

# A Hierarchical Bayesian Approach for Modeling Infant-Mortality and Wearout Failure Modes

Eric Mittman 1\*

Department of Statistics, Iowa State University  
and

Colin Lewis-Beck

Department of Statistics, Iowa State University

April 22, 2017

## Abstract

The text of your abstract. 100 or fewer words.

*Keywords:* 3 to 6 keywords, (don't reuse words appearing in title)

---

\*The authors gratefully acknowledge Bill Meeker for his comments and suggestions

# 1 Introduction

We present an approach to modeling failure data with heavy censoring, multiple censoring and multiple truncation. We extend the generalized limited failure population model (GFLP) of Chan and Meeker which provides a generative model for failure distributions characterized by a "bathtub" hazard function (citation). This model provides an adequate fit to failures of complex systems by differentiating between early and late types of failures as well as the propensity toward early failure. We discuss the case where the data are grouped, suggesting that we might expand the model, accounting for differences among groups. We show that by modeling the data hierarchically, we can deal with groups of various sample sizes by borrowing information across groups. We illustrate our approach on a large set of failure data for hard disk drives made available by the hard drive backup company, Backblaze (citation).

## 1.1 Background

Products can frequently fail due to more than one reason. For example, there are many parts in a washing machine that can break causing the whole machine to fail. Decorative tree lighting can fail if one bulb in the chain burns out. The general name for such products is a series system where the lifetime of the product is the minimum failure time across  $s$  different components or risks Nelson [2005]. A common assumption in series systems is the time to failure for each risk,  $s$ , is statistically independent. Thus, the overall reliability of a unit can be modeled using the product rule across all  $s$  risks. The GFLP model is a special case of the  $s$  independent competing risks model with 2 modes of failure: defects and wearout. Sometimes the cause of the failure is known; but more often the observations do not contain information about the failure mode. The GFLP model was first introduced to describe the early and late failure modes of CB radios Chan and Meeker [1999]. No failure information was available for this data, which made estimation a challenge using maximum likelihood estimation. Other authors, for example, Basu et al., used a Bayesian approach to fit the GFLP model to masked failure data from an engineering system Basu et al. [2003]. More recently, Ranjan et al. considered a competing risk model for infant mortality and wearout as a mixture of Weibull and exponential failure distributions Ranjan

et al. [2015].

## 1.2 Motivation

The goal of this paper is to model, and compare, the lifetime distribution of the various hard drive brands using the GFLP lifetime model. There are numerous features of the data, however, that make standard estimation approaches for this parametric model problematic. The first issue is many hard drives have been running for quite some time before entering the sample. Even hard drives starting after 2013 appear in the data after hundreds of hours of operation. When an observation is at risk before entering the study, this is known as left truncation. Also, although we have four years of data, hard drives rarely fail so many units are still in service. Thus, we have both left truncation and right censoring. The other challenge is each hard drive model varies in number units in operation and total time on test. Some models have hundreds of units in service; others have as few as two. Combined with the truncation and censoring this makes stable maximum likelihood (ML) estimates difficult to obtain for many drive models.

Another issue is product populations often contain a mixture of defective and nondefective units. The hazard function for this type of population is often described as a bathtub curve: the beginning of the curve corresponds to defective units failing early, followed by a constant hazard, and then an upswing as units fail from wearout. Ignoring this hazard structure, which is known to exist for computer disk drives, could lead to spurious inference when comparing the reliability of hard drive brand models Chan and Meeker [1999]. The general limited failure population model (GFLP) combines different failure time distributions to account for bending hazard functions; for example, one parametric model for defective units (infant mortality) and a second model for wearout Chan and Meeker [1999]. Unfortunately, as often the case with lifetime data, the true causes of failure for the hard drives are unknown, which presents challenges for parameter identification in the GFLP model, as discussed in Chan and Meeker (1999). Therefore, we propose to fit the GFLP model using a hierarchical Bayesian approach. With over 60 brand drive models in testing a hierarchical framework is advantageous as it pools information from across brands to

get more precise estimates of the individual parameters. This is especially helpful for the Backblaze data as there are many drive models with as few as one or two failures.

### 1.3 Overview

The structure of the paper is as follows. Section 1 summarizes the Backblaze data and the unique modeling challenges it presents. Section 2 introduces the GFLP model as a mixture of two Weibull distributions. In Section 3 we fit multiple specifications of the GFLP model to best model the Backblaze data. We also discuss computational and modeling issues that suggests a final model with a common Weibull distribution for defective units and individual Weibull distributions for the wearout failure mode. In Section 4 we compare hard drive brand models based on our final GFLP model. Finally, in Section 5 we review the GFLP model and propose future extensions.

## 2 Data

Backblaze is a company that offers backup storage to protect against hard drive failure. While selling storage space is Backblaze's main business, since 2013 it has been collecting data on hard drives operating at its facility. The purpose is to provide consumers and businesses with reliability information on different hard drive brands and models. The hard drives continuously spin in controlled storage pods where they run until failure. When a hard drive fails it is permanently removed, and new hard drives are regularly added to the storage pods. In addition, the number of storage pods is increasing as Backblaze adds new hard drive brands to the sample. Every quarter Backblaze makes its data publicly available through their website [bac](http://bac.backblaze.com). In addition, Backblaze publishes summary statistics of the different hard drive models currently operating. No other analysis or modeling of the failure data is provided; however, Backblaze does encourage the public to further analyze its data, which for this paper goes through the first quarter of 2016.

As of the first quarter of 2016, Backblaze was collecting data on 63 different hard drive models. Some drive models have been running since 2013, while others have been added at a later date. There are a total of 75,297 hard drives running. The distribution of drives

by model varies: some models have only 1 drive in testing; the maximum number of drives running for a specific drive model is 35,860. We can also look at the distribution of total number of failures and total time on test across all drive models. As seen below, some drive models contain a lot of information, while others have few failures or a little time in operation. [INSERT FIGURE 1].

Probability plotting is an effective method to check the adequacy of various parametric models to the data. Identifying whether failure data follows a specific distribution is difficult to do by eye: especially when looking at pdfs or other non-linear plots. However, if we can linearize the cumulative distribution function it is easier to visually assess a distributional goodness of fit; if a model is appropriate, a non parametric estimate of  $F(\hat{t})$  graphed on linearized probability scales should approximately follow a straight line.

For the hard drive data we first estimate the empirical cdf for each model using the Kaplan-Meier estimator E. L. Kaplan [1958]. With left truncation, however, the Kaplan-Meier estimate,  $\widehat{F(t)}_{KM}$ , is conditional based on hard drive  $i$  surviving up to time  $\tau_i^L$  where  $L$  is the amount of time the hard drive was running before Backblaze starting monitoring the drive. To get the unconditional distribution we correct the non-parametric estimates using a parametric adjustment outlined by Turnbull, and given in more detail by Meeker and Escobar Turnbull [1976], Meeker and Escobar [1998]. For each hard drive model we select  $\tau_{\min}^L$ , the smallest left truncated time in the sample. Then, using the parametric estimates of the parameters, we calculate  $Pr(T > \tau_{\min}^L)$  the probability a hard drive has survived up to  $\tau_{\min}^L$ . We then apply this correction to the non parametric estimates, which results in the unconditional distribution of time to failure.

In Figure 2 we plot the Kaplan-Meier adjusted cdf for hard drive model X on Weibull paper. Each point on the plot corresponds to a hard drive failure. Censored drives are not plotted. As mentioned in the introduction, the population of hard drives exhibits two primary failure modes. One mode is a result of manufacturing defects, which cause early failures, known as infant mortality. The second mode is non-defective hard drives that

eventually fail due to wearout. Evidence of at least two failure modes is seen in the Kaplan Meier plot with a kink occurring around hour Z. Therefore, fitting a single Weibull model would not be flexible enough to model the failure distribution.

### 3 Hierarchical GLFP model

For modeling the lifetime of hard-drives, we select the Generalized Limited Failure Population model of Chan and Meeker [1999]. Let  $T_{d,i}$  be the time of failure for the  $i^{th}$  drive of drive-model  $d$ . We assume that  $T_{d,1}, \dots, T_{d,n_d}$  are independent and have a probability distribution with cdf given by

$$P(T_{d,i} \leq t) = 1 - (1 - \pi_d F_{d1}(t))(1 - F_{d2}(t)), \text{ for } t > 0, \text{ and where } \pi_d \text{ is in } (0, 1).$$

As in Chan and Meeker [1999], we assume  $F_{dj}$  is a member of the Weibull family of cdfs and parameterize in terms of a log-location parameter  $\mu_{dj}$  and log-scale parameter  $\sigma_{dj}$  so that

$$F_{dj}(t) = 1 - \exp \left\{ - \exp \left\{ \frac{\log(t) - \mu_{dj}}{\sigma_{dj}} \right\} \right\}, \quad t > 0, j = 1, 2$$

For the purpose of exposition, we will refer to the probability distributions  $F_{d1}$  and  $F_{d2}$  as “failure modes.” Specifically, we will refer to  $F_{d1}$  as the “early failure mode”, and  $F_{d2}$  as the “main failure mode.” The interpretation of the parameter  $\pi_d$  is to represent the proportion of units susceptible to early failure, hence susceptible to both failure modes. Here the cause of failure is not assumed to be known, thus units of the same drive-model are exchangeable.

To borrow strength across models, we can either share parameters across drive-models, or model the drive-model specific parameters hierarchically, allowing the data to inform the hyperparameters. For the second option, we model the scales,  $\sigma_j$ , quantiles,  $t_{p_j,d,j} = \Phi^{-1}(\mu_j)$ , and proportions,  $\pi_d$ , of the component distributions as follows:

$$\sigma_{d,j} \stackrel{ind.}{\sim} \text{Lognormal}(\eta_{\sigma,j}, \tau_{\sigma,j}^2) \text{ for } j = 1, 2 \text{ } d = 1, \dots, D$$

$$t_{p_j,d,j} \equiv \mu_{d,j} + \sigma_{d,j} \Phi^{-1}(p_j) \stackrel{ind.}{\sim} \text{Normal} \left( \eta_{t_{p_j},j}, \tau_{t_{p_j},j}^2 \right) \text{ for } j = 1, 2 \text{ } d = 1, \dots, D$$

$$\text{logit } \pi_d \stackrel{ind.}{\sim} \text{N}(\eta_{p_i}, \tau_{p_i}) \text{ for } d = 1, \dots, D.$$

Here,  $\Phi^{-1}$  is the quantile function of the standard log-Weibull distribution. The decision to parameterize in terms of a quantile other than the log-location parameter,  $\mu = \Phi^{-1}(.5)$ , is that lifetime data often features heavy right-censoring where inferences about the location parameter are extrapolations beyond the range of the data. For this data we selected  $p_1 = 0.5$ , (the median), and  $p_2 = 0.2$ .

We consider the following set of restrictions:

Model 1:  $\pi_d = \pi$ ,  $\mu_{d1} = \mu_1$ ,  $\sigma_{d1} = \sigma_1$ ,  $\mu_{d2} = \mu_2$ ,  $\sigma_{d2} = \sigma_2$

Model 2:  $\pi_d = \pi$ ,  $\mu_{d1} = \mu_1$ ,  $\sigma_{d1} = \sigma_1$ ,  $\sigma_{d2} = \sigma_2$

Model 3:  $\pi_d = \pi$ ,  $\mu_{d1} = \mu_1$ ,  $\sigma_{d1} = \sigma_1$

Model 4:  $\mu_{d1} = \mu_1$ ,  $\sigma_{d1} = \sigma_1$

The set of models to compare was chosen based on the data, interpretation of the model, as well as estimation considerations. For all model specifications the location and scale parameter of the first failure mode was fixed across drive brand models. Partly this was due to a lack of early failures in the data. Moreover, though, from a consumer or business standpoint, comparing the lifetime distribution of infant mortality across hard drive brand models is not particularly informative. The more practical questions are: is this hard drive likely to suffer from infant mortality? and how long should it expect to run before it fails from wearout? Assuming a constant failure distribution for infant mortality allows the first question to be answered since  $\pi$ 's are comparable across brands, which would not be the case if each drive brand model had a unique Weibull distribution fit to the early failure mode. Flexibility in the parameters for the 2nd failure mode is necessary as initial plots of the data indicate heterogeneity across brands in the right tails of the failure distribution. Going from a fixed model for all hard drive brands and gradually increasing

the complexity of the model, we end up at Model 4, which has the most parameters. Model 4 allows for the probability of infant mortality as well as the shape and scale parameters for the 2nd failure mode to vary by brand drive model.

### 3.1 Priors

To complete the full probability model, we need to select prior distributions for the parameters governing the hierarchical model. We select proper priors to ensure a proper posterior. For the scale hyper-parameters we follow the recommendations of Gelman et al. [2014] and use half-Cauchy priors. As for the location hyper-parameters, we selected mildly informative priors consistent with our prior understanding of hard-drives. We believe that using such priors improves the identifiability of the model parameters while still allowing for substantial Bayesian learning about the hierarchical distributions.

$$\eta_{\pi} \sim \text{Normal}(-3, 1)$$

$$\tau_{\pi} \sim \text{Cauchy}^+(0, 1)$$

$$\eta_{\sigma,2} \sim \text{Normal}(0, 2)$$

$$\tau_{\sigma,2} \sim \text{Cauchy}^+(0, 1)$$

$$\eta_{t_{p2},2} \sim \text{Normal}(9, 2)$$

$$\tau_{t_{p2},2} \sim \text{Cauchy}^+(0, 1)$$

### 3.2 Computation

Discuss the MCMC approach, RStan, etc.

## 4 Data analysis

### 4.1 Results

Show some of the fits of the GFLP Model. Also present the unpooled output for a few models. Show parameter estimates with uncertainty. Mention some convergence statistics



for the MCMC.

## 4.2 collapsing $\mu_1, \sigma_1$

Based on posterior plots, we might fit a simpler model...

## 4.3 Brand Comparisons

How can we use this model to compare brands? Present comparisons of Quantiles, other parameters. Highlight how this model could be used in an applied sense. Quantiles by brand, etc.

# 5 Concluding Remarks and Extensions

Review the advantages of fitting the GFLP model and offer future ideas.

## SUPPLEMENTARY MATERIAL

Put R Stan code here

## References

Backblaze hard drive data sets. <https://www.backblaze.com/b2/hard-drive-test-data.html>. Accessed: 2016-4-1.

Sanjib Basu, Ananda Sen, and Mousumi Banerjee. Bayesian analysis of competing risks with partially masked cause of failure. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 52(1):77–93, 2003. ISSN 00359254, 14679876. URL <http://www.jstor.org/stable/3592633>.

Victor Chan and William Q Meeker. A failure-time model for infant-mortality and wearout failure modes. *IEEE Transactions on Reliability*, 48(4):377–387, 1999.

- Paul Meier E. L. Kaplan. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958. ISSN 01621459. URL <http://www.jstor.org/stable/2281868>.
- Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.
- W.Q. Meeker and L.A. Escobar. *Statistical Methods for Reliability Data*. Wiley Series in Probability and Statistics. Wiley, 1998. ISBN 9780471673279. URL <https://books.google.com/books?id=VQOXGr1lkDgC>.
- Wayne B Nelson. *Applied life data analysis*, volume 577. John Wiley & Sons, 2005.
- Rakesh Ranjan, Sonam Singh, and Satyanshu K Upadhyay. A bayes analysis of a competing risk model based on gamma and exponential failures. *Reliability Engineering & System Safety*, 144:35–44, 2015.
- Bruce W. Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295, 1976. ISSN 00359246. URL <http://www.jstor.org/stable/2984980>.