GPSRL: Learning Semi-Parametric Bayesian Survival Rule Lists from Heterogeneous Patient Data

Ameer Hamza Shakur
Dept. of Industrial and Systems Engineering
University of Washington
Seattle, USA
ameerhshakur@gmail.com

Xiaoning Qian
Dept. of Electrical and Computer Engineering
Texas A&M University
College Station, USA
xqian@tamu.edu

Zhangyang Wang
Dept. of Computer Science and Engineering
Texas A&M University
College Station, USA
atlaswang@tamu.edu

Bobak Mortazavi
Dept. of Computer Science and Engineering
Texas A&M University
College Station, USA
bobakm@tamu.edu

Shuai Huang
Dept. of Industrial and Systems Engineering
University of Washington
Seattle, USA
shuaih@uw.edu

***Abstract*—Survival data is often collected in medical applications from a heterogeneous population of patients. While in the past, popular survival models focused on modeling the average effect of the covariates on survival outcomes, rapidly advancing sensing and information technologies have provided opportunities to further model the heterogeneity of the population as well as the non-linearity of the survival risk. With this motivation, we propose a new semi-parametric Bayesian Survival Rule List model in this paper. Our model derives a rule-based decision-making approach, while within the regime defined by each rule, survival risk is modelled via a Gaussian process latent variable model. Markov Chain Monte Carlo with a nested Laplace approximation on the Gaussian process posterior is used to search over the posterior of the rule lists efficiently. The use of ordered rule lists enables us to model heterogeneity while keeping the model complexity in check. Performance evaluations on a synthetic heterogeneous survival dataset and a real world sepsis survival dataset demonstrate the effectiveness of our model.**

I. INTRODUCTION

Survival analysis studies time-to-event data, and consists of important prognostic models to analyze patient morbidity and mortality in almost all medical areas. Many of these models were developed decades ago when medical data were largely collected on paper and were designed for modeling the average effect of risk factors in a population. One of the most commonly used models, the Cox proportional hazards regression [1], was built on a creative nonparametric construct of the baseline hazards function and the use of linear formalism to characterize the relationship between the covariates and the survival outcome, i.e. hazard or risk. With increasing complexity in modern patient medical data that are now available through advances in sensor technologies, the methodological framework of the classic Cox regression model is found to be over-simplified due to its proportional hazards assumption and the imposed linearity of covariate effects. It is possible now to develop better methods for patient survival data analysis by modeling both complex survival effects as well as data heterogeneity. As these two interrelated issues are tackled in the literature as separate issues, in this paper, we propose to tackle both issues in a unified framework that builds on the strength of rule-based learning and semi-parametric Bayesian modeling for survival data. A popular way of modelling survival data is the nonparametric Kaplan-Meier estimator [2] that estimates the survival function from censored and event times but does not incorporate covariate effects. The semi-parametric Cox regression [1] can overcome this problem, but it imposes a strict proportional hazards assumption which is often not valid on real world survival datasets. To ameliorate this limitation, the fully parametric accelerated failure time (AFT) method was proposed where a prior baseline probability distribution is given for the baseline lifetime, and covariate effects directly act on event-times through a link function. Both these methods model linear effects, and can only detect interactions if explicitly specified in the model. On the other hand, the survival trees [3] and random survival forests [4] are non-parametric methods that can implicitly detect interactions and do not assume linearity. Nevertheless, being non-parametric, they are unable to incorporate prior information or quantify uncertainty. Recently, there has been work on using Bayesian semi-parametric methods for survival analysis. Gaussian process extensions to the Cox and AFT models [5], [6] have been introduced, where a latent Gaussian

process is used to model one or more parameters. As GP models define distributions over functions, such models are capable of modelling non-linear effects. Moreover, since they are semi-parametric, they are capable of incorporating prior information, while the Bayesian approach enables them to quantify uncertainty. However, these models aim to model the survival data from a homogeneous population and do not model heterogeneous effects.

Rule-based learning uses spatial partitioning to model heterogeneous data. Rule models have been especially important in medical and healthcare domains where interpretability is critical, e.g., rule based machine learning models have been used for diagnosis of breast [7] and lung cancer [8], sepsis [9], diabetes [10] as well as to study depression profiles [11]. Rule-based methods have also been shown to be effective in identifying subgroups with heterogeneous risk profiles in a patient population [12]. Greedy decision tree models such as CART [13] are typical examples, but they provide a quite restrictive result, i.e., the partitioning is suboptimal. Optimal partitioning such as through integer programming [14] or Bayesian decision trees [15] is NP-hard and computationally demanding due to the exponential search space which severely limits both the depth of the tree and the number of variables that can be considered. Comparing with the tree models, the rule-based methods, which work on the same principle of partitioning, have found a larger flexibility and efficiency in a range of applications. As the purpose of using partitioning is to tackle patient heterogeneity, it can be used to group data into subsets with similar response characteristics that enable subgrouping of subjects and subsequent separate modeling of each subgroup. A recent breakthrough in rule learning is Rulefit [16] for classification and regression modelling, motivated by the sparse regularization techniques, which was generalized to survival outcomes as well [17]. Rulefit generates a sparse list of predictive rules from a large set of rules mined from bootsrapped decision trees. Rules can also be flexibly used or re-organized to make better decisions, one example is the recent development of Bayesian rule lists (BRL) for classification [18] that is able to incorporate Bayesian modeling to quantify uncertainty. BRL is able to strike a balance between greedy and optimal partitioning and provides good generalizability with significantly lower computational load.

In this paper, we propose an integrative framework that uses ordered rule lists to derive a rule-based decision-making approach, while within the regime defined by each rule, survival risk is modelled via a Gaussian process latent variable model. The computational challenges are overcome by a tailored Markov Chain Monte Carlo algorithm with a nested Laplace approximation for the latent variable model to search over the posterior of the rule lists efficiently. The use of ordered rule lists enables our method to model data heterogeneity while simultaneously keeping the model complexity low and providing interpretability. Moreover, since basic GP survival models require $O(N^3)$ matrix inversion operations, the data partitioning approach may lower computational demands once a rule list is found.

The remainder of the paper is organized as follows: Section II summarizes the related works; the proposed method is described in Section III; experimental results are presented in Section IV; finally, conclusions and future works are summarized in Section V. Note that, in this paper, we use lower or upper case letters, e.g., x or X, to represent scalars, bold-face lower-case letters, e.g., $\mathbf{x}$, to represent vectors, bold-face upper-case letters, e.g., $\mathbf{X}$, to represent matrices.

II. RELATED WORK

Survival analysis analyzes time-to-event data, which are typically represented in the form $\{(t_i, \delta_i, \mathbf{x}_i) \mid \mathbf{x}_i \in \mathbb{R}^P, i \in 1:N\}$. Here, t_i is the event time or censored time, $\mathbf{x}_i$ are covariates, and δ_i denotes whether or not the event has occurred for an observation. Survival data are often incomplete and the event time may be either right, left or interval censored. The objective of survival analysis is to model the *survival* function and the *hazard* function and understand how the covariates impact these functions. Assuming that the time-to-event (failure time), T, is a continuous random variable with a density function, $f(t)$, the *survival* function, $S(t)$, which is the probability that the event has not yet occurred by time t is,

$$S(t) = \Pr\{T \geq t\} = \int_t^\infty f(x)dx. \tag{1}$$

Further, the *hazard* function, $h(t)$ is the instantaneous rate of the occurrence of the event at time t,

$$h(t) = \lim_{dt \to 0} \frac{\Pr\{t \leq T < t + dt \mid T \geq t\}}{dt} = \frac{f(t)}{S(t)} \tag{2}$$

The data likelihood can be derived through these functions as:

$$L = \prod_{i=1}^{K:\delta=1} f(t_i) \prod_{j=1}^{M:\delta=0} S(t_i) = \prod_{i=1}^{N} h(t_i)^{\delta_i} S(t_i). \tag{3}$$

A. Generalized Linear Models for Survival Analysis

Different approaches aim to model the hazard function based on different premises. In the Cox proportional hazards regression [1], the hazard function is modelled as the product of a nonparametric baseline hazard, $h_0(t)$ which is a function of time and a relative hazard term, $h_R(\mathbf{x})$ that is a log linear function of covariates,

$$h(t \mid \mathbf{x}_i) = h_0(t)\exp(g(\mathbf{x}_i)), \tag{4}$$

where $g(\mathbf{x}_i) = \mathbf{w}'\mathbf{x}_i$. The Cox model makes an assumption that the hazards are proportional, i.e., the ratio of hazards of any two observations remains constant over time and only depends on the covariates. This restrictive assumption often does not hold on real survival datasets. Further, since the baseline hazard, $h_0(t)$ is not estimated, the Cox model can only describe how the hazards of two observations relate with each other, and not describe the hazard or survival function

of a given observation directly. The accelerated failure time (AFT) model is a popular parametric approach that relaxes the proportional hazards assumption by modelling the covariate effects as directly influencing the failure time of the observations,

$$\log T_i = g(\mathbf{x}_i) + \epsilon, \tag{5}$$

where ϵ is the error term with a specified distribution that determines the baseline failure density. A common modelling approach is to assign a logistic distribution to ϵ, which is equivalent to assigning a log-logistic (*LL*) distribution to the baseline failure time, T_0, as we can see by rewriting (5) as, $T_i = \exp(g(\mathbf{x}_i))T_0$ and $T_0 = \exp(\epsilon)$. In this case, the failure time of an observation, T_i can therefore be seen as a sample from a log-logistic distribution with scale parameter, $\alpha_i = \exp(g(\mathbf{x}_i))$ depending on covariates and a common shape parameter, β for all observations, i.e. $T_i \sim LL(\alpha_i, \beta)$, where $\alpha_i = \exp(\mathbf{w}'\mathbf{x}_i)$. The parameters, $(\mathbf{w})$ of the Cox regression or $(\mathbf{w}, \beta)$ of the AFT model are estimated through Maximum Likelihood estimation. These two models fall under the framework of generalized linear model as the covariate effects are assumed to be linear, therefore, they cannot deal with non-linearities and interaction effects unless explicitly specified in the model terms. A popular approach to deal with some of these limitations are survival trees [3] and the bootstrap aggregation of trees, i.e., the random survival forest [4] which can implicitly deal with interactions. An added advantage of the tree based approaches is that they are interpretable and can easily be decomposed into rules.

B. Gaussian Processes for Survival Analysis

To ameliorate the limitations of the generalized linear model framework, which can only include nonlinear effects through explicit model specification, a well known approach has been to use Gaussian processes. Several works in literature have applied Gaussian processes to extend the standard models of survival analysis. The GP-Cox model [5] extends the Cox regression by the replacing linear effects term, $g(\mathbf{x})$ with a GP, $\mathbf{f}$ to model nonlinear covariate effects while assuming the baseline hazard to be piecewise constant. This model is extended in [6] by smoothing the piecewise constant baseline hazard with a second GP. Recently, the Gaussian Process framework was used to extend the AFT model [19] to nonlinear effects. We discuss this GP-AFT model in further detail since it is related to our work.

1) GP-AFT survival model: The GP extension to the log-logistic AFT model (Section II-A) includes nonlinear effects by imposing a log GP prior on the scale parameter (α), instead of a log linear relation to the variables, i.e., $\boldsymbol{\alpha} = \exp(\mathbf{f})$ where $\mathbf{f} \sim \mathcal{GP}(0, \mathbf{K})$. GP-AFT models the scale of the AFT distribution as dependent on the variables through the GP, while the shape parameter, β, is considered to be the same for all observations, and does not depend on covariates. However, this model carries over the AFT assumption that the shape parameter of the failure time density does not change with respect to the covariates. This is restrictive as it assumes the hazard function to be either exponential or unimodal for all observations in the data, which does not hold in heterogeneous datasets.

Work to ameliorate the limitations of Gaussian processes on heterogeneous data have focused on partitioning approaches. [20] propose Bayesian treed partitioning, with GP being fit to the terminal nodes of a Bayesian CART model, while [21] fit separate GP's at each element of a voronoi tessellation. However, these efforts still use exclusive tree structures for modeling heterogeneity, and do not address nonlinear survival models such as GP-AFT. In contrast to these fully probabilistic GP models, our work is semi-parametric which enables us to achieve a balance between computational demand and performance. Our proposed work also addresses the limitation of the GP-AFT survival to heterogeneity by relaxing the common shape assumption, and varying the shape parameter with respect to covariates, i.e., our rule list approach allows us to learn failure time densities with varying shapes for different partitions, as well as identify the covariates that cause heterogeneity.

III. METHODS

Let $\mathcal{R}$ be the pre-mined rule set containing a total of K rules. We generate $\mathcal{R}$ by extracting rules of various cardinalities from trees in a random survival forest. The cardinality of a rule is defined as the number of interacting covariates in the rule (ex. $1, 2, ..$ etc.). Rules that are endorsed by at least a given minimum number of observations in the dataset are selected, i.e., rules that apply to very few observations are filtered out. Our goal is to tackle heterogeneity by building an *ordered* rule list, d which is a subset of $\mathcal{R}$ of size m where $m \ll K$. Priors on the number of rules in the list, m and the number of covariates interacting in each rule ensure that the rule list is sufficiently sparse and complex. We utilize an MCMC scheme similar to that in BRL [18] to obtain a posterior over the ordered rule list given data; however, the objective in BRL was multivariate classification, while our goal is survival analysis with Gaussian processes. In what follows, we describe our predictive model, Gaussian process survival rule lists (GPSRL), and inference procedure to learn the model.

A. Formulation of Gaussian Process Survival Rule Lists

An ordered rule list, d with m rules will divide the dataset into $m + 1$ non overlapping partitions as follows: each observation in the dataset which endorses at least one of the m rules belongs to the partition associated with the *first* rule in the ordered list, d, that the observation endorses. Observations not endorsing any of the m rules will belong to the $m+1$-th partition. Thus, an ordered rule list of length m will divide the data into $m + 1$ exclusive partitions. For each of the partitions determined by d, we fit a log-logistic (LL) Gaussian process AFT model (II-B1) with its scale parameter $(\boldsymbol{\alpha})$ dependent on covariates and modelled via a Gaussian process, and a shared shape (β) parameter with

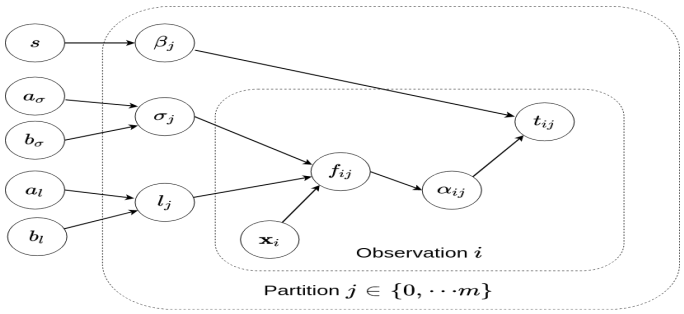


Fig. 1: Caption

a log uniform prior. The same priors are adopted across partitions to control the model complexity. An illustration of our Bayesian GPSRL model is given in (6).

$$\begin{aligned} \textit{if } r_1 \textit{ then } \mathbf{t_1} &\sim LL(\boldsymbol{\alpha_1}, \beta_1) \\ \textit{else if } r_2 \textit{ then } \mathbf{t_2} &\sim LL(\boldsymbol{\alpha_2}, \beta_2) \\ &\cdot \\ &\cdot \\ \textit{else if } r_m \textit{ then } \mathbf{t_m} &\sim LL(\boldsymbol{\alpha_m}, \beta_m) \\ \textit{else } \mathbf{t_0} &\sim LL(\boldsymbol{\alpha_0}, \beta_0) \end{aligned} \tag{6}$$

The priors on the parameters of the rule list and survival models on each partition ($j \in \{0, \cdots m\}$) are:

$$\begin{aligned} m &\sim TP(\lambda, 0, |R|) & \boldsymbol{\alpha_j} &= \exp\big(\mathbf{f}_j(\mathbf{X}_j)\big) \\ \log\, \beta_j &\sim U(0, s) & \mathbf{f}_j(\mathbf{X}_j) &\sim \mathcal{GP}(0, \mathbf{K}_j) \\ s &\sim IG(a, b) & \mathbf{K}_j(\mathbf{x}, \mathbf{x}) &= \sigma_j^2 \exp\big(-\frac{|\mathbf{x} - \mathbf{x}'|^2}{2l_j^2}\big) \\ \sigma_j &\sim IG(a_\sigma, b_\sigma) & l_j &\sim IG(a_l, b_l) \end{aligned} \tag{7}$$

Here, LL, U, TP, IG, N are the Log-logistic, Uniform, Truncated-Poisson, Inverse-Gamma and Normal distributions, respectively. Our model seeks to combine the interpretability of rule-based models with the modelling flexibility of Gaussian process survival models. Given the pre-mined set of rules $\mathcal{R}$, we seek to obtain the posterior distribution of the ordered Bayesian rule list, d, and the associated posterior distributions of parameters of the Gaussian survival processes that model the survival response at each of the corresponding partitions defined by d.

B. Bayesian inference

Given covariate data, $\mathbf{X}$, and survival response, $\mathbf{y}$, Our goal is to obtain the posterior distribution of the ordered Bayesian rule list. The posterior probability density of the rule list, d is proportional to the product of data likelihood and prior probability:

$$p(d \mid \mathbf{X}, \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{X}, d) p(d). \tag{8}$$

1) Prior probability: Similar to the prior probability of the BRL [18] model, the prior probability of the GPSRL rule list is defined hierarchically as,

$$p(d) = p(m \mid \lambda) \prod_{j=1}^{m} p(c_j \mid c_1 \cdots c_{j-1}, \eta) p(r_j \mid r_1, \cdots r_{j-1}, c_j). \tag{9}$$

Here m is the number of rules in the list, and c_j denotes the cardinality of rule r_j. Truncated-Poisson (TP) priors are selected for both m and each of $c_j \mid c_1 \cdots c_{j-1}$. The TP prior on m has a given mean value λ and is truncated on the total number of rules that are available, K. It may be written as follows.

$$p(m \mid \mathcal{R}, \lambda) = \frac{(\lambda^m / m!)}{\sum_{j=1}^{K} (\lambda^j / j!)}, \quad m = 0, 1, \ldots K \tag{10}$$

The TP prior on $c_j \mid c_1 \cdots c_{j-1}$ has a mean η and is truncated to only include the cardinalities of rules that are currently available. $\mathcal{R}_j = \mathcal{R} \setminus \{r_1, r_2 \ldots r_{j-1}\}$ are the rules available after sampling $j-1$ rules. The probability of selecting the cardinality, c_j may be written as:

$$p(c_j \mid c_1 \cdots c_{j-1}, \eta) = \frac{\eta^{c_j} / c_j!}{\sum_{c \in C_j} \eta^c / c!}, \tag{11}$$

where C_j is the set of cardinalities of rules in $\mathcal{R}_j$. Once c_j is sampled, a uniform probability is chosen over all the available rules with cardinality c_j to sample the j-th rule in d, r_j. That is,

$$p(r_j \mid r_1, \cdots r_{j-1}, c_j) = \frac{1}{|\{r_i \mid r_i \in \mathcal{R}_j, c_i = c_j\}|} \tag{12}$$

The first term in the prior (9) is the probability of obtaining a rule list with m rules given the mean number of rules λ. In the subsequent product over the rules $\{r_1 \cdots r_m\}$ in d, each term denotes the probability of obtaining a rule r_j with a cardinality of c_j given a mean cardinality η multiplied by the probability of choosing rule r_j from all the available rules with this cardinality.

2) Likelihood: We fit a log-logistic (LL) Gaussian process AFT model (II-B1) at each of the partitions defined by the rule list, d. Here, the data likelihood (3) of partition $j \in \{1, 2, \ldots m+1\}$ given response $\mathbf{y} = (\mathbf{t}, \boldsymbol{\delta})$, covariate data, $\mathbf{X}$, and parameters $(\boldsymbol{\alpha}, \beta)$ is as follows:

$$p(\mathbf{y} \mid \boldsymbol{\alpha}, \beta) = \prod_{i=1}^{M:\boldsymbol{\delta}=1} \frac{(\frac{\boldsymbol{\alpha}_i}{\beta})(\frac{\mathbf{y}_i}{\boldsymbol{\alpha}_i})^{\beta-1}}{(1 + \frac{\mathbf{y}_i}{\boldsymbol{\alpha}_i}\beta)^2} \prod_{j=1}^{N:\boldsymbol{\delta_j}=0} \frac{1}{1 + (\frac{\mathbf{y}_i}{\boldsymbol{\alpha}_i})^\beta}. \tag{13}$$

The scale parameter, $\boldsymbol{\alpha} = \exp(\mathbf{f})$ is defined as an exponential of a Gaussian process (GP) with prior $\mathbf{f} \mid \mathbf{X} \sim \mathcal{GP}(0, \mathbf{K_X})$, while the shape parameter has a log uniform prior, $\log\, \beta \sim U(0, s)$. Since this likelihood (13) is not Gaussian or conjugate-Gaussian, the posterior density of $\mathbf{f}$, i.e., $p(\mathbf{f} \mid \mathbf{X}, \mathbf{y})$, and consequently the marginal data likelihood, $p(\mathbf{y} \mid \mathbf{X}) = \int p(\mathbf{y} \mid \mathbf{f}) p(\mathbf{f} \mid \mathbf{X}) d\mathbf{f}$ is not analytically tractable and must be approximated. A popular method to approximate

the posterior in case of non-Gaussian likelihood with latent Gaussian processes such as those arising in survival analysis is the Laplace approximation, which obtains a Gaussian distribution approximation to the posterior density of the GP around the mode of the true distribution. The Laplace approximation [22] obtains a Gaussian density, q, which approximates the true non-Gaussian posterior density of the GP, i.e., the approximate posterior density, $q(\mathbf{f}) \approx p(\mathbf{f} \mid \mathbf{y}, \mathbf{X})$ given by,

$$q(\mathbf{f}) = \mathcal{N}(\mathbf{f} \mid \hat{\mathbf{f}}, \mathbf{A}^{-1}) \tag{14}$$

where $\hat{\mathbf{f}} = \text{argmax}_{\mathbf{f}} p(\mathbf{f} \mid \mathbf{X}, \mathbf{y})$ is the mode of the posterior and $\mathbf{A} = -\nabla\nabla \log\, p(\mathbf{f} \mid \mathbf{X}, \mathbf{y})$ is the Hessian of the negative log posterior at the mode. This approximate density can be used in lieu of the true GP posterior to calculate an approximated marginal likelihood, $p(\mathbf{y} \mid \mathbf{X}) \approx p_L(\mathbf{y} \mid \mathbf{X}) = \int p(\mathbf{y} \mid \mathbf{f}) q(\mathbf{f}) d\mathbf{f}$. The total approximated marginal likelihood of the data can be written as:

$$p_L(\mathbf{y} \mid \mathbf{X}, d) = \prod_{j=0}^{m} p_L(\mathbf{y}_j \mid \mathbf{X}_j), \tag{15}$$

where $\mathbf{y}_j$ and $\mathbf{X}_j$ is the response and covariate data, respectively, belonging to each of the $m+1$ partitions (indexed by j) defined by our rule list.

3) MCMC Sampling: We use Metropolis-Hastings sampling to infer the posterior distribution of the rule list, $p(d \mid \mathbf{X}, \mathbf{y})$. The sampling sequence of rule lists starts with an initial random list, d^0 sampled from the prior, $p(d)$. After initialization, the sequence proceeds as follows: At step t in the sequence, with a rule list d^t of length m^t, a proposal distribution Q is used to propose the next list in the sequence $d^{t+1} \sim Q(d^t)$. The new rule list is generated through one of three equally likely operations: i) adding a rule to the bottom of d^t, ii) removing a randomly selected rule from d^t, iii) moving a randomly selected rule to a different position in d^t. The proposal distribution denotes the probability of sampling d^{t+1} from d^t,

$$Q(d^{t+1} \mid d^t, \mathcal{R}) = \begin{cases} \frac{1}{(K-m^t)(m^t+1)} & \text{if a rule is added} \\ \frac{1}{m^t} & \text{if a rule is removed} \\ \frac{1}{m^t(m^t-1)} & \text{if a rule is moved.} \end{cases} \tag{16}$$

The proposed sequence, d^{t+1} is then accepted with an acceptance probability, $\pi(d^{t+1} \mid d^t)$, defined as follows:

$$\pi(d^{t+1} \mid d^t) = \min\Big\{ \frac{Q(d^{t+1}, d^t)}{Q(d^t, d^{t+1})} \frac{p(\mathbf{y} \mid \mathbf{X}, d^{t+1}) p(d^{t+1})}{p(\mathbf{y} \mid \mathbf{X}, d^t) p(d^t)}, 1 \Big\}. \tag{17}$$

Here $p(\mathbf{y} \mid \mathbf{X}, d^t)$ and $p(\mathbf{y} \mid \mathbf{X}, d^{t+1})$ are marginals that may be evaluated approximately as shown in Section III-B2. For a sufficiently long chain, the sequence will sample rule lists from the posterior density of the Bayesian rule list. Gaussian approximates to the marginal likelihood have been proposed in literature to increase the speed of the MCMC algorithm when the likelihood evaluation is costly [23], where it was shown that using an approximate likelihood may take more MCMC steps to reach convergence though the total time of convergence reduces due to faster sampling enabled by the approximations. However, using approximate likelihood evaluations does not theoretically guarantee convergence of the MCMC algorithm though the algorithm will push the sequence towards an area with a high approximate posterior.

C. Predictive Inference

Given the posterior distribution of the rule lists obtained from the MCMC sequence, $p(d \mid \mathbf{X}, \mathbf{y})$, we can obtain a point estimate of the rule list, d, and the model defined by the Laplace-posterior approximations of the latent GP, $q_j(\mathbf{f})$ and shape parameter of the log likelihood distribution, β_j at each of the partitions. The predictive density (under the Laplace approximation) of a new observation $(y^*, \mathbf{x}^*)$ which falls in, say, the j-th partition as defined by d can be evaluated as follows: first, the distribution of the latent GP at the new observation (see [22] for derivation) under the Laplace approximation is computed. Since,

$$\begin{bmatrix} f^* \\ \mathbf{f} \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \mathbf{K}(\mathbf{X}) & \mathbf{K}(\mathbf{X}, \mathbf{x}^*) \\ \mathbf{K}(\mathbf{x}^*, \mathbf{X}) & \mathbf{K}(\mathbf{x}^*) \end{bmatrix}\right) \tag{18}$$

and therefore the conditional density is

$$f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{f} = \mathcal{N}\Big(\mathbf{K}(\mathbf{x}^*, \mathbf{X})\mathbf{K}(\mathbf{X})^{-1}\mathbf{f}, \\ \mathbf{K}(\mathbf{x}^*) - \mathbf{K}(\mathbf{x}^*, \mathbf{X})\mathbf{K}(\mathbf{X})^{-1}\mathbf{K}(\mathbf{X}, \mathbf{x}^*)\Big). \tag{19}$$

The posterior predictive density of the GP under the Laplace approximation, $q_j(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y})$, can then be evaluated as,

$$\begin{aligned} p_j(f^* \mid \mathbf{X}, \mathbf{y}, \mathbf{f}) &= \int p_j(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{f}) p(\mathbf{f} \mid \mathbf{X}, \mathbf{y}) d\mathbf{f} \\ &\approx \int p(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{f}) q_j(\mathbf{f} \mid \mathbf{X}, \mathbf{y}) d\mathbf{f} \\ &= q_j(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) \end{aligned} \tag{20}$$

and is therefore a Gaussian distribution that is analytically tractable. Then, the predictive density of the observation $(y^*, \mathbf{x}^*)$ under the Laplace approximation is given by the integral,

$$\begin{aligned} p(y^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) &= \int p(y^* \mid \alpha^*, \beta_j) p_j(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) df^* \\ &\approx \int p(y^* \mid \alpha^*, \beta_j) q_j(f^* \mid \mathbf{x}^*, \mathbf{X}, \mathbf{y}) df^* \end{aligned} \tag{21}$$

Since the first term of this integral is the log logistic data likelihood (13), and is not conjugate-Gaussian, the integral is not analytically tractable and must be calculated using numerical methods.

IV. EXPERIMENTS

Performance validation and comparison with several existing GP-based survival models such as GP-AFT survival Gaussian process model with Laplace approximation [24] SGP(L), survival GP model with variational approximation [25] SGP(V), and a recent work that proposed to model both shape and scale parameters with GP's, chained Gaussian process model [26], CHGP. Experiments are performed on a synthetic heterogeneous survival dataset and a real-world survival dataset of sepsis patients from the MIMIC-III (Medical Information Mart for Intensive Care) [27] database. Performance is evaluated using the negative log predictive density (NLPD) (21) and concordance index (C-INDEX) [28]. NLPD is calculated from the predictive density as shown in Section III-C,

$$\text{NLPD}(\mathbf{y}^*, \mathbf{X}^*) = \frac{\sum_{i=1}^{N} p(y_i^* \mid \mathbf{x}_i^*, \mathbf{X}, \mathbf{y})}{N}. \tag{22}$$

C-INDEX is the measure of a model's ability to rank survival times. It estimates the probability that in a randomly selected pair of test observations, the one with the lower response time has the lower predictive response time. In our experiments, for each observation, we take the average C-INDEX calculated over 100 predicted times sampled from the predictive log logistic distributions in the obtained GPSRL rule lists (6). NLPD is lower in a superior model while C-INDEX is higher. We use the GPy [29] software to train the GP models used in these experiments.

A. Synthetic Data

We simulated a heterogeneous survival dataset, $D = (\mathbf{y}, \delta, \mathbf{X})$, consisting of $N = 1000$ observations with $P = 4$ variables. The covariates of each observation $\mathbf{x} \in \mathbf{X}$ are generated by sampling from uniform distribution ($\mathbf{x}_i \sim U(0,1)\ \forall i \in 1 : P$). Event times for all observations, $\mathbf{t}$, are simulated by sampling from a log-logistic distribution, $\mathbf{t}_i \sim LL(\alpha, \beta)$ with the scale α, and shape β parameters generated as follows:

$$\alpha(\mathbf{x}) = I_1\alpha_1(\mathbf{x}) + I_2\alpha_2(\mathbf{x}) + I_3\alpha_3(\mathbf{x}),$$
$$\beta(\mathbf{x}) = I_1\beta_1(\mathbf{x}) + I_2\beta_2(\mathbf{x}) + I_3\beta_3(\mathbf{x}),$$

where I_1, I_2, I_3 are indicator functions to denote certain conditions on the covariate data being satisfied and α_i, β_i are different complex functions of the covariates of the following form:

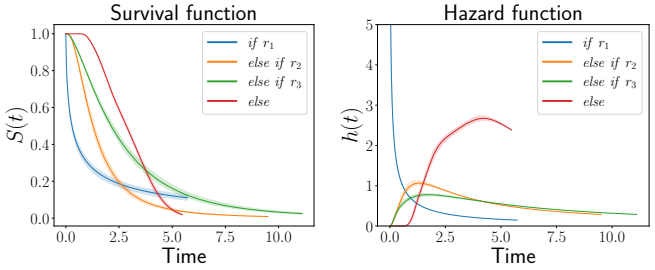
$$\alpha_i(\mathbf{x}) = a_1 \exp\Big(a_2\big(\sum_{k=1}^{2} \exp(a_3(x[k] - a_4)^2)\big) + \sum_{k=3}^{4} \sin(\pi \mathbf{x}[k]^2)\Big),$$

$$\beta_i(\mathbf{x}) = b_1 \exp\Big(\sum_{k=1}^{2} \sin(2\pi \mathbf{x}[k]^2) + \sum_{k=3}^{4} \cos(2\pi \mathbf{x}[k]^2)\Big),$$

with varying values of a_1, a_2, a_3, a_4 and b_1 for each $i \in \{1, 2, 3\}$. We assume that 35% of the simulated data is censored ($\boldsymbol{\delta}_i = 0$). To account for this, the simulated event times, t of a random subset consisting of 35% of the data are multiplied with a uniform random variable to simulate the response times, $\mathbf{y}$, i.e., $\mathbf{y}_i = \rho_i \mathbf{t}_i$ if $\boldsymbol{\delta}_i = 0$ else $\mathbf{y}_i = \mathbf{t}_i$ where $\rho_i \sim U(0,1)$. Performance comparison is carried out by evaluating model performance on a testing dataset consisting of 250 observations that was simulated in a similar manner as the training data. A large set of rules are mined from a survival random forest and rules that are followed by at least 10% of the data are selected to generated the initial rules, $\mathcal{R}$.

TABLE I: Estimate of ordered rule list d from the posterior

Rules	
r_1	$x_3 \leq 0.259$
r_3	$x_4 > 0.596$ & $x_3 > 0.196$
r_3	$x_4 \leq 0.677$ & $x_4 > 0.192$

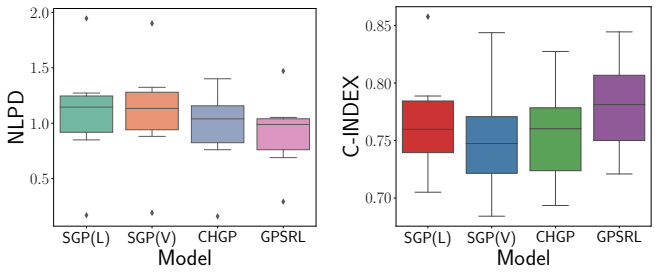


(a) (b)

Fig. 2: (a) Mean survival function and (b) Mean hazard function of GPSRL model on each of the partitions defined by rule list in Table I

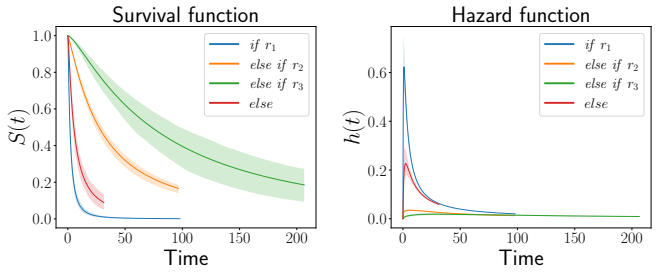
Results: We used hyperparameter values of $\lambda = 3$ for the mean length of the rule list, and $\eta = 2$ for the mean number of variables in each of the rules. The MCMC chain was simulated until convergence, which took approximately 4000 iterations. An example of the rule list obtained from the posterior in one of the MCMC chains is shown in Table I. Fig. 2a shows the mean survival function learnt at each of the four data partitions defined by the rule list, and Fig. 2b shows the mean hazard function. It is noted that the hazard function learnt by GPSRL is multimodal, i.e., in the first partition, the hazard is exponential, signifying that the shape parameter, $\beta <= 1$ while in the other partitions it is unimodal meaning $\beta > 1$. The standard GP-AFT models discussed in Section II assume the same value of β for all observations and hence learn either an exponential or a unimodal hazard but not both, while the partitioning approach of GPSRL allows us to model both unimodal and multimodal hazards, which is a typical aspect of heterogeneous medical datasets.

The performance comparison results in Fig. 3 show that GPSRL model outperforms the other models on the synthetic



(a) (b)

Fig. 3: (a) NLPD and (b) C-INDEX comparison over 10 cross-validation folds with 250 replicates for different survival GP models trained on synthetic data



(a) (b)

Fig. 4: (a) Mean survival function and (b) Mean hazard function of GPSRL model on each of the partitions defined by rule list in Table II

dataset. The box plot in Fig. 3a shows the NLPD obtained by each model on the testing data over 10 folds of the testing data. We observe that GPSRL achieves the lowest values of NLPD as compared to other models, demonstrating its effectiveness. The other survival model that provides a way to model heterogeneity, chained survival Gaussian process performs second best. The GP-AFT models with Laplace approximation, SGP-L and with sparse GP variational approximation, SGP-V achieve a comparable performance. In Fig. 3b, the comparison of C-INDEX for different models on the 10 testing folds is shown. Once again, GPSRL outperforms other models and has the highest mean C-INDEX. The C-INDEX follows the same trend as NLPD, however SGP(V) achieves the lowest C-INDEX.

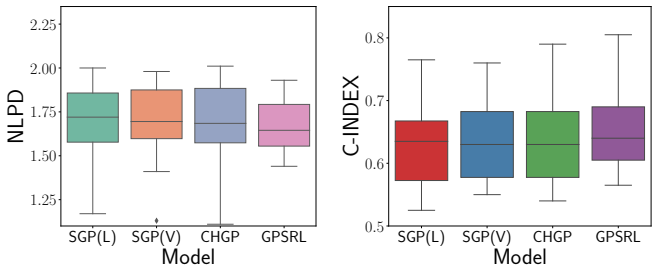
B. Sepsis Data

MIMIC-III is a comprehensive database comprising anonymized information relating to patients admitted to the Beth Israel Deaconess Medical Center in Boston, MA between 2001 and 2012. The data consists of over 53,000 adult ICU admissions during this time period. In this paper, we utilize a subset of inpatient admissions which were diagnosed with sepsis conditions. We consider 9 variables consisting of patient characteristics and physiological measurements, which are age, heart rate, diastolic and systolic blood pressure, saturated oxygen, arterial-pH etc. We choose a subset of 1200 observations for training the model and a testing dataset of 400 observations.

TABLE II: Estimate of ordered rule list d from the posterior

Rules	
r_1	artpH-(mean) <= 7.249
r_2	O2sat-(sd) <= 4.66 & diaBP-(mean) <= 61.26
r_3	O2sat-(sd) <= 4.8

Results: We used hyperparameter values of $\lambda = 3$ for the mean length of the rule list, and $\eta = 2$ for the mean number of variables in each of the rules. The MCMC chain was simulated until convergence, which took approximately 4700 iterations. An example of the rule list obtained from the posterior in one of the MCMC chains is shown in Table II. Fig. 4a shows the mean survival function learnt at each of the three data partitions defined by the rule list, and Fig. 4b shows the mean hazard function. The performance comparison results in Fig. 5 show that on this dataset, performance of all the models is comparable though a small gain is achieved by using GPSRL. The box plot in Fig. 5a shows the NLPD achieved by each model on the testing data over 10 folds of 40 observations each. GPSRL does achieve a slightly lower median NLPD as compared to other models, though standard GP models are satisfactory on this dataset. The same applies to Fig. 5b which compares the C-INDEX achieved by the models on the 10 testing folds. Once again, the values obtained are nearly equal.



(a) (b)

Fig. 5: a) NLPD and (b) C-INDEX comparison over 10 cross-validation folds with 400 replicates for different survival GP models trained on sepsis data.

TABLE III: Summary of NLPD comparison of different models

Data	SGP(L)	SGP(V)	CHGP	GPSRL
Synthetic	1.08 ± 0.2	1.09 ± 0.2	0.93 ± 0.1	0.9 ± 0.1
Sepsis	1.76 ± 0.2	1.75 ± 0.2	1.73 ± 0.2	1.67 ± 0.2

A summary of the mean NLPD obtained via crossvalidation by the various survival GP models on both the synthetic and survival test datasets is provided in Table III, and a summary of the mean C-INDEX is provided in Table IV. As

TABLE IV: Summary of C-INDEX comparison of different models

Data	SGP(L)	SGP(V)	CHGP	GPSRL
Synthetic	0.77 ± 0.05	0.75 ± 0.05	0.75 ± 0.05	0.78 ± 0.04
Sepsis	0.61 ± 0.07	0.62 ± 0.06	0.62 ± 0.07	0.63 ± 0.07

can be seen, our model achieves clearly better performance in both average NLPD and average C-INDEX on the synthetic heterogenous dataset while the all the GP models have more or less similar performance on the sepsis data.

V. CONCLUSION

In this paper, we propose a novel and effective method to model heterogeneity in survival data analysis. Our model, 'Gaussian Process Survival Rule Lists (GPSRL)', utilizes a semi-parametric Bayesian framework to partition the data into subsets with different survival characteristics. This allows us to address some of the limitations of standard survival Gaussian process models, and also provides a degree of interpretability. Experimental results on the synthetic dataset and the MIMIC sepsis survival dataset demonstrate the efficacy of our model by outperforming existing survival GP models. In future work, exploring speedups to GPSRL through either screening for bad proposals of rule lists or computationally efficient approximations, such as stochastic approximations to the latent marginal can improve the model further.

REFERENCES

[1] D. R. Cox, "Regression Models and Life-Tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, pp. 187–220, 1972.

[2] E. L. Kaplan and P. Meier, "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958. [Online]. Available: https://www.jstor.org/stable/2281868

[3] I. Bou-Hamad, D. Larocque, H. Ben-Ameur *et al.*, "A review of survival trees," *Statistics surveys*, vol. 5, pp. 44–71, 2011.

[4] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, M. S. Lauer *et al.*, "Random survival forests," *The annals of applied statistics*, vol. 2, no. 3, pp. 841–860, 2008.

[5] S. Martino, R. Akerkar, and H. Rue, "Approximate Bayesian Inference for Survival Models," *Scandinavian Journal of Statistics*, vol. 38, pp. 514–528, 2011.

[6] H. Joensuu, A. Vehtari, J. Riihimäki, T. Nishida *et al.*, "Risk of recurrence of gastrointestinal stromal tumour after surgery: an analysis of pooled population-based cohorts," *The Lancet. Oncology*, vol. 13, pp. 265–274, 2012.

[7] S. S. Abu-Naser and B. G. Bastami, "A proposed rule based system for breasts cancer diagnosis," *WWJMRD*, 2016.

[8] A. N. Nguyen, M. J. Lawley, D. P. Hansen, R. V. Bowman *et al.*, "Symbolic rule-based classification of lung cancer stages from free-text pathology reports," *Journal of the American Medical Informatics Association*, vol. 17, pp. 440–445, 2010.

[9] E. J. Giamarellos-Bourboulis, A. Norrby-Teglund, V. Mylona, A. Savva *et al.*, "Risk assessment in sepsis: a new prognostication rule by APACHE II score and serum soluble urokinase plasminogen activator receptor," *Critical Care*, vol. 16, p. R149, 2012.

[10] Y. Lin, X. Qian, J. Krischer, K. Vehik, H.-S. Lee, and S. Huang, "A Rule-Based Prognostic Model for Type 1 Diabetes by Identifying and Synthesizing Baseline Profile Patterns," *PLOS ONE*, vol. 9, p. e91095, 2014.

[11] Y. Lin, S. Huang, G. E. Simon, and S. Liu, "Data-based Decision Rules to Personalize Depression Follow-up," *Scientific Reports*, vol. 8, 2018.

[12] M. Haghighi, S. B. Johnson, X. Qian, K. F. Lynch *et al.*, "A Comparison of Rule-based Analysis with Regression Methods in Understanding the Risk Factors for Study Withdrawal in a Pediatric Study," *Scientific Reports*, vol. 6, p. 30828, 2016.

[13] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[14] D. Bertsimas and J. Dunn, "Optimal classification trees," *Machine Learning*, vol. 106, pp. 1039–1082, 2017.

[15] H. A. Chipman, E. I. George, and R. E. McCulloch, "Bayesian Treed Models," *Machine Learning*, vol. 48, pp. 299–320, 2002.

[16] J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," *The Annals of Applied Statistics*, vol. 2, pp. 916–954, 2008.

[17] M. Fokkema, "Fitting Prediction Rule Ensembles with R Package pre," *arXiv:1707.07149 [stat]*, 2017.

[18] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model," *The Annals of Applied Statistics*, vol. 9, pp. 1350–1371, 2015.

[19] J. E. Barrett and A. C. C. Coolen, "Gaussian process regression for survival data with competing risks," *arXiv:1312.1591 [math, stat]*, 2014.

[20] R. B. Gramacy and H. K. H. Lee, "Bayesian Treed Gaussian Process Models With an Application to Computer Modeling," *Journal of the American Statistical Association*, vol. 103, pp. 1119–1130, 2008.

[21] H.-M. Kim, B. K. Mallick, and C. Holmes, "Analyzing nonstationary spatial data using piecewise gaussian processes," *Journal of the American Statistical Association*, vol. 100, no. 470, pp. 653–668, 2005.

[22] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, ser. Adaptive computation and machine learning. Cambridge, Mass: MIT Press, 2006.

[23] V. Gómez-Rubio and H. Rue, "Markov chain monte carlo with the integrated nested laplace approximation," *Statistics and Computing*, vol. 28, no. 5, pp. 1033–1051, 2018.

[24] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. CRC press, 2013.

[25] J. Hensman, A. Matthews, and Z. Ghahramani, "Scalable variational gaussian process classification," *JMLR*, 2015.

[26] A. D. Saul, J. Hensman, A. Vehtari, and N. D. Lawrence, "Chained gaussian processes," in *Artificial Intelligence and Statistics*, 2016, pp. 1431–1440.

[27] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, May 2016. [Online]. Available: https://doi.org/10.1038/sdata.2016.35

[28] J. Harrell, Frank E., R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the Yield of Medical Tests," *JAMA*, vol. 247, pp. 2543–2546, 1982.

[29] GPy, "GPy: A gaussian process framework in python," http://github.com/SheffieldML/GPy, since 2012.