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

June 19, 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

Failure distributions of high-reliability products, can be difficult to assess. Accelerated testing tries to address this problem, but requires strong assumptions to extrapolate to use conditions in the field (citation). When field data are available, information may be very limited. In this paper, we propose the use of Bayesian hierarchical modeling to borrow information to improve inferences on lifetime distributions where information is limited by right-censoring and left-truncation.

Failure data for series systems may be collected at the system level by end-users interested only in the lifetime of the system. Chan and Meeker [5] proposed the generalized limited failure population model (GLFP), for failure times in a population where some units are susceptible to infant mortality. This model covers a subset of the class of distributions known as bathtub distributions [14]. Bathtub distributions are characterized by a U-shaped hazard function, where early failures are due to infant mortality and late failure are due to wearout. While complex systems actually exhibit more than two modes of failure, suitable parametric models which can suitably approximate the overall lifetime distribution are useful.

In the case of complex, high-reliability systems where cause of failure is not available, a model which

a. has parameters that can be estimated from the observed data

b. is flexible enough to fit the observed data

c. in interpretable with respect to available theory of the underlying process

is desired. We present a Bayesian, hierarchical modeling approach for grouped failure data. As a benefit of working with Monte Carlo samples, we can easily make inference on a wide range of quantities of interest while accounting for uncertainty. We demonstrate this approach on real data hard drive failure data from a cloud-based storage company.

1.1 Background

In engineering applications, a product can often fail from one out of a set of possible malfunctioning components. For example, a computer system can fail if the mother board, disc drive or power supply stop working. Circuit boards (CB) can fail due to a manufacturing defect or later as a result of wearout. 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 [13]. 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 is modeled using the product rule across all s risks. Parameter estimation is straightforward if the cause of failure is known for each observation. With engineering systems data, however, the exact cause of failure is frequently unknown or masked from the researcher.

Previous papers have employed various data assumptions and methodologies to model masked lifetime failure data. When modeling computer system failures Reiser et al. assumed each observed failure came from a known subset of failure modes, and estimation was performed using a Bayesian approach [16]. Chan and Meeker labeled the cause of circuit board failures as infant mortality, unknown, or wearout based on the time of observed failures. This helped identify parameters when using maximum likelihood (ML) estimation. Extending Chan and Meeker's analysis, Basu et. al performed a Bayesian analysis with informative priors to better identify early versus late failure modes without making any data assumptions [2]. Berger and Sun introduced the Poly-Weibull distribution where cause of failure is the minimum of a several Weibull distributions [3]. More recently, Ranjan et al. considered a competing risk model for infant mortality and wearout as a mixture of Weibull and exponential failure distributions [15]. Treating the unknown failure modes as incomplete data, an expectation maximization algorithm with ML was used, in addition to Bayesian estimation.

1.2 Motivation

The goal of this paper is to model, and compare, the lifetime distribution of different hard drive brands using the GLFP lifetime model. Product populations often contain a mixture of defective and non defective 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 and upswing as units fail from wearout. Ignoring this hazard structure, which exists for series systems, could lead to spurious inference when comparing the reliability of hard drive-models [14]. The GLFP 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 [5].

While the GLFP model allows for multiple hazard functions, it does not account for the group structure of the hard drive data. A naive approach would combine all drive-models and fit an overall GLFP model. This model can be easily estimated, but a common GLFP model for all drive-models fits poorly: the model captures drive-models with lots of observed failures, while the predicted time to failure for drive-models with less information is over or under estimated. At the other extreme, fitting individual models to each drive-mode is difficult for drive-models with few failures. Due to multimodality, in these cases the model fit is unstable unless strongly informative priors are chosen.

We therefore propose to fit the GLFP model using a hierarchical Bayesian approach. With over 60 drive-models in operation, a hierarchical model is advantageous as allows for borrowing of information to improve estimation of the GLFP parameters for each drive-model. Moreover, the hierarchical model is a nice compromise between the aggregate and individual modeling approaches: drive-models with lots of failures are estimated precisely; and in cases where little data is available the hierarchical model borrows information from other drive-models to obtain stable estimates. This allows us to include more of the hard drive data and more importantly, make valid inferences for each group of hard drive brands.

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 GLFP model as a mixture of two Weibull distributions. In Section 3 we fit multiple specifications of the GLFP model to best describe the Backblaze data. We also discuss computational and modeling issues that suggest 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 the different model specifications, in addition to the hard drive-models based on our final GLFP model. Finally, in Section 5, we review the GLFP model and propose future extensions.

2 Data

Backblaze is a company that offers backup storage to protect against data loss. 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-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 drive-models to the sample. Every quarter Backblaze makes its data publicly available through their website [1]. 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 drive-models. Some drive-models have been running since 2013, while others were added at a later date. There are a total of 75,297 hard drives in operation. The distribution of drives by drive-model varies: some drive-models have only have a service record for a single drive whereas the maximum number of service records for a single drive model is 35,860. Figure 1 shows the distribution of total number of failures and total observed running time for drive-models with at least 3 failures. For model identification, a minimum of 3 failures was the criterion for hard drive brands to be included in the GLFP model.

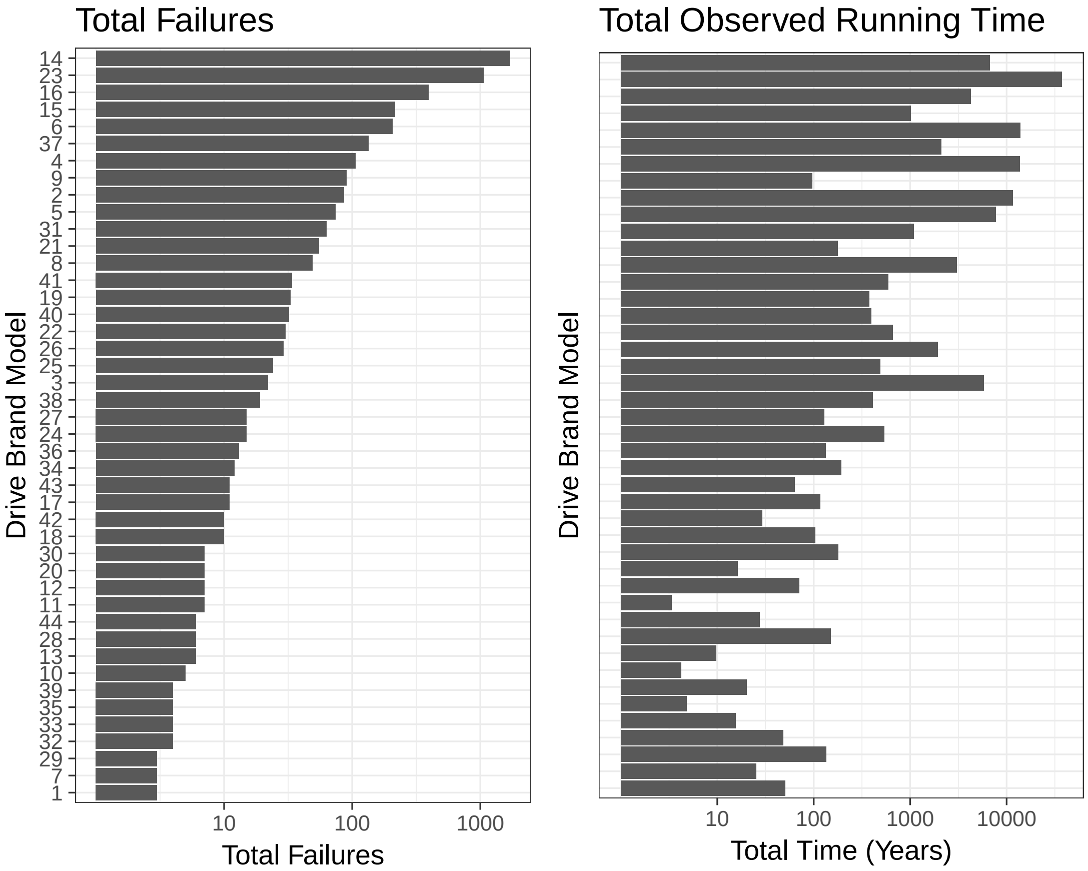

Figure 1: Left: Total number of failures (log scale) for drive-models with at least 3 observed failures in the Backblaze data. Right: Total observed time in operation (log scale) for each of the drive-models in the Backblaze data. Both plots are sorted based on the total number of failures.

Probability plotting is a simple method to assess the adequacy of (log)location-scale families of probability models to the data. Identifying whether failure data are consistent with a specific family of distributions is difficult to do by examining a histogram or other density estimate. By transform the y-axis of the estimated cumulative distribution function (cdf), denoted by $\widehat{F(t)}$, by the inverse cdf of for the standardized distribution, a visual comparison of the plot to a straight line can be an effective informal test for distributional goodness of fit.

Applying this method to the hard drive data, we first must estimate the empirical cdf for each drive-model using the Kaplan-Meier estimator [7]. With left truncation, however, the standard Kaplan-Meier estimator for drive-model d, denoted by $\widehat{F_d(t)}_{KM}$, is conditional on survival up to $t^L_{d,\min}$, the shortest reported running time of all units of drive-model d for which records are available. To produce unconditional estimates, we adapt the adjustment method outlined by Turnbull, and given in more detail by Meeker and Escobar [17, 12]. For each drive-model we select $t^L_{d,\min}$, the smallest left truncated time in the sample. By sampling from the full posterior distribution, since $Pr(T > t^L_{d,\min}|\theta_d)$ (the probability a hard drive has survived up to $t^L_{d,\min}$) is a function of the model parameters, we can easily compute its posterior median, $\widehat{A}_{\text{med}} = \widehat{Pr}(T > t^L_{d,\min}|\theta_d)$. We compute the adjusted estimate by

$$\widehat{F(t)}_{KMadj} = \widehat{A}_{\text{med}} + \left(1 - \widehat{A}_{\text{med}}\right) \widehat{F_d(t)}_{KM},\ t > t^L_{d,\min}.$$

While this adjustment is negligible for the majority of drive-models, five drive-models receive upward adjustments of greater than 5 percent and the estimated time to failure distribution of one drive-model (30) was adjusted by nearly 16 percent, in part because the shortest truncation time for all observed units was approx. 2.3 years.

In Figure 2 we plot the Kaplan-Meier adjusted cdf for drive-model 14 on Weibull paper. Drive-model 14 had the most observed failures of all the brands. Each point on the plot corresponds to an observed failure. Censored drives are not plotted. Standard error bands are calculated using Greenwood's method [9]. 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 between 8000 and 20,000 hours. Therefore, fitting a single Weibull model is not flexible enough to model the failure distribution.

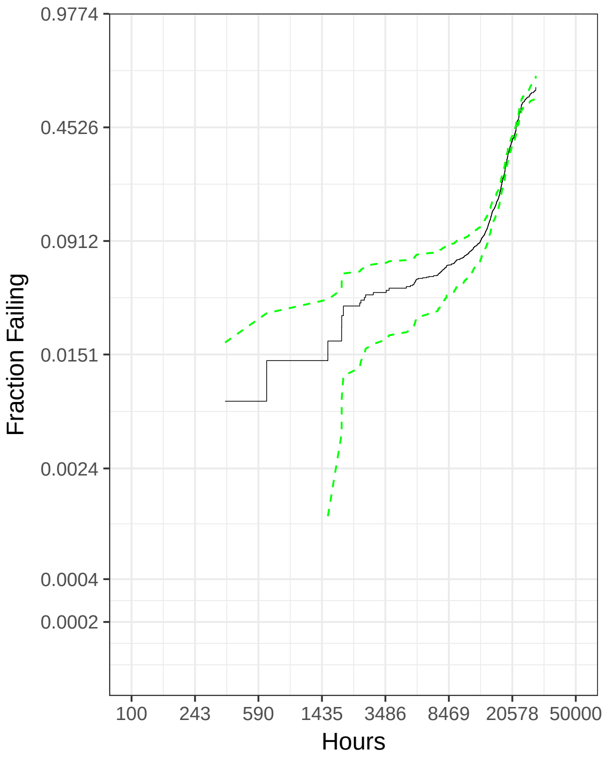

Figure 2: Truncated Adjusted Kaplan-Meier plot showing nonparametric estimate of fraction failing for drive-model 14. Dashed lines are Greenwood standard errors. Plotted on Weibull paper.

3 Hierarchical GLFP model

For modeling the lifetime of hard drives, we select the Generalized Limited Failure Population model of Chan and Meeker [5]. 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 [5], we assume F_{dj} is a member of the Weibull family of cdfs and parameterize in terms of a log-location parameter μ_{dj} and log-scale parameter σ_{dj} so that

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

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 π_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 exchangable.

To borrow strength across drive-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, σ_j, quantiles, $t_{p_j,d,j}$, and proportions, π_d, of the component distributions as follows:

$$\sigma_{d,1} \overset{ind.}{\sim} \text{Lognormal}\left(\eta_{\sigma,1}, \tau^2_{\sigma,1}\right) \text{ for } d = 1, \ldots, D$$

$$\sigma_{d,2} \overset{ind.}{\sim} \text{Truncated-Lognormal}\left(\eta_{\sigma,2}, \tau^2_{\sigma,2}, 0, 1\right) \text{ for } d = 1, \ldots, D$$

$$t_{p_j,d,j} \equiv \mu_{d,j} + \sigma_{d,j}\, \Phi^{-1}(p_j) \overset{ind.}{\sim} \text{Normal}\left(\eta_{t_{p_j},j}, \tau^2_{t_{p_j},j}\right) \text{ for } j = 1,2 \; d = 1, \ldots, D$$

$$\text{logit}\, \pi_d \overset{ind.}{\sim} \text{N}(\eta_\pi, \tau_\pi) \text{ for } d = 1, \ldots, D.$$

Here, Φ^{-1} is the quantile function of the standard log-Weibull distribution. We truncate $\sigma_{d,2}$ at 1, appealing to the logic that,by nature, wearout produces an increasing hazard function. The decision to parameterize in terms of a quantile other than the log-location parameter, $\mu = t_{0.632}$, is that lifetime data often, as is true of the data presented here, 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 models with the following set of restrictions (from most to least restrictive):

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

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

3. $\pi_d = \pi, \quad \mu_{d1} = \mu_1, \quad \sigma_{d1} = \sigma_1$

4. $\mu_{d1} = \mu_1, \quad \sigma_{d1} = \sigma_1$

The set of model specifications were chosen based on the data, interpretation of the model, as well as estimation considerations. Drive brand specific parameters for the the wearout failure mode (μ_{d2}, σ_{d2}), and the proportion defective (π_d), are considered as a means to account for heterogeneity across brands in the right tails of the failure distribution. Going from a common model for all hard drive brands and gradually increasing the complexity of the model, we end up at Model 4. 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 drive-model.

For all of the models we consider, the parameters for the infant mortality failure mode are held in common across drive-models. We found that there was insufficient information in the data to model these parameters hierarchically. Moreover, assuming a common distribution for infant mortality provides a meaningful interpretation and comparison of π_i and π_j $(i \neq j)$.

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 models 1-4, different sets of restrictions required different prior specifications, which were assigned as follows:

1. This model constrains all drive-models to the same failure distribution. For this "reduced" model we assume the priors :

$$\text{logit}^{-1}\pi_1 \sim \text{Normal}(-3, 2), \quad \sigma_1 \sim \text{Lognormal}(0, 2.5), \quad t_{0.5,1} \sim \text{Normal}(8, 4)$$

$$\sigma_2 \sim \text{Lognormal}(0, 2.5) \quad t_{0.2,2} \sim \text{N}(10, 4)$$

The prior on π implies with 90% probability π is in the interval $(.002, .572)$. The prior on σ_1 and σ_2 put it on the interval $(0.02, 61.1)$ with 90% probability. (Note that, due to symmetry, the Weibull shape parameter has the same prior interval and that a value of 1, corresponding to a constant hazard, is the median of this prior distribuion.) $t_{0.5,1}$ has a prior 90% interval of $(1.4, 14.6)$. For $t_{0.2,2}$ it is $(3.4, 16.6)$. Since these last two parameters are on the scale of log-hours, admitting values from hours to decades, we consider them to be relatively uninformative.

2. We allow t_{pd} to vary by drive-model. To help with model identifiability, we tighten the priors on the defective mode:

$$\text{logit}^{-1}\pi \sim \text{Normal}(-3, 1), \quad \sigma_1 \sim \text{Lognormal}(0, 1), \quad t_{p1} \sim \text{Normal}(7, 2),$$

   implying now a prior 90% probability that σ_1 is between 0.19 and 5.18 and $t_{0.5,1}$ between 3.7 and 10.3, which, translating from the log-scale, is equivalent to an interval from 1.5 days to 3.4 years.

3. σ_{d2} is allowed to vary. Priors for constrained parameters are the same as for model 2.

4. π_d varies. Priors for constrained parameters are the same as for model 2.

For the scale hyper-parameters we follow the recommendations of Gelman et al. [8] and use half-Cauchy priors. As for the location hyper-parameters, we select weakly informative priors consistent with our prior information on hard-drives. The prior for η_π puts 95% of prior mass on the interval $(0.006, 0.27)$ for the median proportion defective (What justification? Should maybe be lower.) For $\eta_{\sigma,2}$, 95% of prior mass is on values greater than 0.037. Since σ_d is the reciprocal of the Weibull shape parameter, this correponds roughly to an assumption that the median Weibull shape parameter is less than $1/0.037 = 27$. The prior for $\eta_{t_{p_2},2}$ implies that the median 20th percentile for non-defective units is less than greater than 3 days and less than 24 years.

$$\begin{aligned}
\eta_\pi &\sim \text{Normal}(-3, 1) \\
\tau_\pi &\sim \text{Cauchy}^+(0, 1) \\
\eta_{\sigma,2} &\sim \text{Normal}(0, 2) \\
\tau_{\sigma,2} &\sim \text{Cauchy}^+(0, 1) \\
\eta_{t_{.2_2},2} &\sim \text{Normal}(9, 2) \\
\tau_{t_{.2_2},2} &\sim \text{Cauchy}^+(0, 1)
\end{aligned}$$

3.2 Computation

Each model was fit using the `rstan`[10] package in `R` [11], which implements a variant of Hamiltonian Monte Carlo (HMC) [4]. Multiple chain were run for 1500 iterations after 1500 warmup iterations. Convergence was assessed by examining posterior plots and checking that potential scale reduction factors (Rhat) [8] were less than 1.1. Four chains were run for Models 1,2 and 3 and 16 chains were run for Model 4.

4 Data analysis

4.1 Model Comparisons

The right side of Figure 3 shows that the set of estimated failure curves for all four models that we fit. The upper left panel shows the estimated failure curve fit to all data ignoring drive-model as a factor (Model 1). The upper left plot, which represents Model 2, demonstrates that there is appreciable differentiation between drive-models; note that, since the scale parameter, $\tau_{t_{p2}}$ is learned hierarchically from the data, we can ascertain that there is evidence in the data that this variability is real.

The lower left panel, representing Model 3, suggests differing shape parameters in the wearout failure mode which mean that the ranking of drive-models with respect to fraction failing differs over time.

Finally, the lower right panel, corresponding to Model 4, shows that, by allowing π_d to vary by drive-model, there is greater variability in the early rate of failure and, for some drive-models, the forecasted fraction failing is adjusted downward.

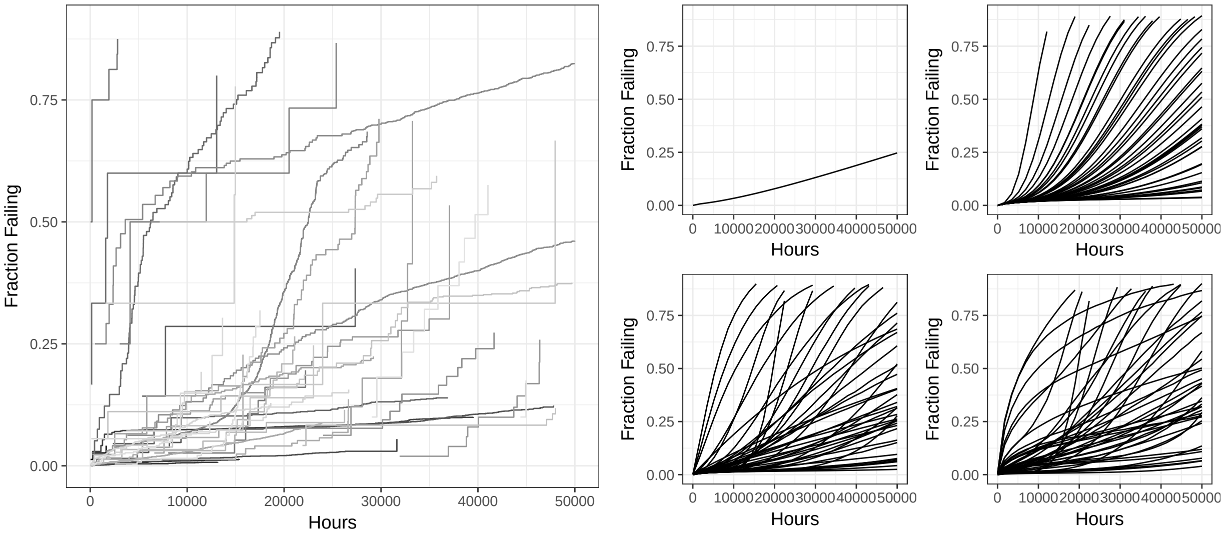

Figure 3: Left: Kaplan-Meier estimates of the time to failure for each of the drive-models in the Backblaze data. Right: Pointwise posterior median time to failure curves for Models 1-4, ordered left to right, top to bottom.

We can also compare the model fits for each drive-model individually. Using the set of parameters that maximize the posterior log probability, we examine the GLFP cdf for each of the 4 model specifications. The GLFP fit for drive-models 2, 9, 14, and 40 are presented in Figure 4. The black step function again corresponds to the adjusted Kaplan-Meier estimates. As model complexity increases, the GLFP curves become more flexible and are better able to fit the observed failure data. All the GLFP models appear to overestimate the proportion of failed hard drives for drive-model 2, but Model 4 agrees the most with the nonparametric estimate. For drive-model 14, Models 1-3 underestimate the proportion of drives failing, and for drive-model 40, they slightly overestimate the data. Conversely, Model 4 follows the observed failures quite closely–accurately capturing failures at the lower and upper ends of the distribution. The difference in accuracy between Models 3 and 4 are and Models 1 and 2 are visually quite apparent. Differentiating between Model 3 and 4 is less clear; for instance, the fits from Model 3 and 4 line almost exactly on top of each other for drive-model 40.

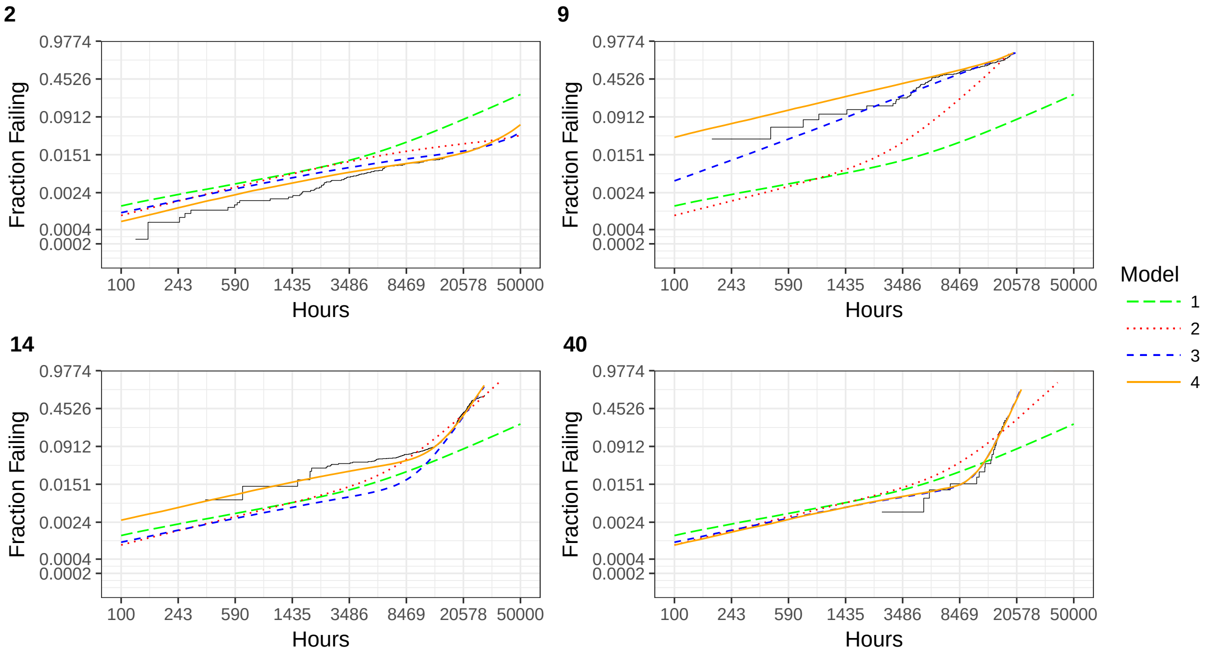

Figure 4: Estimated GLFP for 4 drive-models plotted on Weibull paper. Black step function is the truncated adjusted Kaplan-Meier points.

To statistically compare the GLFP Models, we estimated the predictive accuracy of the 4 different models using an approximate leave-one-out cross validation (LOO) method. As outlined in Vehtari et. al, the predictive accuracy from a fitted Bayesian model can be estimated simply with posterior draws of the model parameters rather than re-fitting the model with different data sets [19]. For each observation, i, the log point-wise predictive density is calculated over the full set of posterior samples with the ith data point removed. The final expected log point-wise predictive density (elpd) is the sum over all observations (elpd $= \sum \log p(y_i|y_{-i})$). We computed the elpd for all 4 models using the R package loo [18]. When n is large, the distribution of the elpd is approximately normal and different models can be statistically compared. We calculated the difference in expected predictive accuracy for Model 4 vs 3, Model 3 vs 2, and Model 2 vs 1, as well as the standard error of the difference (Table 1). As the model complexity increased, the expected predictive accuracy improved. Of all the models, Model 4 had the best predictive accuracy and was significantly better than the other 3 model specifications.

	ELPD	Difference in ELPD	SE of the Difference
Model 4	-13309.5	40.7	11.3
Model 3	-13350.2	458.8	31.0
Model 2	-13809.0	3674.6	96.2
Model 1	-17483.6		

Table 1: Expected Log Pointwise Predictive Density (ELPD) for each model specification. Each model is compared to the more parsimonious model below. The estimated difference of expected leave-one-out prediction errors between the two models, as well as the standard error of the difference, is also presented.

4.2 Drive-model Comparisons

We consider the problem of ranking drive-models. From a business perspective, it is clear that we should prefer the drive-models which will provide more years of service. For ease of exposition, we will assume that the purchase price of a hard-drive is the same across drive-models.

There are two sources of variation at play; the posterior uncertainty in the parameters, which largely depends on the sample size, and the uncertainty in future observations conditional on the parameters. The *posterior predictive* distribution incorporates both sources.

$$p(t_{d,new}|t) = \int p(t_{d,new}|\theta_d)p(\theta_d|t)\, d\theta_d$$

We can sample from this distribution by drawing $t_{d,new}^{(s)}$ from GLFP$(\pi_d^{(s)}, \mu_1^{(s)}, \sigma_1^{(s)}, \mu_{2,d}^{(s)}, \sigma_{2,d}^{(s)})$, for $s = 1, ..., S$, using the saved posterior draws for the model parameters. The expected time-to-failure (TTF) for drive-model d can be estimated by

$$\frac{1}{S}\sum_{s=1}^{S} t_{d,new}^{(s)}.$$

While TTF is clearly important, due to the anticipation of advancements in technology, we can expect that the relative value of computer hardware will depreciate rather quickly. To account for this depreciation, rather than using TTF as the metric for drive-model comparison, we use the value at replacement. The IRS considers computer hardware to be “5-year” equipment[6]. Using“the declining balance method” the rate of depreciation is 40% per year.

Let $L(t) = e^{-.4t}$ represent the value at the time of failure relative to a new unit. We can rank the drive-models by

$$\mathrm{E}(L(t_{d,new})|t)) \approx \frac{1}{S}\sum_{s=1}^{S} L(t_{d,new}^{(s)}).$$

We can also use the posterior draws to compute quantiles of interest, which in reliability applications is typically more informative than a central measure of tendency. On the left panel of Figure 5 we plot B10, the .10 quantile (i.e., the amount of time it takes for 10% of hard drives to fail), for all drive-models. The best six drive-models ranked by TTF are also the drive-models with the longest time to B10. Drive-models 5, 2, 4, and 3 seem to really stick out as the most reliable as there is a larger gap separating them from the rest of the group. Another interesting feature to compare is π, the proportion failing due to infant mortality. One the right panel of Figure 5 are the posterior credible intervals of π for each drive-model. The plot is again sorted according to B10 so it can be compared to the left hand side. While for many of the drive-models the ordinal ranking is the same as on the left, there are some drive-model comparisons, for example 23 and 18, that differ in ranking if compared using B10 or π. Depending on a user’s application or the hard drive warranty length, using the π’s as a comparison tool may be a relevant parameter of interest.

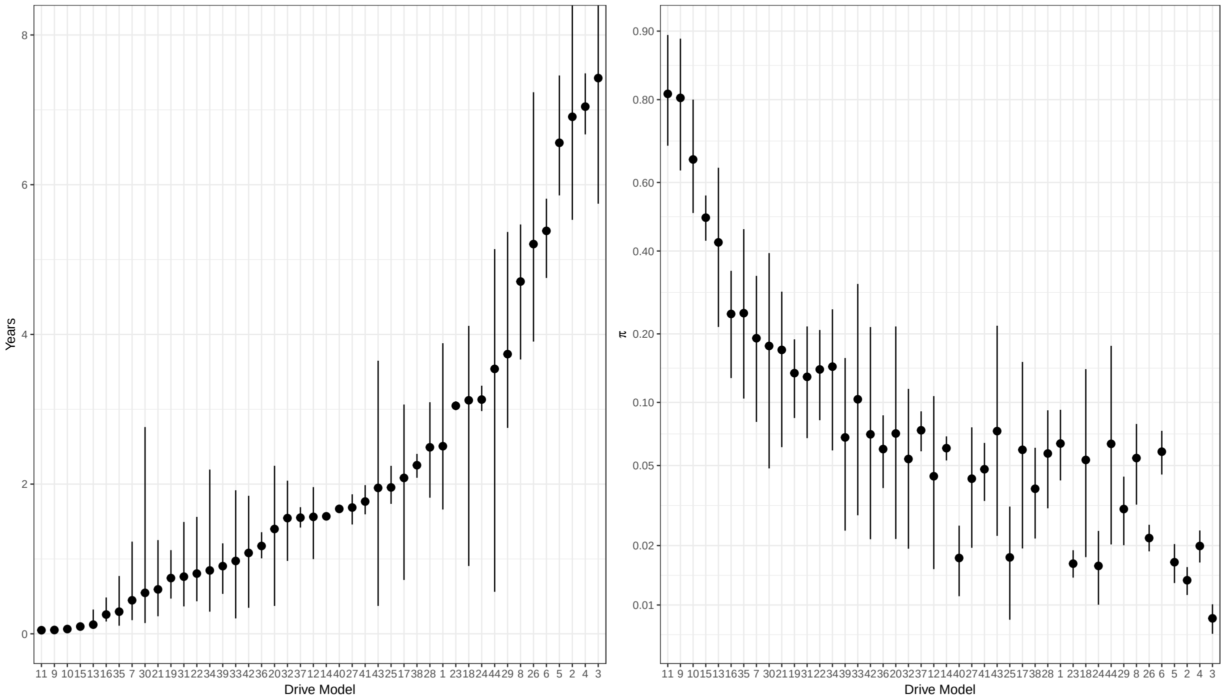

Figure 5: Left: 50% credible intervals for B10 (time in years till 10% of the drives fail). Right: 50% credible intervals for π plotted on the logit scale. Both plots are sorted based on the median value of B10.

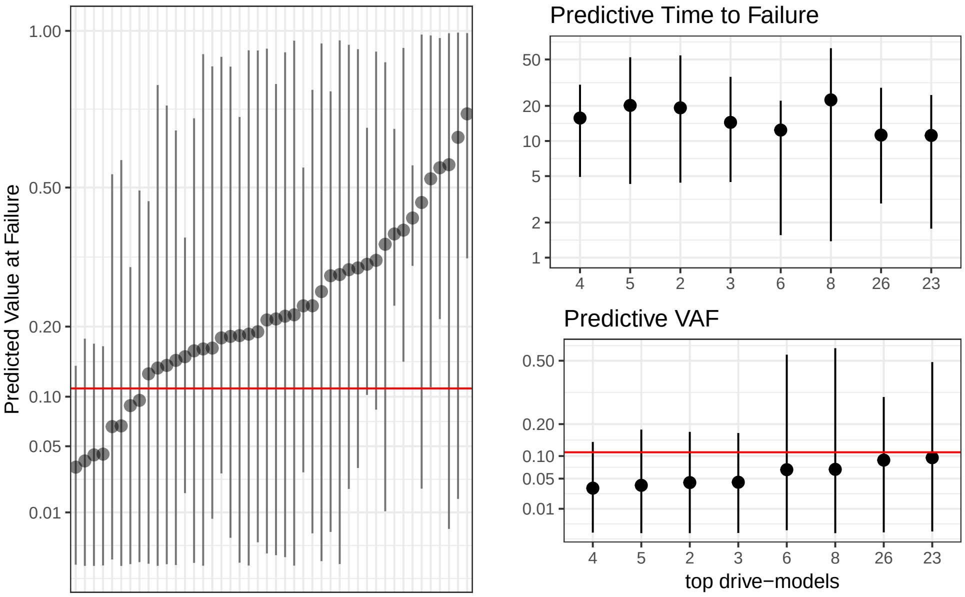

Figure 6: Left: Posterior predictive relative value-at-loss for new unit, by drive-model. Right: TTF (top) and VAF based on 40% annual depreciation (bottom), for best 11 drive-models.

5 Conclusions

We should point out some assumptions made in this analysis. First, we assume exchangability of units within drive-models. This assumption is due to our ignorance about the causes of failure and of potentially important covariates. If we had information about distinguishing characteristics of the two identified failure modes, then we should use this information. Also, there could be many environmental or human-related factors that impact lifetime, such as operating conditions associated with location in the data center, which are likely confounded with drive-model.

Second, it is probably true that many modes of failure occur and that the heterogeneity that we observe among drive models is due to different combinations of these. Our ignorance of the causes of failure leads us to approximate the main failure mode with a single drive-model specific Weibull distribution. Because of this, we suspect that our forecasts are too confident; especially for "well-estimated" models, our uncertainty in the right tail of the distribution is probably too small. On the other hand, our model borrows strength across drive-models, including those with information about late failures. In addition, we suggest that using metrics for comparing drive-models that reduce the importance of the right tail of the distribution are appropriate for these data. With respect to decisions based on such metrics, we believe the assumptions in the previous paragraph to be the most critical. Those issues are typical to observational data such as these.

Consumers of high-tech manufactured goods usually lack the expertise to perform a failure analysis to determine the cause of failure. The ability to fit marginal models, such as GLFP, can allow investigators to answer many pertinent questions while still appropriately accounting for uncertainty and avoiding substantial model bias.

Advances in technology have made computationally intensive Bayesian methods for fitting hierarchical models feasible in practice. Hierarchical modeling can be an effective way to regularize a large number of models fits by borrowing information. Because all the data contribute some information about the global parameters, it is often sufficient to use weakly informative priors on the global parameters, making the resulting shrinkage data dependent. Using samples from the full joint posterior avoids the need for potential problematic asymptotic approximations.

Analysis using HMC is accessible to practitioners by Stan. HMC has been shown to be very efficient in effective samples per iteration because its joint updating scheme.

SUPPLEMENTARY MATERIAL

Put R Stan code here

References

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

[2] 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`.

[3] James O Berger and Dongchu Sun. Bayesian analysis for the poly-weibull distribution. *Journal of the American Statistical Association*, 88(424):1412–1418, 1993.

[4] Michael Betancourt and Mark Girolami. Hamiltonian monte carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79:30, 2015.

[5] 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.

[6] IRS Department of Treasury. Instructions for form 4562, 2016. URL `https://www.irs.gov/pub/irs-pdf/i4562.pdf`.

[7] 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`.

[8] 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.

[9] Major Greenwood. The errors of sampling of the survivorship tables. *Reports on public health and statistical subjects*, 33(1):26, 1926.

[10] J Guo, M Betancourt, M Brubaker, B Carpenter, B Goodrich, M Hoffman, D Lee, M Malecki, and A Gelman. Rstan: The r interface to stan, 2014.

[11] Ross Ihaka and Robert Gentleman. R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314, 1996.

[12] 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=VQOXGrllkDgC`.

[13] Wayne B Nelson. *Applied life data analysis*, volume 577. John Wiley & Sons, 2005.

[14] Sujata Rajarshi and M. B. Rajarshi. Bathtub distributions: A review. *Communications in Statistics: Theory and Methods*, 17:2597–2621, 1988.

[15] 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.

[16] B Reiser, Irwin Guttman, Dennis KJ Lin, Frank M Guess, and John S Usher. Bayesian inference for masked system lifetime data. *Applied Statistics*, pages 79–90, 1995.

[17] 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`.

[18] Aki Vehtari, Andrew Gelman, and Jonah Gabry. loo: Efficient leave-one-out cross-validation and waic for bayesian models, 2016. URL `https://CRAN.R-project.org/package=loo`. R package version 1.1.0.

[19] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 2016. doi: 10.1007/s11222-016-9696-4.