Application of Deep Neural Networks to Estimating Flexible Survival Curves

By: Zachary McCaw (SUID: 05933518)[1]

TA: Ruta Joshi (Final project meeting: 2021-11-30)

Introduction

Time to event data arise frequently in the context of randomized clinical trials [1]. Such data are characterized by the presence of right censoring [5]. Starting from the time of randomization, let T denote the time until the event of interest (e.g. death), and C the time until censoring (e.g. study closure). The observed data take the from (U, δ) where $U = \min(T, C)$ and $\delta = \mathbb{I}(T \leq C)$. If $\delta = 1$, then $U = T$ and the event was observed. If $\delta = 0$, then $U = C$, and the event is only known to have occurred after time C. Fitting parametric survival curves to the event-time data is useful for projecting a patient's expected survival experience beyond the duration of study follow-up. A particularly flexible parametric survival distribution is the generalized gamma [4]:

$$f_\theta(t) = \frac{\beta\lambda}{\Gamma(\alpha)}(\lambda t)^{\alpha\beta-1}e^{-(\lambda t)^\beta}, \; t > 0. \tag{1.1}$$

The specific aim of this project was to develop a model for estimating the parameters $\theta = (\alpha, \beta, \lambda)$ of a generalized gamma distribution from right-censored event-time data.

Methods

2.1 Code

Replication code is available on GitHub at `https://github.com/zrmacc/cs230`. Data generation was performed using the `Temporal` package in `R`. Model training and evaluation were performed in `Tensorflow` v2.6.0.

2.2 Data

Generative values of $\theta = (\alpha, \beta, \lambda)$ were drawn from independent uniform distributions on $[0, 4]$, and the standard exponential distribution $\theta = (1, 1, 1)$ was included as a base-case. A total of 100 values of θ were selected. At each θ, event times T were independently simulated for 10^3 training, 10^2 validation, and 10^2 testing patients. Censoring times were independently generated such that 15% of patients in the training, validation, and testing cohorts were censored (in expectation). The total sample sizes were $n_{\text{train}} = 10^5$, $n_{\text{val}} = 10^4$, and $n_{\text{test}} = 10^4$.

[1] I worked independently on this project.

2.3 Model Overview

Various model architectures were explored, all having the following features in common. The goal was to learn the mapping $(U, \delta) \mapsto (\alpha, \beta, \lambda)$. The input dimension was $(m, 2)$, where m is the batch size and the columns corresponded to the observation time U and the status indicator δ. The output dimension was $(m, 3)$, where the columns corresponded to $(\ln \alpha,\ \ln \beta,\ \ln \lambda)$; the parameters were estimated on log scale because each must be strictly positive. An effective estimation procedure should produce parameter estimates $\hat{\theta}$ that are near to the truth θ. Therefore, model performance was evaluated with respect to the mean absolute error (MAE).

2.4 Optimization

Training was performed using the Adam optimizer with default hyperparameters. The learning rate was initialized at 10^{-4}, then scheduled to decay by a factor of 0.5 every 10 epochs thereafter. The batch size was 128 and upto 100 epochs were allowed for training. Because early stopping was permitted if the validation loss failed to improve across 10 consecutive epochs, the maximum number of epochs was never actually reached. The best model (based on validation loss) obtained during training was check-pointed and its weights were used to evaluate performance.

Experiments

3.1 Model Architecture

For this experiment, the loss was the total MAE. The following architectures for the hidden layers were compared:

i. 1 dense layer with 256 units and ReLU activation.

ii. 1 dense layer with 256 units and tanh activation.

iii. 4 dense layers with (256, 256, 256, 128) units, ReLU activation throughout.

iv. 4 dense layers with (256, 256, 256, 128) units. ReLU activation for layers 1-2, tanh for layers 3-4, batch normalization prior to each tanh layer.

Figure (5.3) compares the training, validation, and testing loss across the 4 models. Within an architecture, the loss was consistent across data sets, providing no evidence of over-fitting. Among the shallow (1 layer) models, ReLU activation led to a lower loss. The performance of the two deep (4 layer) models was statistically equivalent, but the deep tanh architecture (iv) was selected to move forward because its performance was numerically better. Figure (5.4) suggests that estimation of the shape parameters α and β was more difficult than estimation of the scale parameter λ. Figure (5.5) provides a visualization of how the model output varies by U and δ. Note that the model has learned to account for censoring, as evidenced by the differing prediction curves censored and uncensored patients.

3.2 Choice of Loss Function

In statistics, parametric survival curves are typically estimated via maximum likelihood. The right censored generalized gamma log likelihood is:

$$\ell(\theta) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \ln f_\theta(u_i) + (1-\delta_i)\ln S_\theta(u_i), \tag{3.1}$$

where f_θ is the density presented in equation (1.1) and $S_\theta(t) = \Gamma\{\alpha, (\lambda t)^\beta\}/\Gamma(\alpha)$.[2] We hypothesized that a model that incorporates the negative log likelihood into its loss function might outperform a model trained simply to minimize the MAE because the former has access to additional information; namely, the likelihood of the distribution that generated the data. For this experiment, a custom loss function of the following form was defined:

$$L(\theta, \hat{\theta}) = (1-\omega)\cdot\left\{-\ell(\hat{\theta}) + n^{-1}||\ln\hat{\theta}||_2^2\right\} + \omega\cdot n^{-1}||\ln\hat{\theta} - \ln\theta||_1. \tag{3.2}$$

In the loss definition, θ is the true parameter value (the target of the model) and $\hat{\theta}$ is the estimated parameter value (the output of the model). $\omega \in [0,1]$ is a hyperparameter that trades off between the penalized log likelihood loss in blue and the MAE loss in red. Larger ω places more weight on the MAE. An L_2-penalty was added to the term in blue because, with additional regularization, gradient descent often diverged (with 1 or more elements of $\hat{\theta}$ becoming excessively large) for ω near zero.

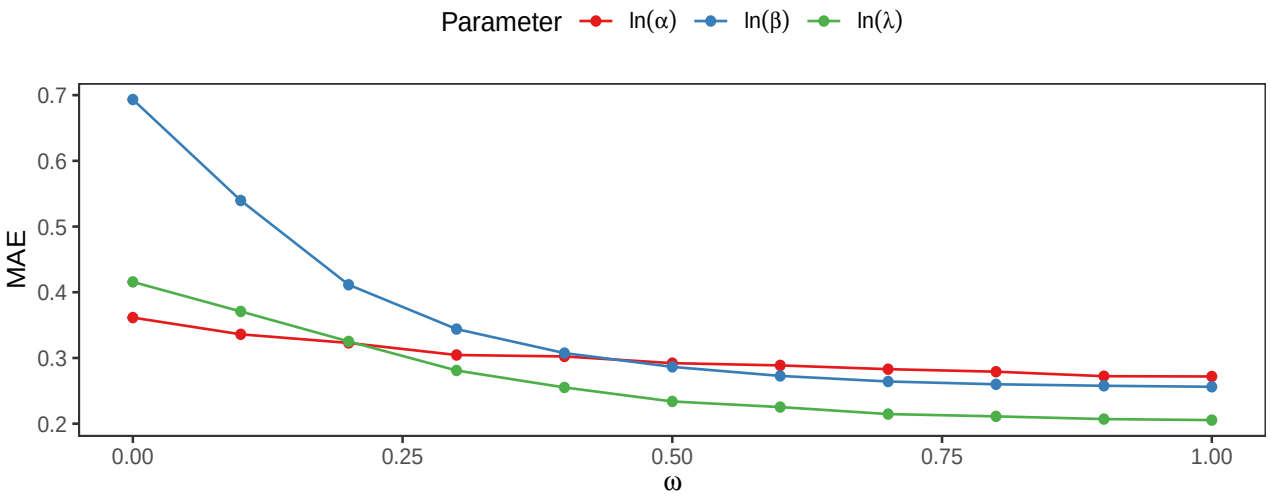

Figure 3.1: Evaluation set MAE across deep tanh models trained with different loss functions. ω refers to the weight given to the MAE term in (3.2).

Figure (5.6) compares the training, validation, and testing loss across models trained with loss functions corresponding to different values of ω. Although the loss curves are not comparable with one another, the consistency of the loss across training, validation, and test sets provides

[2] See section 5.2.1 for a note on numerical evaluation of $S_\theta(t)$.

no evidence of over-fitting. Figure (3.1) demonstrates that having access to the form of the log likelihood did not markedly improve model performance, as the evaluation MAE monotonically decreased with increasing ω. Estimation of β was most affected by ω, while estimation of α was least affected.

3.3 Practical Evaluation

In the context of clinical trials, the treatment arms are often compared with respect to the typical time from randomization to the event of interest. The advantage of fitting parametric survival curves is the ability to extrapolate the mean or median survival times, which often lie beyond the duration of the study. The mean and median[3] of the generalized gamma are respectively given by:

$$\mu(\theta) = \frac{\Gamma(\alpha + \beta^{-1})}{\lambda\Gamma(\alpha)}, \qquad \nu(\theta) = \inf_u\{u : S_\theta(u) \geq 0.5\}.$$

Two sets of evaluation data ($n_{\text{eval}} = 10^4$) were simulated with generative parameters $\theta_1 = (1, 1, 1)$ and $\theta_2 = (2, 2, 3\sqrt{\pi}/4)$. These distributions have the same mean but different medians. The best model from section (5.1), using MAE loss, was applied to estimate $\hat{\theta}_i$ for each example. The per-example estimates were averaged to obtain the final estimate $\hat{\theta}$, from which the mean $\hat{\mu} = \mu(\hat{\theta})$ and median $\hat{\nu} = \nu(\hat{\theta})$ of the generative distribution were calculated.

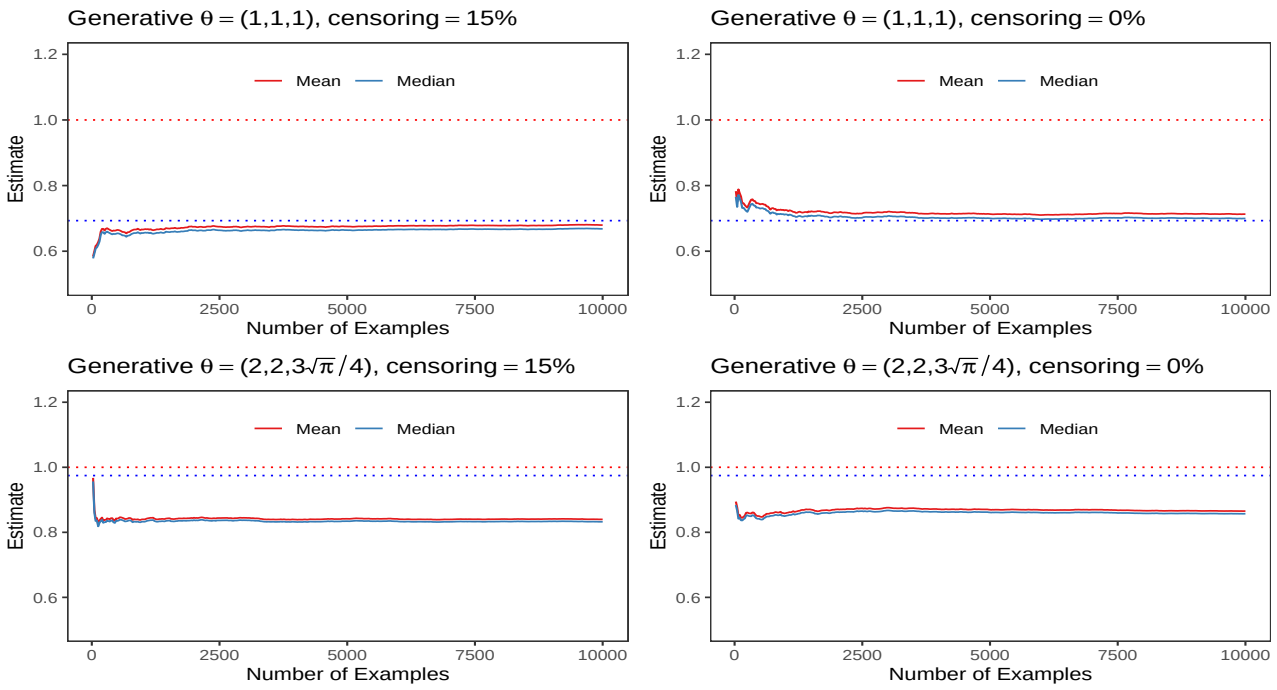

Figure 3.2: Estimated mean and median of the generative distribution as a function of the number of examples. The dotted red and blue lines demarcate the true mean and median. The solid lines are estimates.

[3]The median lacks a closed form but may be found numerically.

Figure (3.2) plots the estimated mean and median of the generative distribution against the number of examples provided. Several trends are noteworthy. First, the estimated mean and median are very close even when the true mean and median are markedly different, as in the case of θ_1. Second, the estimated mean and median stabilize after relatively few $< 10^2$ examples. Third, the estimated mean and median are not yet highly accurate. For θ_1, the relative estimation error was -32% for the mean, but only -3.6% for the median. For θ_2, the relative estimation error was -16% for the mean and -15% for the median.

Regarding point 3, we hypothesized that censoring might be responsible for the model's inaccurate estimation of μ and ν. To test this hypothesis, all training, validation, and evaluation data were re-simulated, starting from the same seed, but without censoring, then mean and median estimation were repeated. Contrary to expectations, censoring had negligible impact on either the MAE (5.7) or the estimated mean and median (3.2), suggesting the model had properly learned to account for it. Rather, from figure (5.8), it appears the major obstacle to accurate estimation of the generative mean/median was inaccurate estimation of the individual components of θ, particularly β.

Discussion

This project attempted to learn a mapping from a potentially right-censored observation time (U, δ) to the parameters $\theta = (\alpha, \beta, \lambda)$ of the generative distribution. The problem may be viewed as attempting to identify the distribution of a particular form most likely to have produced a given example. In general, such a problem may seem under-determined. Given a single (U, δ), there are infinitely many generalized gamma distributions from which it could have arisen. However, the example is *not* equally likely under all these distributions, and the principle of maximum likelihood (ML) asserts that the distribution which makes the example most likely should be selected.

The overarching hypothesis of this project was that, given many examples of (U, δ) and the corresponding generative parameters θ, a deep learning model should be able to learn a mapping from a single instance of (U, δ) to the most likely θ. Figure (5.8) indicates that while this aim has not yet been achieved, some degree of learning did take place, as evidenced by the fact that the estimated $(\hat{\alpha}, \hat{\beta}, \hat{\lambda})$ tended to increase with the generative values. If the accuracy can be refined, a key advantage of the deep learning-based approach is that, whereas ML-based estimation often produces non-sensible $\hat{\theta}$ if initialized poorly, the deep model trained here always provided reasonable estimates, if sometimes inaccurate.

The presence of censoring could not account for failure to learn the mapping $(U, \delta) \mapsto \theta$ because MAE did not markedly improve when censoring was removed. A potential explanation is that, although the model was trained on a sizable volume of data ($n_{\text{train}} = 10^5$), the training data were insufficiently diverse, containing only 10^2 distinct values of θ. To evaluate this hypothesis, the next step would be to simulate data from a greater number of distinct θ, retain the model, then look for an improvement in evaluation set MAE relative to the current results.

Appendix

5.1 Model Architecture

5.1.1 Loss

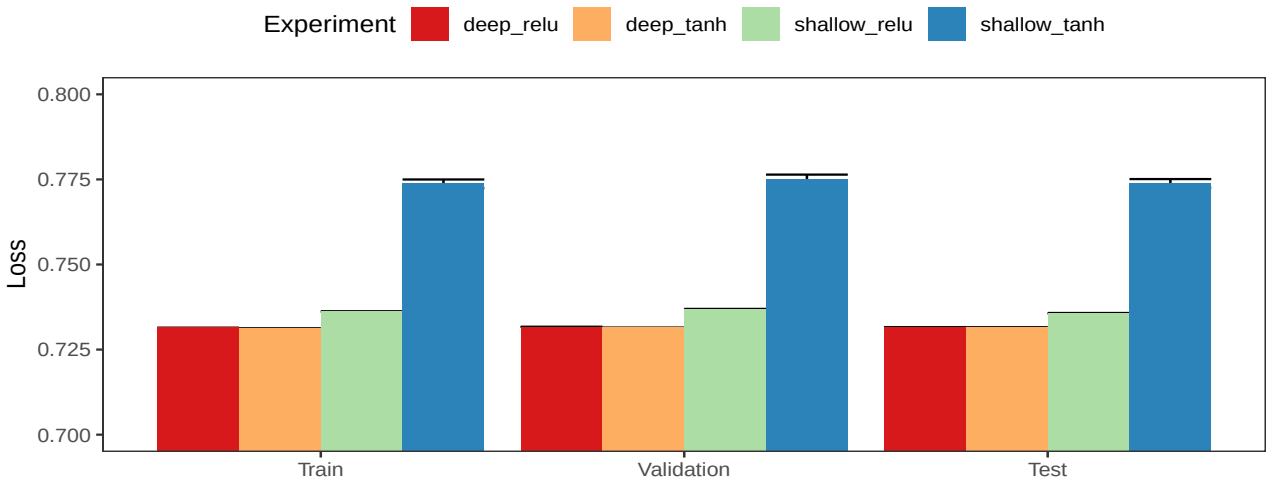

Figure 5.3: Training, validation, and testing loss across different model architectures. The loss is total MAE across the three elements of θ (on log scale).

5.1.2 MAE

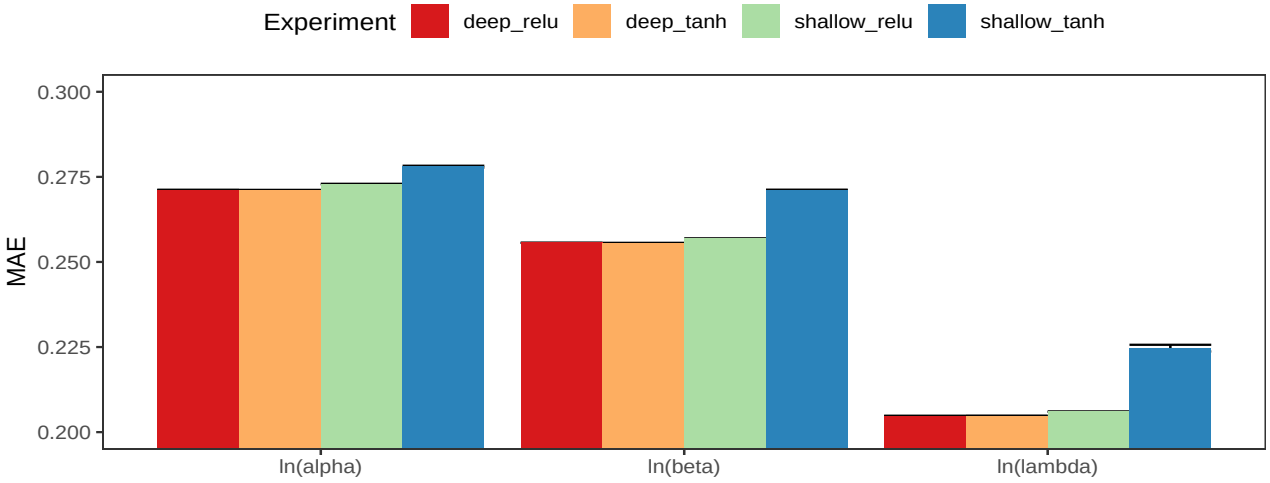

Figure 5.4: MAE by the component of θ across different model architectures.

5.1.3 Estimation Curves

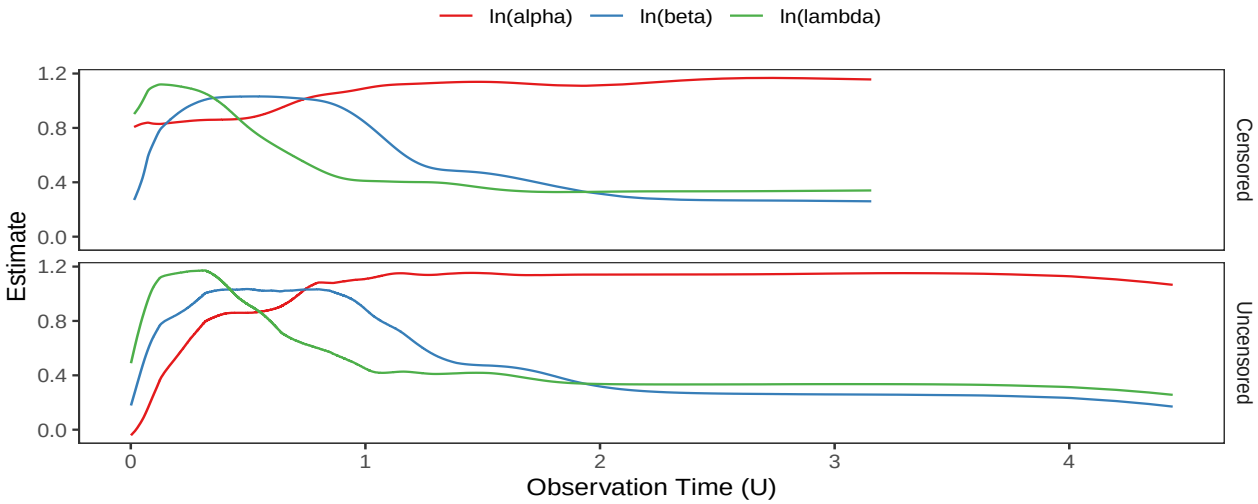

Figure 5.5: Model output by observation time U and censoring status δ.

5.2 Choice of Loss Function

5.2.1 Asymptotic Expansion of the Incomplete Gamma Function

The survival function of the generalized gamma $S_\theta(t) = \Gamma\{\alpha, (\lambda t)^\beta\}/\Gamma(\alpha)$ makes use of the upper incomplete gamma function:

$$\Gamma\{\alpha, (\lambda t)^\beta\} = \int_{(\lambda t)^\beta}^{\infty} s^{\alpha-1} e^{-s} ds.$$

For increasing values of the lower limit $(\lambda t)^\beta$, $\Gamma\{\alpha, (\lambda t)^\beta\} \to 0$, and $\ln S_\theta(t) \to -\infty$. To prevent the right-censored log likelihood (3.1) from diverging, the following heuristic was employed. The lower limit $z = (\lambda t)^\beta$ was first evaluated. If $z < 1$, then the upper incomplete gamma was safely evaluated using the built-in Tensorflow function `tf.math.igammac(·,·)`. For $z \geq 1$, the incomplete gamma function was approximated using the asymptotic expansion:

$$\ln \Gamma(\alpha, z) = (\alpha - 1) \ln z - z \left\{ 1 + \frac{\alpha - 1}{z} + \mathcal{O}(z^{-2}) \right\}.$$

The asymptotic expansion becomes increasingly accurate for $z \to \infty$, which is precisely the domain in which numeric evaluation of $\Gamma(\alpha, z)$ becomes unstable.

5.2.2 Loss by ω

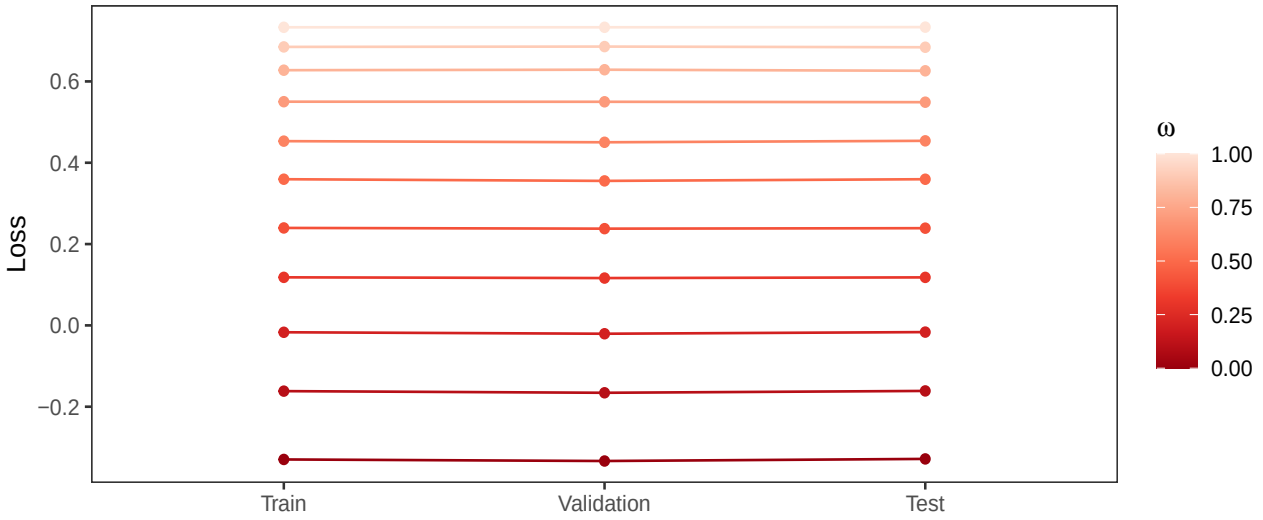

Figure 5.6: Training, validation, and testing loss across deep tanh models trained with different loss functions. ω refers to the weight given to the MAE term in (3.2). The curves should not be compared to one another, as the loss functions differ. Rather, the consistency of a given loss across data sets should be noted.

5.3 Practical Evaluation

5.3.1 MAE by Censoring

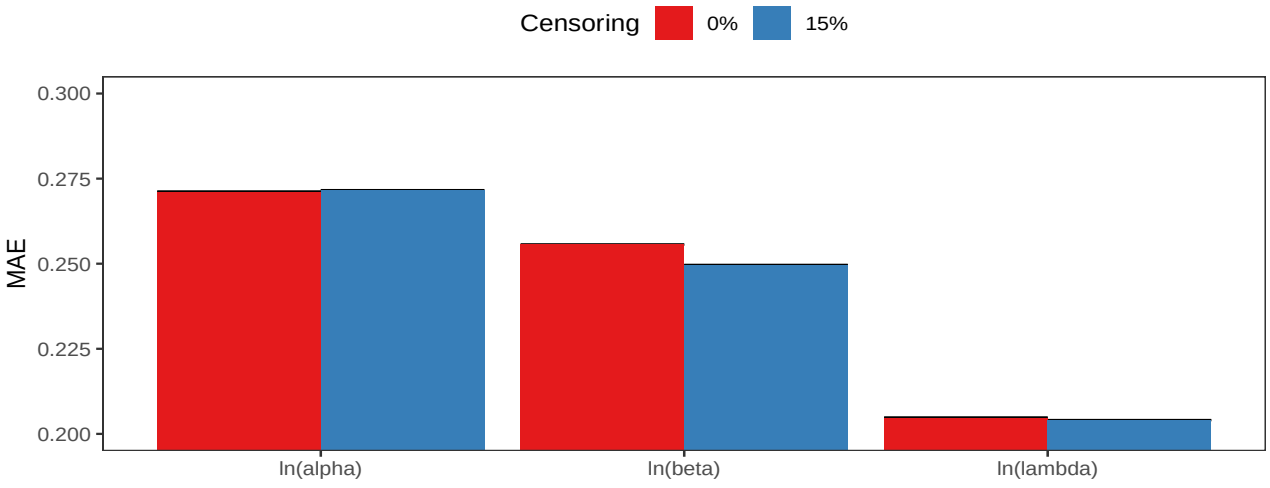

Figure 5.7: MAE by the component of θ in the presence and absence of censoring, using the deep tanh architecture.

5.3.2 Estimated vs. Generative Parameter Values

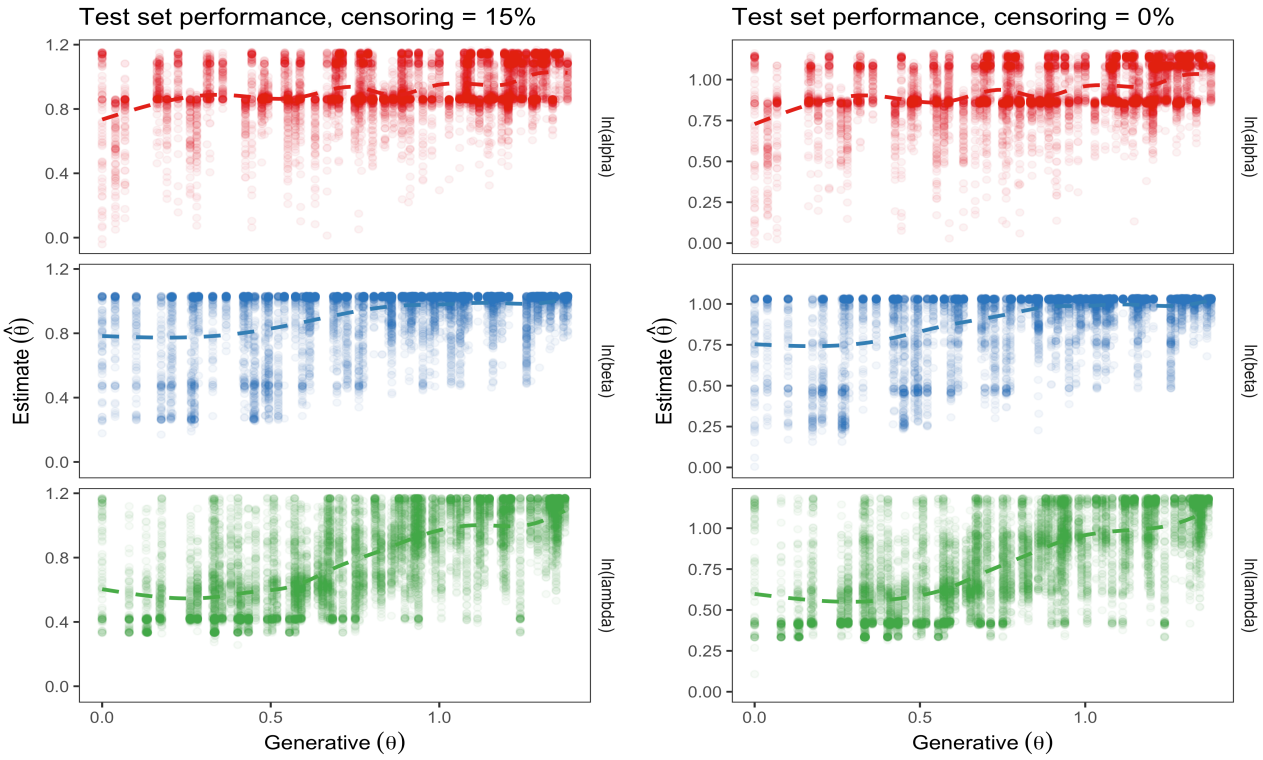

Figure 5.8: Estimated $\hat{\theta}$ vs. generative θ for the best-fitting deep tanh model across the testing data set in the presence vs. absence of censoring.

References

[1] Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-trial-endpoints-approval-cancer-drugs-and-biologics. Accessed: 2021-10-01.

[2] P Guyot, AE Ades, MJ Ouwens, and NJ Welton. Enhanced Secondary Analysis of Survival Data: Reconstructing the Data from Published Kaplan-Meier Survival Curves. *BMC Medical Research Methodology*, 12(9), 2012.

[3] H Luo, J Lu, , Y Bai, et al. Effect of Camrelizumab vs Placebo Added to Chemotherapy on Survival and Progression-Free Survival in Patients With Advanced or Metastatic Esophageal Squamous Cell Carcinoma: The ESCORT-1st Randomized Clinical Trial. *Journal of the American Medical Association*, 326(10):916–925, 2021.

[4] EW Stacy. A Generalization of the Gamma Distribution. *Annals of Mathematical Statistics*, 33(3):1187 – 1192, 1962.

[5] TM Therneau and PM Grambsch. *Modeling Survival Data*. Springer, 2000.