**REFEREE REPORT ON MANUSCRIPT TCH-18-013**

“A Hierarchical Model for Heterogeneous Reliability Field Data”

This work developed a hierarchical Bayesian framework for modeling reliability field data, allowing data to be left-truncated, right-censored and from heterogeneous sub-populations. A GLFP model was adopted to handle two types of failures with evidence shown in the probability plot. Though the GLFP model is not new, to combine with the hierarchical structure enhances its usefulness on information sharing in meta analysis. The application to hard drive data is an interesting one, in which the Bayesian inference is well taken as well as the diagnostic procedures. This manuscript shows a high standard research work in a good writing (but slightly too long). Some concerns and comments are given in the following.

* The authors used an alternative parameterization (*tp,σ*) to represent the Weibull distribution. One of the reason is that, with a proper chosen *p*, the resulting estimates of (*tp,σ*) would be approximately uncorrelated. As suggested in the manuscript, the choice of *p* relates to the percentage of failures in the observed data. It is more informative to provide some quantitative comparisons in the applications (such as in Section 4.2) to address the advantage of using (*t*0*.*5*,*1*,t*0*.*2*,*2) over the other choices.

Not sure about this one.

* In the manuscript, the notation h·*,*·i is used to specify the prior in p.11 and p.17, but the standard notations are used elsewhere. Such inconsistency should be avoided. To provide more information about the prior distribution and to show its diffusion, a better way is to visualize the density directly. Therefore, I will suggest to use the standard way to specify the prior and simultaneously provide the density plots for priors, possibly compared across the nested models considered in Section 5.4.

But it is too long already … We can put the plots in the supplementary though.

* In Figure 3, the 90% pointwise credible bands associated with the GLFP posterior are slightly wilder than the confidence bands of the same level associated with the nonparametric K-M estimates. The former is based on a parametric model fitting with proper chosen priors. It seems to be unusual that the estimation accuracy is not improved and even worse than the nonparametric approach. Please explain it.

Technically, it is the estimation precision which is not improved, not the accuracy. I checked for obvious coding mistakes. It seems that the nonparametric model makes different assumptions than the parametric model and, while proper, diffuse priors were used. Interesting comment…

* Provide further evidence about

“The joint hierarchical modeling among multiple drive models does improve the *overall accuracy* for estimating the characteristics of interest (such as MTTF, VAF, or a certain quantile) for individual models, compared to the approach without the hierarchical structure.”

This seems to be the primary goal of doing this project.

This assertion needs to be reviewed. I disagree that this is necessarily the primary goal.

* Table 2 shows the goodness-of-fit for different models in terms of elpd. The associatedd model complexity (degrees of freedom) for each fitted model would also be of interest and informative to understand differences between models, which can be numerical approximated using the MCMC draws analogously.

We can look into this. I have heard of, but am not familiar with this.

* typos:
  + p16, line 9: should it be

... approximating *p*(*tg,i*|*t*−(*g,i*)) (remove log) with(change

the first summation index from *i* to *s*)...

* + Following the 2nd equation in p22, the argument *trep,g* is understood as a *ng*-

dimensional vector. But, in the following statement “... draw *treg,g* from *f*(*tnew*|*tLg,i,cg,i,θ*(*s*))”, *trep,g* is understood as a scalar. Please verify.

Reviewer’s Comments on TCH-18-013: “A Hierarchical Model for Heterogenous Reliability

Field Data”

The paper proposed a hierarchical Bayesian estimation procedure for the GLFP model and applied the procedure to hard drive failure data which are publically available at the Backblaze website. Generally speaking, the paper is well written. The analysis is comprehensive and the insights from the analysis are well discussed. I have the following comments for the authors to improve the paper.

It seems that the proposed hierarchical Bayesian procedure, including the likelihood function in Section 3.4, is tailored to the Backblaze hard drive data. To make the model more generic, it would be good to consider general data types rather than the left truncated and right censored data. For example, for failure data collected from the field, the data are usually interval censored.

We could lay out the equations in a more general way, or we could just add a comment in the discussion.

In the paper, the SEV reparameterization of the Weibull distribution is not used in the subsequent analysis. It should be removed from the paper so that a readers will not be confused between this SEV reparameterization and the p-quantile reparameterization.

Second, the reparameterization of the Weibull is not an original contribution of the paper. Therefore, proper citation of the original paper that proposes this parametrization is needed. Third, the reparametrization involves a quantity p which should be determined by the user. The authors may like to discuss how to select p, and it there rigorous justification for the determination of p.

Similar to previous reviewer’s comment.

The selection of the prior distributions are somewhat arbitrary. The lognormal distribution is chosen as prior distributions for most parameters. Discussion of the sensitivity of the analysis to the prior selection is desired. When there is no prior information, a noninformative prior is more appropriate. We may choose the parameters in the prior distribution so that the ratio between the standard deviation and the mean is very large to make the prior approximately noninformative. In lognormal prior, this can be achieved by using a large scale parameter sigma, which is adopted in the paper. However, it is unclear how the location parameter is chosen in the paper. From the priors chosen in Page 11, the mean of the lognormal prior seems to be deliberately chosen to tally with the MTTF of the data. If this is the case, the data are used twice (for prior selection and for the likelihood), and the estimation errors of the parameters and reliability quantities will be underestimated.

There are a lot of parameters… how to carry out a sensitivity analysis?

The proposed estimation procedure is applied to the HDD failure data, but the small sample performance should be verified using simulation. Personally, I am interested in the following questions that may require simulation verification: (i) what is the performance of MCMC in the Bayesian computation? There are five parameters in the model, and the convergence can be an issue when implementing MCMC. The authors used the Gelman-Rubin’s potentialscale reduction factor in the data analysis for diagnostics and it didn’t indicate problems for this dataset. However, this is partially because the mean of the prior is not very different from the posterior mean. A simulation may be needed to provide insights on the effect of prior selection on the convergence. For example, when the mean of these lognormal priors are fixed at 1, or when an improper uniform distribution is used, then what is the convergence behavior? (ii) What is the performance of the model selection method using the estimated elpd?

We do some simulation with a single population, if need be.

In Bayesian model selection, a popular method is the Bayesian factor. The authors may discuss why elpd is preferred in the HDD data analysis.

I think we can get the Bayes factors from the posterior and check. Bayes factors are sensitive to the choice of prior though.

In addition to the HDD data, the authors may like to discuss to what kinds of products the proposed model is applicable. The current methodological development centers around the HDD data and does not look generic.

Don’t we do this?

Editor’s comment:

In addition to the comments from the Referees and AE, I have a comment of my own that I would like you to address in the revision:  On page 7 you point out how your GLFP model can be viewed as a mixture of two distributions F2 and 1-(1-F1)(1-F2). What are the pros/cons of this particular mixture vs. a mixture of any two distributions, e.g., two Weibull distributions with different modes?  Allowing a mixture of any two survival distributions would be more general. Is the main advantage of your particular mixture that it leads to more tractable analyses? If so, how important is this, considering that you are using MCMC anyway?  Or does your particular mixture have some physical justification that suggests it might fit many real data sets? Please include some discussion on this to convince readers that the route you are taking is the best route.

This particular mixture has a meaningful interpretation and avoids “label-switching” from non-identifiability of the two failure modes that would be present in a general mixture.