Variational Inference of Joint Models using Multivariate Gaussian Convolution Processe

Xubo Yue[1] and Raed Al Kontar[1]
[1]*Industrial & Operations Engineering, University of Michigan, Ann Arbor, MI, USA*

Abstract

We present a framework for event prediction based on joint modeling of longitudinal and time-to-event data. Our approach exploits a Gaussian convolution process to model signals and a Cox model for survival prognosis. Parameters are estimated by variational inference. Experiments show that our model outperforms approaches built on two-stage inference.

1 Introduction

In recent years, the multivariate Gaussian process (MGP) has drawn significant attention as an efficient non-parametric approach to predict longitudinal signal trajectories. The MGP draws its roots from multitask learning where transfer of knowledge is achieved through a shared representation between training and testing signals. One neat approach that achieves this knowledge transfer, employs convolution processes to construct the MGP. Specifically, each signal is expressed as a convolution of latent functions drawn from a Gaussian process (GP). Commonalities amongst training and testing signals are then captured by sharing these latent functions across the outputs (Álvarez and Lawrence, 2011). Consequently, the multiple signals can be expressed as a single output from a common multivariate Gaussian convolution process (MGCP). Indeed, many recent studies have demonstrated the MGCP ability to account for non-trivial commonalities in the data and provide accurate predictive results (Zhao and Sun, 2016; Cheng, 2018).

In this article we exploit the MGCP to explore the following question: can we use both time-to-event data (also known as survival data) along with longitudinal signals to obtain a reliable event prediction? This is illustrated in Figure 1. As shown in the figure, our goal is to utilize both survival data and longitudinal signals from training units to predict the survival probability of a partially observed testing unit. Naturally, the aforementioned question is often encountered in a wide range of applications, including: event prediction using vital health signals from monitored patients at risk, remaining useful life (RUL) estimation of operational units/machines and failure prognosis in connected manufacturing systems (e.g., nuclear power plants). In order to link survival and longitudinal data, state-of-the-art methods have focused on joint models. The seminal work of Rizopoulos

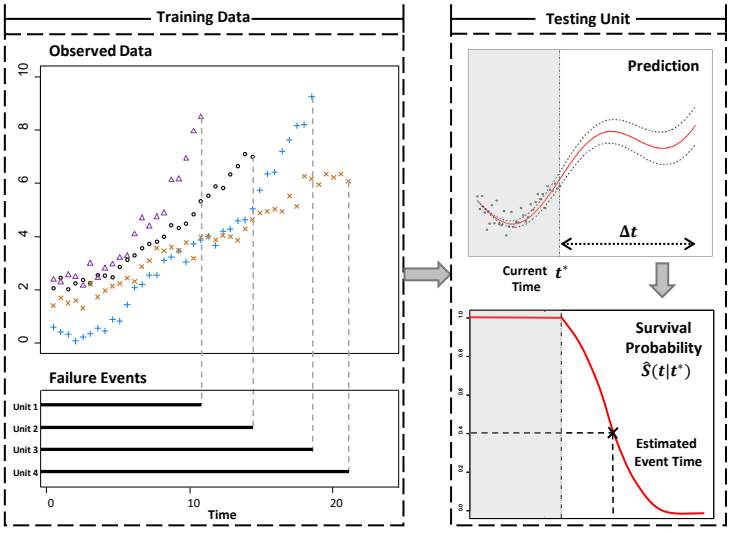

Figure 1: Joint modeling of longitudinal and time-to-event data.

(Rizopoulos, 2012) laid a foundation for joint models where a linear mixed effects model is used to model longitudinal signals. The coefficients of the mixed model are then used in a Cox model to compute the probability of event occurrence conditioned on the observed longitudinal signals. This idea provided the bases for many extensions and applications in the literature (Proust-Lima et al., 2014; Rizopoulos et al., 2017). It is important to note here that joint methods are in general built using a two-stage inference procedure. In two-stage inference, features from the longitudinal data are first learned, these estimated features are then inserted into a survival model to predict event probabilities. Indeed, many papers have shown that this two-stage procedure can produce competitive predictive results (Zhou et al., 2014). Nevertheless, the foregoing works are based on strong parametric assumptions where signals are assumed to follow a specific parametric form and all the signals (training and testing) exhibit that same functional form. In other words, signals behave according to a similar trend but at different rates (i.e., different parameter values). However, parametric methods are restrictive in many applications and if the specified form is far from the truth, predictive results will be misleading. Furthermore, the assumption that all signals possess the same functional form may not hold in real-life applications. For instance, units operated under different environmental conditions may exhibit different signal evolution rates and trends. Some recent efforts aimed to relax strong parametric assumptions using splines, continuous time Markov chains and the GP. Unfortunately, these methods still assume homogeneity across the population and focus on merely imputing the longitudinal data rather than predicting signal evolution within a time interval of interest (Soleimani et al., 2018). We here note that there has been some recent attempts at rebuilding the Cox model using a GP (Kim and Pavlovic, 2018). However these approaches are only based on survival data and do not handle joint modeling, which is the focus of this article. To address the aforementioned challenges, we propose a flexible joint modeling approach denoted as MGCP-Cox. Our approach exploits the MGCP to model the evolution of longitudinal signals and a Cox model to map time-to-event data with longitudinal data modeled through MGCP. Event occurrence probability is then derived within any future interval Δt as shown in Figure 1. We also propose a variational inference framework using pseudo-inputs (Snelson and Ghahramani, 2006) to simultaneously estimate parameters in the joint MGCP-Cox model. This facilitates scalability to large data settings and safeguards against model overfitting. Finally, the advantageous features of the proposed method are demonstrated through numerical studies and a case study with real-world data in the application to finding the remaining useful lifetime (RUL) of NASA Aero-propulsion engines.

2 Background: Survival Analysis

Survival analysis is a branch of statistics for analyzing time-to-event data and predicting the probability of occurrence of an event. For each individual unit i, the associated data is $\boldsymbol{\mathcal{D}}_i = (V_i, \delta_i, \boldsymbol{Y}_i, \boldsymbol{w}_i)$, where $V_i = \min\{T_i, C_i\}$ is the event time (the unit failed at time T_i or was censored at time C_i), $\delta_i \in \{0, 1\}$ is an event indicator ($\delta_i = 1/0$ indicates the unit has failed/censored), $\boldsymbol{Y}_i$ are the noisy observed longitudinal data (e.g., vital signals collected from patients) corresponding to the underlying latent values $\boldsymbol{f}_i$, and $\boldsymbol{w}_i$ is a set of time-invariant features (e.g., patient's gender). Typically, the continuous random variable T_i is characterized by a survival function $S(t) = P(T \geq t)$ which represents the probability of survival up to time t. Another important term is the hazard function $h(t) = \lim_{\Delta \to 0} \frac{1}{\Delta} P(t < T \leq t + \Delta | T \geq t) = -\frac{d}{dt} \log S(t)$ and can be thought of as the instantaneous rate of occurrence of an event at time t. It is easy to show that $S(t) = \exp\{-\int_0^t h(u)du\}$. The term $\int_0^t h(u)du$ is called cumulative hazard function and is denoted

by $H(t)$. The basic scheme of survival analysis is to find suitable models to explain relationships between the hazard function $h_i(t)$ and collected data $\boldsymbol{\mathcal{D}}_i$. These models are defined as survival models. Many survival models have been developed to analyze time-to-event data. They typically model the hazard function as a function of some time-varying and fixed features. One class of prevailed survival models is called the Cox model, which has the form $h_i(t) = h_0(t)\exp[\boldsymbol{\gamma}^T\boldsymbol{w}_i + \beta f_i(t)]$, where $h_0(t)$ is a baseline hazard function shared by all individuals, and is typically modeled by the Weibull or a piecewise constant function, $\boldsymbol{\gamma}$ is a vector of coefficients for the fixed covariates (features), $f_i(t)$ is the feature estimated by a longitudinal model (e.g., Gaussian Process), and β is a scaling parameter for the time-varying covariates. Parameters in the Cox model are typically estimated by maximizing the log-likelihood function

$$\sum_{i=1}^{N}\log p(V_i,\delta_i|\boldsymbol{w}_i,\boldsymbol{f}_i) = \sum_{i=1}^{N}\{\delta_i\log\Big[h_0(V_i)\exp[\boldsymbol{\gamma}^T\boldsymbol{w}_i+\beta f_i(V_i)]\Big] - \int_0^{V_i} h_0(u)\exp[\boldsymbol{\gamma}^T\boldsymbol{w}_i+\beta f_i(u)]du\}.$$

Given an estimate of parameters from the Cox model, we can then obtain the event (failure) probability within a future time interval Δt given the fact that the testing unit i survives non-shorter than the current time instance t^*. This probability, denoted $\hat{P}_{\Delta t}$, is estimated as follows:

$$\begin{aligned}\hat{P}_{\Delta t} &= 1-\hat{S}(t^*+\Delta t|t^*,\boldsymbol{w}_i,\boldsymbol{f}_i) = 1-\frac{\hat{S}(t^*+\Delta t|\boldsymbol{w}_i,\boldsymbol{f}_i)}{\hat{S}(t^*|\boldsymbol{w}_i,\boldsymbol{f}_i)}\\ &= 1-\exp\Big\{-\int_{t^*}^{t^*+\Delta t}\hat{h}_0(u)\exp\Big[\hat{\boldsymbol{\gamma}}^T\boldsymbol{w}_i+\hat{\beta}f_i(u)\Big]du\Big\}.\end{aligned} \tag{1}$$

where $\boldsymbol{w}_i$ and $\boldsymbol{f}_i$ are features for a testing unit i.

3 Joint Modeling and Variational Inference

3.1 The multivariate Gaussian convolution process (MGCP)

Assume data have been collected from N units and let $\mathcal{I} = \{1, 2, \ldots, N\}$ denote the set of all units. For unit i, its associated data is $\boldsymbol{\mathcal{D}}_i = \{V_i, \delta_i, \boldsymbol{Y}_i, \boldsymbol{w}_i\}$. The observed longitudinal signal is denoted by $\boldsymbol{Y}_i = (y_i(t_{i1}), \ldots, y_i(t_{il_i}))^T$, where l_i represents the number of observations and $\{t_{ir} : r = 1, \ldots, l_i\}$ denotes the inputs. We decompose the longitudinal signal as $y_i(t) = f_i(t) + \epsilon_i(t)$, where $f_i(\cdot)$ is a mean zero GP and $\epsilon_i(t)$ denotes additive noise with zero mean and σ_ϵ^2 variance.

To obtain an accurate predictive result, we need to capture the intrinsic relatedness among N signals. Particularly, we resort to the convolution process to model the latent function $f_i(t)$. We consider K independent latent functions $\{X_k(t)\}_{k=1}^K$ and NK different smoothing kernels $\{G_{i,k}(t) : i \in \mathcal{I}\}_{k=1}^K$. The latent functions are assumed independent GPs with covariance $\text{cov}[X_k(t), X_k(t')] = \kappa_k(t,t')$. We set $G_{i,k}(t) = \alpha_{i,k}\mathcal{N}(0,\xi_{i,k}^2)$ to be scaled Gaussian kernels and $\kappa_k(t,t')$ to be squared exponential covariance functions.

$$\kappa_k(t,t') = \exp\left[-\frac{1}{2}\frac{(t-t')^2}{\lambda_k^2}\right] = \sqrt{2\pi\lambda_k^2}\mathcal{N}(0,\lambda_k^2) \coloneqq C_k\mathcal{N}(0,\lambda_k^2), \tag{2}$$

The GP $f_i(t)$ is then constructed by convolving the shared latent functions with the smoothing kernel as shown in (3). This is the underlying principle of the MGCP, where the latent functions

$\{X_k(t)\}_{k=1}^K$ are shared across different outputs through the corresponding kernels $G_{i,k}(t)$. Since convolutions are linear operators on a function and since the latent function, a GP, is shared across multiple outputs then all outputs can be expressed as a jointly distributed GP, an MGCP. A key feature is that information is shared through different parameters encoded in the kernels $G_{i,k}(t)$. Outputs then can possess both shared and unique features. Thus, accounting for heterogeneity in the longitudinal data.

$$f_i(t) = \sum_{k=1}^{K} \int_{\mathbb{R}} G_{i,k}(t-u) X_k(u) du. \tag{3}$$

Based on equation (3), the covariance function between f_i and f_j, and the covariance function between f_i and X_k, can be calculated as

$$\begin{aligned}
\kappa_{f_i,f_j}(t,t') &= \sum_{k=1}^{K} \int_{\mathbb{R}} G_{i,k}(t-u) \int_{\mathbb{R}} G_{j,k}(t'-u') \kappa_k(u,u') du' du \\
&= \sum_{k=1}^{K} \alpha_{i,k}\alpha_{j,k} \sqrt{\frac{\lambda_k^2}{\eta_{i,j,k}^2}} \exp\Big(-\frac{1}{2}\frac{(t-t')^2}{\eta_{i,j,k}^2}\Big), \\
\kappa_{f_i,X_k}(t,u) &= \int_{\mathbb{R}} G_{i,k}(t-u')\kappa_k(u,u')du' = \alpha_{i,k}\sqrt{\frac{\lambda_k^2}{\eta_{i,k}^2}} \exp\Big(-\frac{1}{2}\frac{(t-u)^2}{\eta_{i,k}^2}\Big),
\end{aligned} \tag{4}$$

where $\eta_{i,j,k}^2 = \xi_{i,k}^2 + \xi_{j,k}^2 + \lambda_k^2$ and $\eta_{i,k}^2 = \xi_{i,k}^2 + \lambda_k^2$.

Denote the underlying latent values as $\boldsymbol{f} = \{\boldsymbol{f}_1^T, \ldots, \boldsymbol{f}_N^T\}^T$, where $\boldsymbol{f}_i = \{f_i(t_{i1}), \ldots, f_i(t_{il_i})\}^T$. The density function of $\boldsymbol{f}$ can be obtained as $p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}})$, where $\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}$ sized $(\sum_{i=1}^N l_i) \times (\sum_{i=1}^N l_i)$ is the covariance function. The likelihood of $\boldsymbol{f}$ involves inverting the large matrix $\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}}$. This operation has computational complexity $\mathcal{O}((\sum_{i=1}^N l_i)^3)$ and storage requirement $\mathcal{O}((\sum_{i=1}^N l_i)^2)$. To alleviate computational burden, we introduce M pseudo-inputs from the latent functions denoted as $\boldsymbol{X}_k(\boldsymbol{\mathcal{Z}}) = [X_k(z_1), \ldots, X_k(z_M)]^T$ where $\boldsymbol{\mathcal{Z}} = \{z_i\}_{i=1}^M$. Since the latent functions are GPs, then any sample $\boldsymbol{X}_k(\boldsymbol{\mathcal{Z}})$ follows a multivariate Gaussian distribution. Conditioned on $\boldsymbol{X}_k(\boldsymbol{\mathcal{Z}})$, we next sample from the conditional prior $p(X_k(u)|\boldsymbol{X}_k(\boldsymbol{\mathcal{Z}}))$. In equation (3), $X_k(u)$ can be approximated well by the expectation $\mathbb{E}(X_k(u)|\boldsymbol{X}_k(\boldsymbol{\mathcal{Z}}))$ as long as the latent functions are smooth (Álvarez and Lawrence, 2011). Denote by $\boldsymbol{X} = \{\boldsymbol{X}_1^T(\boldsymbol{\mathcal{Z}}), \ldots, \boldsymbol{X}_K^T(\boldsymbol{\mathcal{Z}})\}^T$. The probability distribution of $\boldsymbol{X}$ can be expressed as $p(\boldsymbol{X}|\boldsymbol{\mathcal{Z}}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}})$, where $\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}$ is a block-diagonal matrix such that each block is associated with the covariance of X_k in (2). By multivariate Gaussian identities, the probability distribution of $\boldsymbol{f}$ conditional on $\boldsymbol{X}, \boldsymbol{\mathcal{Z}}$ is $p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\mathcal{Z}}) = \mathcal{N}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{X}}\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1}\boldsymbol{X}, \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})$ where $\boldsymbol{Q} = \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{X}}\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1}\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{f}}$. Therefore, $p(\boldsymbol{f})$ can be approximated by $p(\boldsymbol{f}|\boldsymbol{\mathcal{Z}})$, which is given as $p(\boldsymbol{f}|\boldsymbol{\mathcal{Z}}) = \int p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\mathcal{Z}})p(\boldsymbol{X}|\boldsymbol{\mathcal{Z}})d\boldsymbol{X}$. Therefore, $p(\boldsymbol{Y})$ can then be approximated by $p(\boldsymbol{Y}|\boldsymbol{\mathcal{Z}}) = \int p(\boldsymbol{Y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\mathcal{Z}})p(\boldsymbol{X}|\boldsymbol{\mathcal{Z}})d\boldsymbol{f}d\boldsymbol{X}$.

3.2 Joint Model and Variational Inference

Now following our convolution construction in (3), the hazard function at time t is given as

$$h_i(t) = h_0(t) \exp\Big[\boldsymbol{\gamma}^T\boldsymbol{w}_i + \beta \sum_{k=1}^{K} \int_{\mathbb{R}} G_{i,k}(t-u)X_k(u)du\Big]. \tag{5}$$

This key equation links the MGCP to the Cox model. We begin with presenting the log-likelihood of the joint model given observed data $\boldsymbol{\mathcal{D}} = \{\boldsymbol{\mathcal{D}}_i\}_{i=1}^N$. The marginal log-likelihood function is

$$\log p(\boldsymbol{\mathcal{D}}) = \log \int p(\boldsymbol{\mathcal{D}}|\boldsymbol{f})p(\boldsymbol{f})d\boldsymbol{f} \approx \int p(\boldsymbol{\mathcal{D}}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\mathcal{Z}})p(\boldsymbol{X}|\boldsymbol{\mathcal{Z}})d\boldsymbol{X}d\boldsymbol{f}.$$

We would like to provide a good approximation of $\log p(\boldsymbol{\mathcal{D}})$ by introducing an evidence lower bound (ELBO) $\mathcal{L}$. This bound is calculated by finding the Kullback-Leibler (KL) divergence between the variational density $q(\boldsymbol{f},\boldsymbol{X}|\boldsymbol{\mathcal{Z}})$ and the true posterior density $p(\boldsymbol{f},\boldsymbol{X}|\boldsymbol{\mathcal{D}},\boldsymbol{\mathcal{Z}})$. Specifically,

$$\begin{aligned} KL(q(\boldsymbol{f},\boldsymbol{X}|\boldsymbol{\mathcal{Z}}) \parallel p(\boldsymbol{f},\boldsymbol{X}|\boldsymbol{\mathcal{D}},\boldsymbol{\mathcal{Z}})) &= \int q(\boldsymbol{f},\boldsymbol{X}|\boldsymbol{\mathcal{Z}}) \log \frac{q(\boldsymbol{f},\boldsymbol{X}|\boldsymbol{\mathcal{Z}})}{p(\boldsymbol{f},\boldsymbol{X}|\boldsymbol{\mathcal{D}},\boldsymbol{\mathcal{Z}})} d\boldsymbol{X}d\boldsymbol{f} \\ &= \log p(\boldsymbol{\mathcal{D}}) - \int q(\boldsymbol{f},\boldsymbol{X}|\boldsymbol{\mathcal{Z}}) \log \frac{p(\boldsymbol{f},\boldsymbol{X},\boldsymbol{\mathcal{D}}|\boldsymbol{\mathcal{Z}})}{q(\boldsymbol{f},\boldsymbol{X}|\boldsymbol{\mathcal{Z}})} d\boldsymbol{X}d\boldsymbol{f} \\ &= \log p(\boldsymbol{\mathcal{D}}) - \mathcal{L} \geq 0. \end{aligned} \tag{6}$$

The variaitonal density is assumed to be factorized as $q(\boldsymbol{f},\boldsymbol{X}|\boldsymbol{\mathcal{Z}}) = p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\mathcal{Z}})q(\boldsymbol{X})$. Maximizing the ELBO with respect to $q(\boldsymbol{X})$ and hyperparameters from the MGCP-Cox model can achieve purposes of variational inference and model selection simultaneously. By equation (6),

$$\begin{aligned} \mathcal{L} &= \int q(\boldsymbol{f},\boldsymbol{X}|\boldsymbol{\mathcal{Z}}) \log \frac{p(\boldsymbol{f},\boldsymbol{X},\boldsymbol{\mathcal{D}}|\boldsymbol{\mathcal{Z}})}{q(\boldsymbol{f},\boldsymbol{X}|\boldsymbol{\mathcal{Z}})} d\boldsymbol{X}d\boldsymbol{f} \\ &= \int q(\boldsymbol{X}) \int p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\mathcal{Z}}) \log p(\boldsymbol{\mathcal{D}}|\boldsymbol{f}) d\boldsymbol{f}d\boldsymbol{X} + \int q(\boldsymbol{X}) \log \frac{p(\boldsymbol{X}|\boldsymbol{\mathcal{Z}})}{q(\boldsymbol{X})} d\boldsymbol{X}. \end{aligned} \tag{7}$$

Furthermore, we can decompose $\log p(\boldsymbol{\mathcal{D}}|\boldsymbol{f}) = \log p(\boldsymbol{Y}|\boldsymbol{f}) + \log p(\boldsymbol{V},\boldsymbol{\delta}|\boldsymbol{w},\boldsymbol{f})$, where $\boldsymbol{V} = \{V_i\}_{i=1}^N$, $\boldsymbol{\delta} = \{\delta_i\}_{i=1}^N$ and $\boldsymbol{w} = \{w_i\}_{i=1}^N$. Based on equation (7), the MGCP propagates uncertainties through the latent processes to the Cox model.

It is desirable to find a closed form of the ELBO in equation (7). Since $p(\boldsymbol{Y}|\boldsymbol{f})$ and $p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\mathcal{Z}})$ are both Gaussian, we can obtain

$$\int p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\mathcal{Z}}) \log p(\boldsymbol{Y}|\boldsymbol{f}) d\boldsymbol{f} = \log \mathcal{N}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{X}}\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1}\boldsymbol{X}, \sigma_\epsilon^2 \boldsymbol{I}) - \frac{1}{2\sigma_\epsilon^2}\text{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q}), \tag{8}$$

where $\text{Tr}(\cdot)$ is a trace operator. Therefore, the ELBO can be simplified as

$$\begin{aligned} \mathcal{L} = &-\frac{1}{2\sigma_\epsilon^2}\text{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q}) + \int q(\boldsymbol{X}) \log \frac{\mathcal{N}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{X}}\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1}\boldsymbol{X}, \sigma_\epsilon^2 \boldsymbol{I}) p(\boldsymbol{X}|\boldsymbol{\mathcal{Z}})}{q(\boldsymbol{X})} d\boldsymbol{X} \\ &+ \int q(\boldsymbol{X}) p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\mathcal{Z}}) \log p(\boldsymbol{V},\boldsymbol{\delta}|\boldsymbol{w},\boldsymbol{f}) d\boldsymbol{f}d\boldsymbol{X} \end{aligned} \tag{9}$$

We compute the optimal upper bound of $\mathcal{L}$ by reversing Jensen's inequality. This gives

$$\begin{aligned} &\mathcal{L}^* \\ &= \log \int \mathcal{N}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{X}}\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1}\boldsymbol{X}, \sigma_\epsilon^2 \boldsymbol{I}) p(\boldsymbol{X}|\boldsymbol{\mathcal{Z}}) d\boldsymbol{X} + \boldsymbol{\zeta} + \int q(\boldsymbol{X}) p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\mathcal{Z}}) \log p(\boldsymbol{V},\boldsymbol{\delta}|\boldsymbol{w},\boldsymbol{f}) d\boldsymbol{f}d\boldsymbol{X} \\ &= \log[\mathcal{N}(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I} + \boldsymbol{Q})] + \boldsymbol{\zeta} + \int q(\boldsymbol{X}) p(\boldsymbol{f}|\boldsymbol{X},\boldsymbol{\mathcal{Z}}) \log p(\boldsymbol{V},\boldsymbol{\delta}|\boldsymbol{w},\boldsymbol{f}) d\boldsymbol{f}d\boldsymbol{X}, \end{aligned} \tag{10}$$

where $\boldsymbol{\zeta} = -\frac{1}{2\sigma_\epsilon^2}\text{Tr}(\boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{Q})$. $\boldsymbol{\zeta}$ can be thought of as a penalization term that regularizes the estimation of the parameters. The first two terms in equation (10) can be computed in $\mathcal{O}((\sum_{i=1}^N l_i)M^2)$.

3.3 Variational Inference on Cox Model

Parameters in the Cox model can be attained by maximizing the following log-likelihood function:

$$\log p(\boldsymbol{V}, \boldsymbol{\delta}|\boldsymbol{w}, \boldsymbol{f}) = \sum_{i=1}^{N} \log p(V_i, \delta_i|\boldsymbol{w}_i, \boldsymbol{f}_i) = \sum_{i=1}^{N} \Big\{\delta_i \log \Big[h_0(V_i) \exp[\boldsymbol{\gamma}^T \boldsymbol{w}_i + \beta \sum_{k=1}^{K} \int_{\mathbb{R}} G_{i,k}(V_i - u) X_k(u) du]\Big] - \int_0^{V_i} h_0(u) \exp[\boldsymbol{\gamma}^T \boldsymbol{w}_i + \beta \sum_{k=1}^{K} \int_{\mathbb{R}} G_{i,k}(u - v) X_k(v) dv] du\Big\}. \tag{11}$$

In equation (10), we obtain the optimal $q^*(\boldsymbol{X})$ to maximize the ELBO. The variational parameters in $q^*(\boldsymbol{X})$ are crucial in computing the likelihood of the Cox model. Specifically, the optimal $q^*(\boldsymbol{X})$ has the form $q^*(\boldsymbol{X}) = \mathcal{N}(\sigma_\epsilon^{-2} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}} (\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}} + \sigma_\epsilon^{-2} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{f}} \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{X}})^{-1} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{f}} \boldsymbol{Y}, \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}} (\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}} + \sigma_\epsilon^{-2} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{f}} \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{X}})^{-1} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}})$. Denote by $q^*(\boldsymbol{X}) \coloneqq \mathcal{N}(\boldsymbol{m}, \boldsymbol{s})$. It is easy to show that $q(\boldsymbol{f}|\boldsymbol{\mathcal{Z}})$ has the normal distribution with parameter $\boldsymbol{\mu}, \boldsymbol{\Sigma}$. Specifically, $\int q^*(\boldsymbol{X}) p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\mathcal{Z}}) d\boldsymbol{X} = q(\boldsymbol{f}|\boldsymbol{\mathcal{Z}}) \coloneqq q(\boldsymbol{f}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{X}} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1} \boldsymbol{m}, \boldsymbol{\Sigma} = \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{X}} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1} (\boldsymbol{I} - \boldsymbol{s} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1}) \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{f}}$. The last integration in equation (10) can be simplified to

$$\begin{aligned} &\int q(\boldsymbol{f}) \log p(\boldsymbol{V}, \boldsymbol{\delta}|\boldsymbol{w}, \boldsymbol{f}) d\boldsymbol{f} = \int q(\boldsymbol{f}) \sum_{i=1}^{N} \log p(V_i, \delta_i|\boldsymbol{w}_i, \boldsymbol{f}_i) d\boldsymbol{f} \\ &= \int q(\boldsymbol{f}) \sum_{i=1}^{N} \Big\{\delta_i \log \Big[h_0(V_i) \exp[\boldsymbol{\gamma}^T \boldsymbol{w}_i + \beta f_i(V_i)]\Big] - \int_0^{V_i} h_0(u) \exp[\boldsymbol{\gamma}^T \boldsymbol{w}_i + \beta f_i(u)] du\Big\} d\boldsymbol{f}. \end{aligned} \tag{12}$$

The first term in equation (12) can be calculated analytically. For each unit i,

$$\int q(\boldsymbol{f}) \delta_i \log[h_0(V_i) \exp[\boldsymbol{\gamma}^T \boldsymbol{w}_i + \beta f_i(V_i)]] d\boldsymbol{f} = \delta_i \Big\{\log h_0(V_i) + \boldsymbol{\gamma}^T \boldsymbol{w}_i + \beta \mathbb{E}_{q(\boldsymbol{f})}[f_i(V_i)]\Big\}, \tag{13}$$

where $\mathbb{E}_{q(\boldsymbol{f})}[f_i(V_i)] = \boldsymbol{K}_{f_i(V_i),\boldsymbol{X}} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1} \boldsymbol{m} \coloneqq \mu_i(V_i)$. The second term in equation (12) can be estimated by the numerical integration. For each unit i,

$$\begin{aligned} &\int q(\boldsymbol{f}) (-\int_0^{V_i} h_0(u) \exp[\boldsymbol{\gamma}^T \boldsymbol{w}_i + \beta f_i(u)] du) d\boldsymbol{f} \\ &= -\int_0^{V_i} h_0(u) \exp[\boldsymbol{\gamma}^T \boldsymbol{w}_i] \int q(\boldsymbol{f}) \exp[\beta f_i(u)] d\boldsymbol{f} du \\ &= -\int_0^{V_i} h_0(u) \exp[\boldsymbol{\gamma}^T \boldsymbol{w}_i] \exp\Big[\beta[\mu_i(u) + \frac{1}{2}\sigma_i^2(u)]\Big] du, \end{aligned} \tag{14}$$

where $\mu_i(u) \coloneqq \boldsymbol{K}_{f_i(u),\boldsymbol{X}} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1} \boldsymbol{m}$, and $\sigma_i^2(u) \coloneqq \boldsymbol{K}_{f_i(u),f_i(u)} - \boldsymbol{K}_{f_i(u),\boldsymbol{X}} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1} (\boldsymbol{I} - \boldsymbol{s} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1}) \boldsymbol{K}_{\boldsymbol{X},f_i(u)}$. We can assume $h_0(t)$ to be an exponential function $\exp(b + \psi(t - \min\{V_i\}_{i=1}^N))$, where b,ψ are parameters to be learned. To obtain a good baseline hazard prediction given the estimated $\hat{b}, \hat{\psi}$, we can calculate the cumulative hazard at time point t as $H(t) = \sum_{u \in \mathcal{F}(t)} \hat{h}_0(u), \forall t$, where $\mathcal{F}(t) \coloneqq \Big\{\{V_{(1)}, V_{(2)}, \ldots, V_{(N-1)}\} \cup \{0, 1, 2, \ldots, V_{(N)}\}\Big\} \cap [0, t]$, and $V_{(i)}$ is the i-th smallest element in $\{V_i\}_{i=1}^N$. Then we fit a regularized smooth spline to $H(t)$. The predicted baseline hazard at $u \in [t^*, t^* + \Delta t^*]$ can be estimated by $\frac{d\hat{H}(t)}{dt}\Big|_{t=u}$.

The $\mathcal{L}^*$ is maximized with respect to the parameters $\Theta = (\boldsymbol{\theta}, \sigma_\epsilon, \boldsymbol{\gamma}, \beta, b, \psi)$, where $\boldsymbol{\theta} = (\{\lambda_k, \xi_{i,k}, \alpha_{i,k}\}_{i=1,k=1}^{N,K})$, by the gradient-based method.

3.4 Event Prediction

Without loss of generality, we focus on predicting the event occurrence probability for unit N. Suppose observations from the testing unit N have been collected up to time t^*. The survival model computes the event probabilities conditioned on the predicted longitudinal features $\boldsymbol{f}_N(u), u \in [t^*, t^* + \Delta t]$. Given estimated parameters, and following (1), we are interested in calculating

$$\begin{aligned} 1 - \hat{S}(t^* + \Delta t | t^*, \boldsymbol{w}_N, \boldsymbol{f}_N) &= 1 - \frac{\hat{S}(t^* + \Delta t | \boldsymbol{w}_N, \boldsymbol{f}_N)}{\hat{S}(t^* | \boldsymbol{w}_N, \boldsymbol{f}_N)} \\ &= 1 - \exp\{-\int_{t^*}^{t^*+\Delta t} \hat{h}_0(u) \exp\left[\hat{\boldsymbol{\gamma}}^T \boldsymbol{w}_N + \hat{\beta} f_N(u)\right] du\}. \end{aligned} \tag{15}$$

Based on equation (15), the accurate extrapolation within Δt is essential. In the MGCP, the predictive distribution for any new input point T is given by

$$\begin{aligned} p(f_N(T^*)|\boldsymbol{Y}) &= \int p(f_N(T^*)|\boldsymbol{X}) p(\boldsymbol{X}|\boldsymbol{Y}) d\boldsymbol{X} = \int \mathcal{N}(\boldsymbol{K}_{f_N(T^*),\boldsymbol{X}} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1} \boldsymbol{X}, \boldsymbol{W}) p(\boldsymbol{X}|\boldsymbol{Y}) d\boldsymbol{X} \\ &= \int \mathcal{N}(\boldsymbol{K}_{f_N(T^*),\boldsymbol{X}} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1} \boldsymbol{X}, \boldsymbol{W}) \frac{p(\boldsymbol{Y}|\boldsymbol{X}) p(\boldsymbol{X})}{p(\boldsymbol{Y})} d\boldsymbol{X} = \mathcal{N}(\boldsymbol{A}\boldsymbol{D}^{-1}\boldsymbol{Y}, \boldsymbol{K}_{f_N(T^*),f_N(T^*)} - \boldsymbol{A}\boldsymbol{D}^{-1}\boldsymbol{A}^T), \end{aligned} \tag{16}$$

where $\boldsymbol{A} = \boldsymbol{K}_{f_N(T^*),\boldsymbol{X}} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{f}}$, $\boldsymbol{W} = K_{f_N(T^*),f_N(T^*)} - \boldsymbol{K}_{f_N(T^*),\boldsymbol{X}} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1} \boldsymbol{K}_{\boldsymbol{X},f_N(T^*)}$ and $\boldsymbol{D} = \boldsymbol{Q} + \sigma_\epsilon^2 \boldsymbol{I}$. We have used $K_{f_N(T^*),f_N(T^*)}$ as a notation to indicate when the covariance matrix is evaluated at the T^*. Consequently, the predicted signal is $\hat{f}_N(T^*) = \boldsymbol{A}\boldsymbol{D}^{-1}\boldsymbol{Y}$.

4 Experiments

4.1 General setting

For the synthetic data we assume that the underlying true path for unit i has the form $y_i(t) = \boldsymbol{z}^T(t)\boldsymbol{b}_i + \epsilon_i(t) = b_{i0} + b_{i1}t + b_{i2}t^2 + \epsilon_i(t)$, where $\epsilon_i(t) \sim \mathcal{N}(0, 0.1)$, $\boldsymbol{z}^T(t) = [1, t, t^2]$ and $\boldsymbol{b}_i = [b_{i0}, b_{i1}, b_{i2}]^T \sim \mathcal{N}(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$ with $\boldsymbol{\mu}_b = [2.5, 0.1, a]^T$ and $\boldsymbol{\Sigma}_b = \begin{bmatrix} 0.2 & -4e-4 & -8e-5 \\ -4e-4 & 3e-6 & 3e-7 \\ -e-5 & 3e-7 & 1e-7 \end{bmatrix}$ where $a \sim$ uniform$(0.003, 0.03)$. Without loss of generality, we assume that the time unit is month and that signals were obtained regularly at each month up to their failure or censoring time. For each unit we specify a time-invariant feature $w_i \in \{0, 1\}$ generated by a Bernoulli distribution with $p = 0.5$. In the Cox model, we use the Weibull baseline hazard rate function $h_0(t) = \lambda\rho t^{\rho-1}$ with $\lambda = 0.001$ and $\rho = 1.05$. We generate the failure time T_i for each unit by rejection sampling using its probability density function $h_i(t)S_i(t)$. We set $\gamma = 0$ and $\beta = 0.5$. Also, we randomly select 5% of the units to be right censored. The number of units generated is $N = 20$ and the experiment is repeated for $Q = 100$ times.

For the real-world case study we use the C-MAPSS dataset provided by the National Aeronautics and Space Administration (NASA). The dataset contains failure time data of aircraft turbofan engines and degradation signals from informative sensors mounted on these engines. Note that in our analysis we standardize all sensor data. We refer readers to Liu et al. (2013) for more details about the data.

We focus on predicting the event probability within a future time interval Δt. We consider $\Delta t = 12, 15, 20$ months in this simulation study. Prediction performance at varying time points t^* for the partially observed unit N is then reported. The time instant $t^* = \alpha T_N$ is defined as the α-observation percentile, where T_N is the failure time of unit N. The values of α are specified as 30%, 50%. Figure 2 shows some examples of units observed up to different percentiles of their failure time. Further, in our simulation studies, we benchmark our method with three other reference methods for comparison: (1) Logistic Regression (LR) classifier: in the LR, event data is transformed into binary labels $\delta_i = 1/0$ denoting whether units failed or not within the time interval $[t^*, \Delta t + t^*]$. The time-fixed covariate w_i and the last observed signal measurement at t^* are used as the model predictors. (2) Support Vector Machine (SVM) classifier: We use the radial basis kernel and determine parameters using 2-fold cross-validation on the training data. (3) Parametric Joint Model (LMM-Joint): we implement a state-of-the-art joint modeling algorithm using the mixed-effect model. The LMM-joint uses a polynomial function whose corresponding degree is determined through an Akaike information criteria to model the signal path. Note that this framework estimates parameters from the mixed-effect model and the Cox model separately (Zhou et al., 2014). Regarding our MGCP-Cox model we set the number of pseudo-inputs to $M = 10$ and the number of latent functions to $K = 1$. The performance of each method is then assessed by the Receiver Operating Characteristic (ROC) curve. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR). Predictive accuracy is then assessed through the area under the curve (AUC). The results from the synthetic data are shown in Figure 3. Due to poor performance of both the LR and SVM on $N = 20$, we also checked whether they can produce comparable results to the MGCP-Cox when $N = 200$. We denote those models as LR-200 and SVM-200.

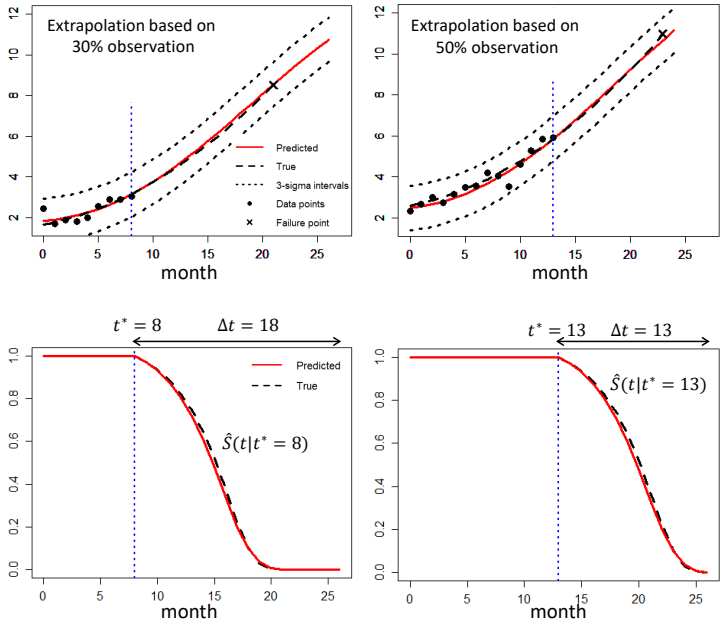

Figure 2: Results of extrapolations and survival probabilities prediction for two testing units.

For the real data, the true survival probabilities are not available since we do not have information about the underlying parameters used to generate the data. Therefore, to evaluate model performance, we calculate the mean remaining lifetime (RL) of the testing unit, which is defined as $\widehat{mrl}(t^*) = \int_{t^*}^{\infty} \hat{S}(u|t^*, \boldsymbol{w}_N, \boldsymbol{f}_N)du$. This integration can be obtained by the Gauss-Legendre quadrature. The performance is assessed by the absolute error $AE = |rl_j - \widehat{mrl}_j|$ where rl_j is the true remaining lifetime of the testing unit. We then report the distribution of the errors across all units using the boxplot in Figure 4. Similar to the synthetic data we use 30% and 50% percentiles to assess prediction accuracy. We also note than we cannot obtain $\widehat{mrl}$ estimates from the SVM and

LR as they transform event prediction into a time series classification problem. Therefore, only results from LMM-joint and MGCP-Cox are reported in Figure 4.

4.2 Results

The illustrative example in Figure 2 demonstrates the behavior of our method. As shown in the figure, our joint model framework can provide accurate predictions of both longitudinal signals and event probabilities. The unique smoothing kernel $G_{i,k}$ for each individual allows flexibility in the prediction, since it enables each training signal to have its own characteristics. This substantiates the strength of the MGCP. Equipped with the shared latent processes, the model can infer the similarities among all units, and predict signal trajectory by borrowing strength from training units. The results in Figures 3 and 4 indicate that our MGCP-Cox model clearly outperforms the benchmarked models. Based on the figure we can obtain some important insights. First, as expected, prediction errors significantly decrease as the lifetime percentiles increase. Thus, the prediction accuracy from the MGCP-Cox model will become more accurate as t^* increases and more data are collected from an online monitored unit. Second, the prediction accuracy slightly decreases as we predict over a longer horizon (i.e. prediction is better for the near future). This is intuitively understandable as accuracy might decrease when predicting over a large region where not many training data might be observed. Third, the results show that the MGCP-Cox clearly outperforms LMM-joint. This result highlights the danger of parametric modeling and demonstrates the ability of our non-parametric approach to avoid model misspecifications. Fourth, even when the LR and SVM had a much larger number of units, the MGCP-Cox was still superior. This observation, also true to the LMM-Joint, highlights the strength of joint models. Lastly, one striking feature, shown in Figures 2, 3 and 4, is that even with a small number of observations (30% observation percentile) from the testing unit we were still able to get accurate predictive results. This crucial in many applications as its allows early prediction of an event occurrence such as a disease or machine failure.

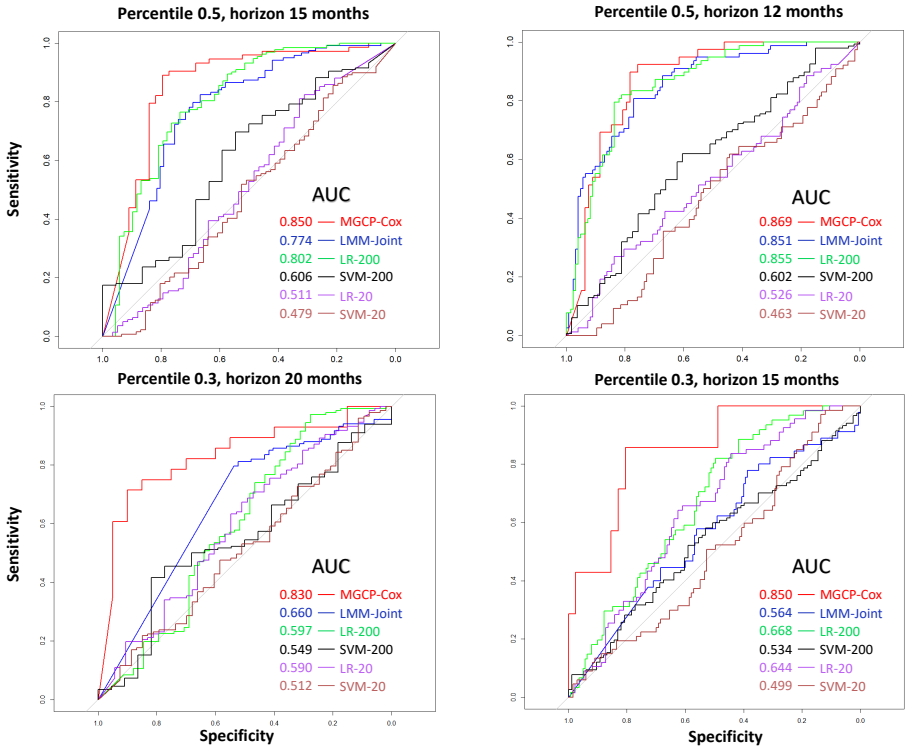

Figure 3: ROC curves from simulation studies under different percentile of observation α.

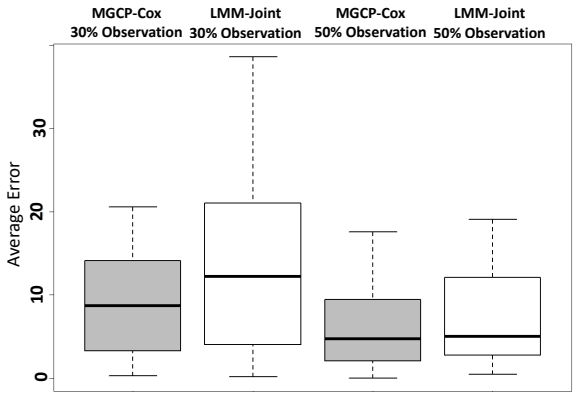

Figure 4: RL prediction accuracy from NASA data.

References

Álvarez, M. and Lawrence, N. D. (2011). Computationally efficient convolved multiple output gaussian processes. *Journal of Machine Learning Research*, 12(May):1459–1500.

Cheng, C. (2018). Multi-scale gaussian process experts for dynamic evolution prediction of complex systems. *Expert Systems with Applications*, 99:25–31.

Kim, M. and Pavlovic, V. (2018). Variational inference for gaussian process models for survival analysis. *Uncertainty in Artificial Intelligence*.

Liu, K., Gebraeel, N. Z., and Shi, J. (2013). A data-level fusion model for developing composite health indices for degradation modeling and prognostic analysis. *IEEE Transactions on Automation Science and Engineering*, 10(3):652–664.

Proust-Lima, C., Séne, M., Taylor, J. M., and Jacqmin-Gadda, H. (2014). Joint latent class models for longitudinal and time-to-event data: A review. *Statistical methods in medical research*, 23(1):74–90.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. Chapman and Hall/CRC.

Rizopoulos, D., Molenberghs, G., and Lesaffre, E. M. (2017). Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal*, 59(6):1261–1276.

Snelson, E. and Ghahramani, Z. (2006). Sparse gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, pages 1257–1264.

Soleimani, H., Hensman, J., and Saria, S. (2018). Scalable joint models for reliable uncertainty-aware event prediction. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):1948–1963.

Zhao, J. and Sun, S. (2016). Variational dependent multi-output gaussian process dynamical systems. *The Journal of Machine Learning Research*, 17(1):4134–4169.

Zhou, Q., Son, J., Zhou, S., Mao, X., and Salman, M. (2014). Remaining useful life prediction of individual units subject to hard failure. *IIE Transactions*, 46(10):1017–1030.