

Extending MKSFitter to Right-Censored Data

Jerzy Wieczorek* and Jong Sung Kim**

Department of Mathematics and Statistics, Portland State University, Portland, OR 97201, U.S.A.

**email*: jerzy@pdx.edu

***email*: jong@pdx.edu

SUMMARY: The MKSFitter algorithm of Weber, Leemis, and Kincaid (2006) computes parameter estimates for several different continuous univariate distributions, using an evolutionary optimization algorithm, and recommends the distribution and parameter estimates that best minimize the Kolmogorov-Smirnov test statistic. We modify this tool by extending it to use the Kaplan-Meier estimate of the cdf for right-censored data. Using simulated data from the most commonly-used distributions in survival analysis, we evaluated the algorithm's ability to select the correct distribution type (at various sample sizes and censoring rates). We also compared this tool's estimates with the right-censored MLE and found the two estimation techniques to have comparable accuracy.

KEY WORDS: Evolutionary algorithm; Goodness-of-fit tests; Right-censored data.

1. Introduction

A common problem in analyzing biological or other survival data is to model the distribution function of a population. It is hoped that the dataset is representative of the population being studied and one attempts to find a model that shows a good fit to this data. Usually one must assume a certain distribution type and then estimate the model parameters that best fit the available data. In some cases prior knowledge of the system strongly suggests that a particular distribution is more appropriate than any other. However, when several models are possible, it can be tedious to fit and compare them all, and thus it is possible to overlook distributions that fit the data better than whatever distribution is chosen.

Furthermore, there are additional difficulties with censored data, which arises commonly in survival analyses. Standard estimators such as MLEs are often challenging to compute analytically and often one is forced to resort to numerical optimization.

In light of some of these challenges, Weber, Leemis, and Kincaid (2006) propose, implement, and test a software tool called MKSFitter that is meant to address this problem of distribution selection and parameter estimation in the case of uncensored data from a univariate continuous distribution. It implements minimum Kolmogorov-Smirnov estimation (MKSE), in which the parameter estimates for a given distribution are chosen so as to minimize the one-sample Kolmogorov-Smirnov test statistic. Gyorfi, Vajda, and Van Der Meulen (1996a, 1996b) explore the idea of MKSE from a theoretical standpoint (and show that parameter estimates are consistent for most common distribution models), but MKSFitter appears to be the first published software tool to implement this approach.

The estimates are found by an evolutionary optimization algorithm, presented in Sobieszczanski-Sobieski, Laba, and Kincaid (1999), whose objective function is minimization of the K-S statistic. An overview of the algorithm, called bell-curve based (BCB) evolutionary optimization, is presented briefly in Weber et al. (2006) and further details of implementation

and performance are explained in Sobieszczanski-Sobieski et al. (1999) and Kincaid, Weber, and Sobieszczanski-Sobieski (2000).

The results of simulation studies by Weber et al. (2006) show that MKSFitter selects the correct distribution fairly well for large sample sizes: depending on the true distribution, the correct selection was made 40 to 70 percent of the time for samples of size $n = 100$. They also run an experiment which suggests that MKSE performance is comparable to maximum likelihood estimation (MLE), in terms of the Euclidean distance (in the parameter space) of the respective parameter estimates from their true values.

However, MKSFitter is only designed to work on uncensored data. Weber et al. (2006) suggest that users can extend this tool to right-censored data by modifying the software to minimize the vertical distance between the fitted CDF and the Kaplan-Meier estimate of the empirical CDF. (The Kaplan-Meier product-limit estimator is a commonly-used empirical estimate of the CDF for right-censored data. In the case where no observations are censored, it is equivalent to the usual empirical CDF.) Since this suggestion has not been followed up with any test results, it seemed to us to be worthwhile to carry out the idea and report its performance.

In other words, if $\mathbf{x} = (x_1, x_2, \dots, x_n)$ are iid from a population with unknown parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$, the original MKSFitter finds estimates $\hat{\boldsymbol{\theta}}$ that minimize the Kolmogorov-Smirnov test statistic

$$D_n = \sup_x |\hat{F}(x|\hat{\boldsymbol{\theta}}) - F_n(x)| \quad (1)$$

where n is sample size, $\hat{F}(x|\hat{\boldsymbol{\theta}})$ is the fitted CDF based on the given set of parameter estimates, and $F_n(x)$ is the empirical CDF based directly on the data. For uncensored data, $F_n(x)$ is a step function increasing by $1/n$ at every observation. For a right-censored dataset, however, we use the Kaplan-Meier product-limit estimator as $F_n(x)$ and otherwise continue exactly as in Weber et al. (2006):

For each proposed set of parameter estimates tested by MKSFitter, it calculates the ‘Kolmogorov-Smirnov’ statistic to evaluate the fitness of that set of parameter estimates. (Technically this is no longer exactly the Kolmogorov-Smirnov test, since the empirical CDF estimate is different. Nonetheless, the same kind of test statistic is calculated the usual way, by finding the absolute vertical difference between the theoretical CDF value at each step point and the empirical CDF values just before and just after that step). Finally the BCB algorithm generates new ‘children’ parameter estimates near the best-fitting ‘parents,’ continuing for a pre-specified number of generations (and hopefully finding the globally-optimal estimates in that time).

2. Implementation

The original MKSFitter source code (in the C language) is available for download on the web at www.math.wm.edu/~leemis. In order to allow the use of uncensored data, we modified the C source code to accept both the observations and their corresponding empirical CDF values as input. Thus, a more user-friendly statistical software package such as R can compute the Kaplan-Meier estimator (or even some other kind of empirical CDF estimate, with other kinds of censoring) and that estimated CDF can be passed in to MKSFitter. For ease of use, and to make efficient use of code already existing in R for computing the Kaplan-Meier estimator, we created an R function that runs MKSFitter and saves the results. In this way we performed several tests of the censored MKSFitter’s performance.

In our tests, we used R to generate censored random samples from several distributions commonly used in survival analysis. The four models we used were Exponential(1), Weibull(1, 1.5), Weibull(1, 0.5), and LogNormal(2, 0.5). Their failure rates are constant, increasing (IFR), decreasing (DFR), and upside-down bathtub (UBT), respectively, which correspond to most of the common hazard function shapes. The distributions are parametrized as in Leemis (1995).

For each test, one of these distributions was used to generate 100 random samples (of size $n = 100$ each) to represent the true (uncensored) survival times, and the same distribution with slightly different parameters was used to generate censoring times. The observations for which the survival time was lower than its censoring time were uncensored, so the true survival time was placed in the dataset; the other observations were treated as censored and only their censoring times entered the dataset. In this way, we formed datasets of $n = 100$ observations each but with variable censoring rates. For instance, if the survival times were generated from $\text{Exp}(1)$ and the censoring times were generated from $\text{Exp}(3)$, then about 25% of the observations were censored on average. For each distribution type, we found censoring distribution parameters that caused 25%, 50%, and 75% mean censoring rates. Thus, we were able to test the influence of censoring rate on MKSFitter performance.

Due to space constraints it is not possible to explain the BCB optimization algorithm in detail here, but we must note that the algorithm does have settings that can be changed by the user. In particular, the number of randomly-generated starting points and the number of generations can be set easily. We had to change the default values (from 50 to 200 for initial population size, and from 100 also to 200 for number of generations) in order to avoid some very bad fits. In practice, users evaluating a single dataset at a time can use the lower values for speedier processing, since they will notice any poor fits (marked by KS statistic values near 0.5, usually meaning the fitted CDF is almost a horizontal line with value 0.5 over the range of the data) and can tweak these parameters. However, when generating and analyzing a hundred datasets in a row, as we did, it is impractical to stop after each one and re-tweak it, so we had to use ‘safer’ settings that unfortunately also increased the run-time significantly (to approximately 30 minutes for 100 datasets). Also worth noting is that the BCB algorithm seems to generate the same random starting values each time, since when we would re-analyze a given dataset that had obviously-bad results, it kept reproducing the

same exact results until we increased the initial population size (at which point the fit would improve dramatically).

3. Results

3.1 *Test 1: Distribution identification*

In our first set of tests, presented in tables 1 through 4, we evaluated how often the MKSFitter identified the correct distribution. As these tables show, the frequency of selecting the correct distribution decreased monotonically as censoring rate increased. Naturally, more censoring means more uncertainty about the data, a worse Kaplan-Meier estimate of the CDF, and fewer uncensored points at which to fit the distribution. However, even for the lowest censoring rate, the highest correct identification rates were around 50% at most, and only for Weibull data at that. Thus it does not seem safe to rely on MKSE to select the correct distribution. However, if you have a particular distribution in mind, presumably MKSFitter could at least help you realize that this choice of distribution is unreasonable if its KS statistic was much higher than that of other distributions.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

3.2 *Test 2: Parameter estimation*

In our second set of tests, shown in tables 5 through 8, we evaluated how well the MKSE matched the correct parameters of the true generating distribution. (Censored MLEs were calculated by the ‘survreg’ function in R.) As censoring rate increases, the fits also get worse. This is to be expected, since there is less and less information about the data, but of

particular importance is the fact that MKSEs get worse more severely than the MLEs. Thus, the censored MLEs appear to be much better parameter estimates than MKSEs except at the lowest censoring rates.

Overall, both tests indicated the worst performance for LogNormal data, which was also the case for uncensored data tests presented in Weber et al. (2006).

[Table 5 about here.]

[Table 6 about here.]

[Table 7 about here.]

[Table 8 about here.]

4. Conclusions

Clearly distribution selection performance degrades at high censoring rates. This fact in itself would not condemn the use of MKSFitter for censored data, since any estimation technique is necessarily going to suffer as more and more of the data is censored. However, since MLE consistently performs at least as well as MKSE (and usually better) in terms of distance from the estimated to the true parameters, MKSE does not appear to have an advantage over the better-studied MLE approach. This is particularly clear for highly-censored data.

The censored MLE's superior performance is probably due to the fact that it incorporates at least a little information about every observation. The MKSE does that only in a roundabout way, by using the Kaplan-Meier estimated cdf which is based on all the data. Otherwise MKSFitter essentially does exactly the same kind of fitting as before, except that now it has only $n*(censoring\ rate)$ effective observations, compared to the n observations the MLE incorporates.

Furthermore, the MKSE approach does not generate any standard errors of the estimates. Using bootstrapping to add this functionality could make MKSFitter a more useful tool

and allow for some hypothesis testing or confidence interval construction for the parameter estimates. However, the distance standard deviations in tables 5 through 8 suggest that MKSEs are more variable than MLEs, again suggesting that MLEs are the better tool.

The sample size of 100 in these tests was meant to represent a large but reasonable number of observations that may be encountered in real datasets. However, it may be worthwhile to explore the asymptotic properties of MKSE by testing much larger samples, and many more of them. Perhaps testing something like 500 samples at a time, of size 300 each, would show whether MKSE can perform as well as MLE at least asymptotically for heavily-censored data.

Donoho and Liu (1988) have shown that for some true distributions arbitrarily close to your assumed model distribution, MKSE can have arbitrarily high variance. However, an estimator based on minimizing the Cramer-VonMises statistic does not have the problem of such inconsistency. Thus it is conceivable that such an estimator may perform better than the KS statistic, including on censored data. This is another potential area for future study.

In any case, one great strength of the MKSFitter software is that it prints out a KS statistic for all of the tested distribution types. This can be useful as a ‘sanity check’ to make sure there is not another distribution with a much better fit than the one proposed *a priori*. Another possible area of strength for MKSE is in simulation studies where it is most important to replicate the properties of the sample data, rather than to know the true form of the original population, and thus the goal is to best match the empirical CDF of the data.

Finally, despite all of the above concerns, MKSFitter may potentially be more useful if expanded to analyze multivariate data. It can be quite challenging to apply maximum likelihood estimation on bivariate and higher-dimension distributions in censored situations. Thus it may be easier to optimize the parameters numerically with a MKSE approach,

although it remains to be seen whether the resulting estimates would be good enough to be useful.

REFERENCES

- Weber, M., Leemis, L., and Kincaid, R. (2006). Minimum Kolmogorov-Smirnov test statistic parameter estimates. *Journal of Statistical Computation and Simulation* **76**, 195-206.
- Gyorfi, L., Vajda, I., and Van Der Meulen, E. (1996a). Minimum Kolmogorov distance estimates of parameters and parametrized distributions. *Metrika* **43**, 237-255.
- Gyorfi, L., Vajda, I., and Van Der Meulen, E. (1996b). Minimum Kolmogorov distance estimates for multivariate parametrized families. *American Journal of Mathematical and Management Sciences* **16**, 167-191.
- Sobieszczanski-Sobieski, J., Laba, K., and Kincaid, R. (1999). Bell-curve based evolutionary optimization algorithm. *Structural Optimization* **18**, 264-276.
- Kincaid, R., Weber, M., and Sobieszczanski-Sobieski, J. (2000). Performance of a bell-curve based evolutionary optimization algorithm. Proceedings of the 41st AIAA Structures, Structural Dynamics, and Materials Conference, Atlanta, GA, 3-6 April, AIAA Paper 2000-1388.
- Leemis, L. (1995). *Reliability: Probabilistic Models and Statistical Methods*. Englewood Cliffs, NJ: Prentice-Hall.
- Donoho, D. and Liu, R. (1988). Pathologies of some minimum distance estimators. *The Annals of Statistics* **16**(2), 587-608.

April 30, 2009.

Table 1
Best distributions identified: Exp(1)

	Censoring rate = .25		Censoring rate = .50		Censoring rate = .75	
	Frequency	KS	Frequency	KS	Frequency	KS
Exponential	0.35	0.062	0.32	0.095	0.16	0.181
Normal	0.01	0.094	0.10	0.113	0.04	0.158
Weibull	0.42	0.057	0.29	0.084	0.08	0.117
Log normal						
Log logistic	0.21	0.059	0.26	0.088	0.22	0.084
Gompertz	0.01	0.054	0.03	0.084	0.25	0.078
Gamma						
Exponential power					0.25	0.100

Table 2
Best distributions identified: Weibull(1, 0.5)

	Censoring rate = .25		Censoring rate = .50		Censoring rate = .75	
	Frequency	KS	Frequency	KS	Frequency	KS
Exponential					0.09	0.270
Normal					0.04	0.191
Weibull	0.52	0.049	0.29	0.067	0.18	0.123
Log normal			0.01	0.139		
Log logistic	0.14	0.054	0.24	0.072	0.20	0.100
Gompertz					0.04	0.083
Gamma	0.04	0.044	0.02	0.057	0.02	0.101
Exponential power	0.30	0.053	0.44	0.078	0.43	0.088

Table 3
Best distributions identified: Weibull(1, 1.5)

	Censoring rate = .25		Censoring rate = .50		Censoring rate = .75	
	Frequency	KS	Frequency	KS	Frequency	KS
Exponential			0.01	0.135	0.12	0.255
Normal			0.06	0.086	0.04	0.194
Weibull	0.48	0.049	0.25	0.064	0.11	0.105
Log normal			0.01	0.139		
Log logistic	0.12	0.057	0.21	0.068	0.27	0.100
Gompertz	0.17	0.050	0.27	0.083	0.32	0.092
Gamma			0.04	0.053		
Exponential power	0.23	0.051	0.15	0.066	0.15	0.097

Table 4
Best distributions identified: LogNorm(2, 0.5)

	Censoring rate = .25		Censoring rate = .50		Censoring rate = .75	
	Frequency	KS	Frequency	KS	Frequency	KS
Exponential						
Normal			0.07	0.078	0.26	0.113
Weibull	0.06	0.051	0.16	0.076	0.10	0.100
Log normal	0.37	0.051	0.27	0.079	0.07	0.125
Log logistic	0.44	0.053	0.29	0.075	0.26	0.090
Gompertz			0.06	0.102	0.23	0.132
Gamma	0.13	0.051	0.13	0.066	0.01	0.095
Exponential power			0.02	0.074	0.07	0.188

Table 5
Distances from parameter estimates (θ) to true values: Exp(1)

	Censoring rate = .25				Censoring rate = .50				Censoring rate = .75			
	Mean	StDev	Min	Max	Mean	StDev	Min	Max	Mean	StDev	Min	Max
MKSE	0.10	0.08	0.00	0.36	0.12	0.08	0.00	0.32	0.21	0.14	0.00	0.71
MLE	0.10	0.08	0.00	0.38	0.11	0.08	0.01	0.37	0.17	0.14	0.00	0.83

Table 6
Distances from parameter estimates (θ_1, θ_2) to true values: Weibull(1, 0.5)

	Censoring rate = .25				Censoring rate = .50				Censoring rate = .75			
	Mean	StDev	Min	Max	Mean	StDev	Min	Max	Mean	StDev	Min	Max
MKSE	0.16	0.11	0.01	0.76	0.27	0.21	0.01	1.49	0.51	0.41	0.08	2.64
MLE	0.14	0.08	0.01	0.40	0.17	0.10	0.02	0.50	0.25	0.14	0.03	0.68

Table 7
Distances from parameter estimates (θ_1, θ_2) to true values: Weibull(1, 1.5)

	Censoring rate = .25				Censoring rate = .50				Censoring rate = .75			
	Mean	StDev	Min	Max	Mean	StDev	Min	Max	Mean	StDev	Min	Max
MKSE	0.23	0.20	0.01	1.25	0.30	0.24	0.01	1.16	0.52	0.40	0.02	1.89
MLE	0.21	0.17	0.00	0.90	0.28	0.23	0.03	1.34	0.43	0.32	0.02	1.79

Table 8
Distances from parameter estimates (θ_1, θ_2) to true values: LogNorm(2, 0.5)

	Censoring rate = .25				Censoring rate = .50				Censoring rate = .75			
	Mean	StDev	Min	Max	Mean	StDev	Min	Max	Mean	StDev	Min	Max
MKSE	0.07	0.04	0.01	0.19	0.19	0.62	0.00	6.26	1.94	7.69	0.04	56.8
MLE	0.06	0.03	0.01	0.19	0.09	0.04	0.01	0.19	0.10	0.07	0.00	0.56