied Statistics*, 38(11):2523–2532, 2011. ISSN 0266-4763. doi: 10.1080/02664763.2011.559209. [p201, 207]

G. Han, M. J. Schell, and J. Kim. Improved survival modeling in cancer research using a reduced piecewise exponential approach. *Statistics in Medicine*, 33(1):59–73, 2014. doi: 10.1002/sim.5915. [p201]

P. He, L. Fang, and Z. Su. A sequential testing approach to detecting multiple change points in the proportional hazards model. *Statistics in Medicine*, 32(7):1239–1245, 2013. doi: 10.1002/sim.5605. [p201]

D. J. Hendry. Data generation for the cox proportional hazards model with time-dependent covariates: A method for medical researchers. *Statistics in Medicine*, 33(3):436–454, 2014. doi: 10.1002/sim.5945. [p201]

C. Hochheimer. *cpsurvsim: Simulating Survival Data from Change-Point Hazard Distributions*, 2021. URL http://github.com/camillejo/cpsurvsim. [p201, 202]

C. J. Hochheimer and R. T. Sabo. Testing for phases of dropout attrition using change-point hazard models. *Journal of Survey Statistics and Methodology*, 09 2021. ISSN 2325-0984. doi: 10.1093/jssam/smab030. URL https://doi.org/10.1093/jssam/smab030. [p207]

J. Kropko and J. J. Harden. *coxed: Duration-Based Quantities of Interest for the Cox Proportional Hazards Model*, 2019. URL https://CRAN.R-project.org/package=coxed. R package version 0.2.4. [p202]

S. Krügel, A. R. Brazzale, and H. Kuechenhoff. *CPsurv: Nonparametric Change Point Estimation for Survival Data*, 2017. URL https://CRAN.R-project.org/package=CPsurv. R package version 1.0.0. [p201, 202]

M. Liu, W. Lu, and Y. Shao. A monte carlo approach for change-point detection in the cox proportional hazards model. *Statistics in medicine*, 27(19):3894–3909, 2008. ISSN 1097-0258. doi: 10.1002/sim.3214. [p201]

L. Meira-Machado and S. Faria. A simulation study comparing modeling approaches in an illness-death multi-state model. *Communications in Statistics - Simulation and Computation*, 43(5):929–946, 2014. ISSN 1532-4141. doi: 10.1080/03610918.2012.718841. [p202]

M. E. Montez-Rath, K. Kapphahn, M. B. Mathur, A. A. Mitani, D. J. Hendry, and M. Desai. Guidelines for generating right-censored outcomes from a cox model extended to accommodate time-varying covariates. *Journal of Modern Applied Statistical Methods*, 16(1):6, 2017a. doi: 10.22237/jmasm/1493597100. [p201]

M. E. Montez-Rath, K. Kapphahn, M. B. Mathur, N. Purington, V. R. Joyce, and M. Desai. Simulating realistically complex comparative effectiveness studies with time-varying covariates and right-censored outcomes. *arXiv*, 09 2017b. doi: https://arxiv.org/abs/1709.10074. [p202]

D. Moriña and A. Navarro. The r package survsim for the simulation of simple and complex survival data. *Journal of Statistical Software*, 59(2):1–20, 2014. doi: 10.18637/jss.v059.i02. [p202]

A. Peters and T. Hothorn. *ipred: Improved Predictors*, 2019. URL https://CRAN.R-project.org/package=ipred. R package version 0.9-9. [p202]

L. Qian and W. Zhang. *Multiple Change-Point Detection in Piecewise Exponential Hazard Regression Models with Long-Term Survivors and Right Censoring*, book section 18. Springer International Publishing, Switzerland, 2014. doi: 10.1007/978-3-319-02651-0_18. [p201]

H. Reulen. *simMSM: Simulation of Event Histories for Multi-State Models*, 2015. URL https://CRAN.R-project.org/package=simMSM. R package version 1.1.41. [p202]

M. L. Rizzo. *Statistical Computing with R*. CRC Press, 2007. ISBN 9781498786591. [p207]

O. Rojas, F. Bulmaro, and J. Hernández. Modelo de riesgo con un punto de cambio y covariables dependientes del tiempo. *Revista Investigación Operacional*, 32:114–122, 2011. [p201]

M.-P. Sylvestre, T. Evans, T. MacKenzie, and M. Abrahamowicz. *PermAlgo: Permutational Algorithm to Generate Event Times Conditional on a Covariate Matrix Including Time-Dependent Covariates*, 2010. URL https://cran.r-project.org/package=PermAlgo. R package version 1.1. [p202]

R. Walke. Example for a piecewise constant hazard data simulation in r. Report, Max Planck Institute for Demographic Research, 2010. URL https://www.demogr.mpg.de/papers/technicalreports/tr-2010-003.pdf. [p201, 203]

T. Welchowski and M. Schmid. *discSurv: Discrete Time Survival Analysis*, 2019. URL https://CRAN.R-project.org/package=discSurv. R package version 1.4.0. [p202]

M. R. Williams and D. Y. Kim. A test for an abrupt change in weibull hazard functions with staggered entry and type i censoring. *Communications in Statistics - Theory and Methods*, 42(11):1922–1933, 2013. ISSN 0361-0926. doi: 10.1080/03610926.2011.600505. [p207]

G. Y. C. Wong, Q. Diao, and Q. Yu. Piecewise proportional hazards models with interval-censored data. *Journal of Statistical Computation and Simulation*, 88(1):140–155, 2018. doi: 10.1080/00949655.2017.1380645. [p201]

C. Q. Wu, L. C. Zhao, and Y. H. Wu. Estimation in change-point hazard function models. *Statistics & Probability Letters*, 63(1):41–48, 2003. ISSN 0167-7152. doi: 10.1016/S0167-7152(03)00047-6. [p201]

Y. C. Yao. Maximum-likelihood-estimation in hazard rate models with a change-point. *Communications in Statistics - Theory and Methods*, 15(8):2455–2466, 1986. ISSN 0361-0926. doi: 10.1080/03610928608829261. [p201]

M. Zhou. Understanding the cox regression models with time-change covariates. *The American Statistician*, 55(2):153–155, 2001. doi: 10.1198/000313001750358491. [p201, 204]

S. Zinn. *MicSim: Performing Continuous-Time Microsimulation*, 2018. URL https://CRAN.R-project.org/package=MicSim. R package version 1.0.13. [p202]

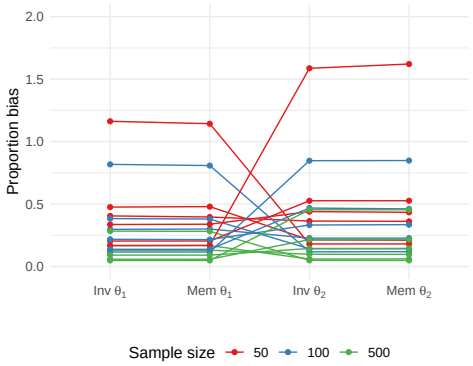
Camille J. Hochheimer, PhD
Department of Biostatistics and Informatics
Colorado School of Public Health
13001 East 17th Place, 4th Floor West, Aurora, CO 80045
USA
ORCiD: 0000-0002-0984-0909
camille.hochheimer@cuanschutz.edu

Roy T. Sabo, PhD
Department of Biostatistics
Virginia Commonwealth University
PO Box 980032, Richmond, VA 23298-0032
USA
ORCiD: 0000-0001-9159-4876
roy.sabo@vcuhealth.org

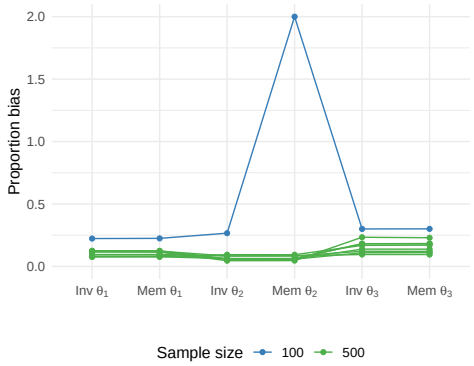
1 Appendix: Analysis of scale parameters

While the change-points can be estimated without knowing the values of the scale parameters, the reverse is not possible. Thus, we used the estimated values of the change-points in order to estimate values of θ_i. As the number of change-points increased, so did the difficulty in estimating values of θ_i, especially with a smaller sample size.

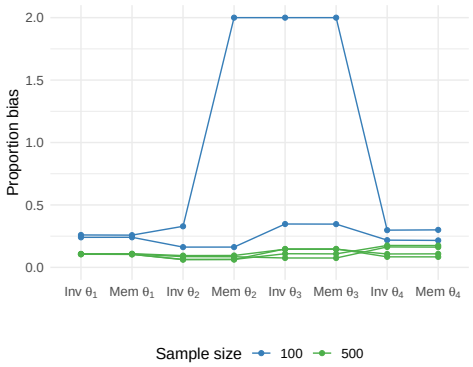
With a few exceptions, the estimates of θ_i for the exponential distribution were similar between both methods (Figure 4). These exceptions were θ_2 in the two change-point (Figure 4b) and three change-point models (Figure 4c), where the memoryless method with a sample size of 100 had a much larger proportion of bias. We were only able to estimate the shape parameters for the four change-point model when the sample size was 500.



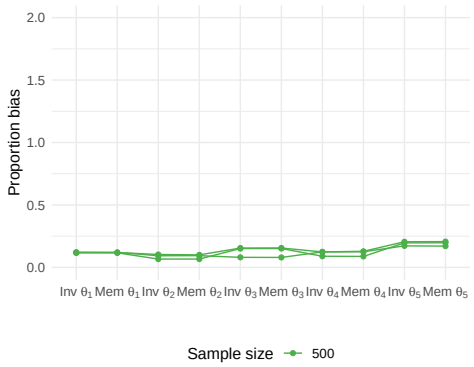
(a) One change-point.



(b) Two change-points.



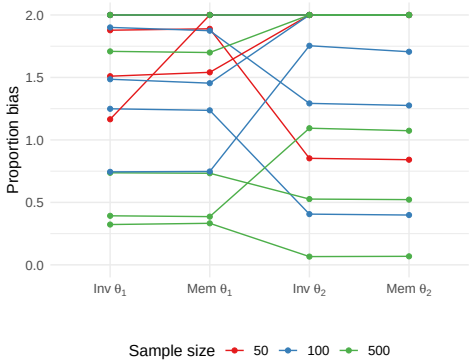
(c) Three change-points.



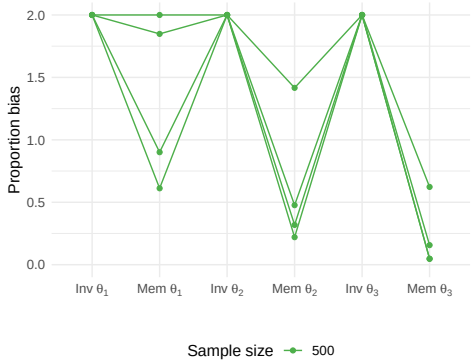
(d) Four change-points.

Figure 4: Accuracy of scale parameter $\hat{\theta}_i$ for the exponential distribution. The y-axis represents the average proportion of bias of $\hat{\theta}_i$ relative to the known value of θ_i. A proportion of bias of 2 represents estimates with at least 200% bias. "Inv" refers to the inverse hazard method and "Mem" refers to the memoryless method. This figure demonstrates that bias was generally similar between simulation methods with a few exceptions where bias was larger using the memoryless method.

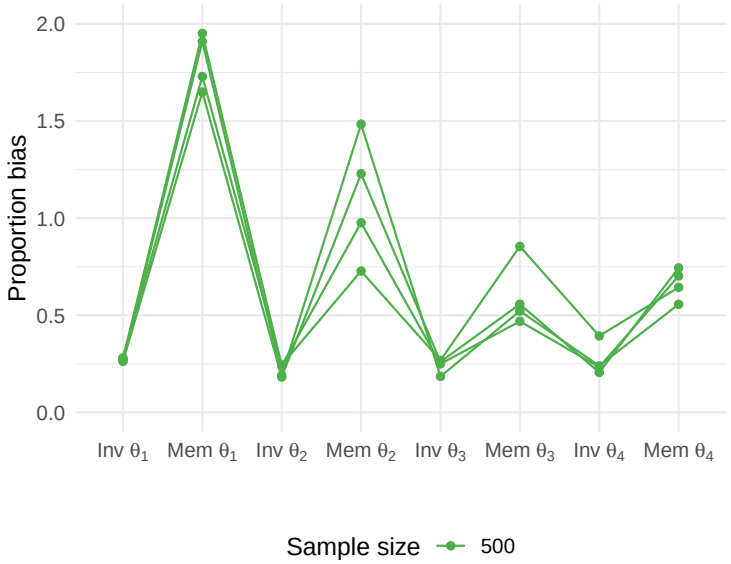
Estimates of θ_i for the one change-point Weibull model were similar across simulation methods but bias was high even when the sample size was large (Figure 5a). Bias was generally smaller when using the memoryless method to estimate θ_i in the two change-point Weibull model (Figure 5b). On the other hand, bias was larger when using the memoryless method to estimate the shape parameter for the three change-point Weibull model (Figure 5c). We were unable to estimate θ using the results from the memoryless method for any of the four change-point simulations.



(a) One change-point.



(b) Two change-points.



(c) Three change-points.

Figure 5: Accuracy of scale parameter $\hat{\theta}_i$ for the Weibull distribution. The y-axis represents the average proportion of bias of $\hat{\theta}_i$ relative to the known value of θ_i. A proportion of bias of 2 represents estimates with at least 200% bias. "Inv" refers to the inverse hazard method and "Mem" refers to the memoryless method. This figure demonstrates similar bias between methods when there is one change-point, smaller bias using the memoryless method when there are two change-points, and smaller bias using the inverse hazard method when there are three change-points.