Multi-output Gaussian Process Modulated Poisson Processes for Event Prediction

Salman Jahani, Shiyu Zhou*, Dharmaraj Veeramani, and Jeff Schmidt

Abstract—Prediction of events such as part replacement and failure events plays a critical role in reliability engineering. Event stream data are commonly observed in manufacturing and teleservice systems. Designing predictive models for individual units based on such event streams is challenging and an under-explored problem. In this work, we propose a non-parametric prognostic framework for individualized event prediction based on the inhomogeneous Poisson processes with a multivariate Gaussian convolution process (MGCP) prior on the intensity functions. The MGCP prior on the intensity functions of the inhomogeneous Poisson processes maps data from similar historical units to the current unit under study which facilitates sharing of information and allows for analysis of flexible event patterns. To facilitate inference, we derive a variational inference scheme for learning and estimation of parameters in the resulting MGCP modulated Poisson process model. Experimental results are shown on both synthetic data as well as real-world data for fleet based event prediction.

Index Terms—Inhomogeneous Poisson processes, Multi-output Gaussian convolution processes, Gaussian process modulated Poisson process, Variational inference, Event prediction.

1 Introduction

Recent advances in information and communication technology are playing a pivotal role in enabling what is referred to as Internet of Things (IoT). An example of IoT technology is teleservice systems. In a teleservice system, the data collected from a fleet of in-field units are transmitted through the communication network to the data processing center where the aggregated data are analyzed for condition monitoring and prognosis of in-field units.

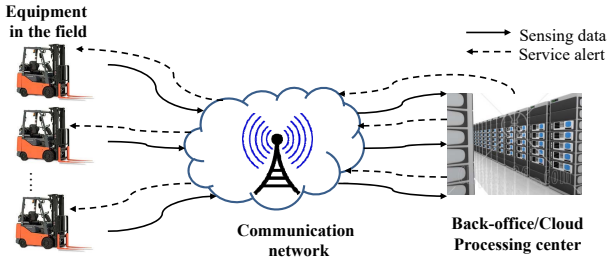

Fig. 1. The structure of a teleservice system

Through the centralized data repository, the teleservice system has access to historical off-line records of events such as part replacements and failure events from all the units. The teleservice system also receives real-time event information from the in-field units. The availability of such a rich set of historical and real-time data in a teleservice system poses significant intellectual opportunities and challenges. On opportunities, since we have observations from potentially a very large number of similar units, we can compare their event patterns, share the information, and extract some common knowledge to enable accurate prediction at the individual level. On challenges, because the data are collected in the field and not in a controlled environment, the data contains significant variation and heterogeneity due to the large variations in working conditions for different units. This requires the developed analytics methods to be stochastic in nature to account for the variations.

This work focuses on event prediction for individual units using the real-time event information collected from the unit under study as well as other units managed by the teleservice system. The event of interest occurs multiple times for each one of similar units during their lifetime. Figure 2 illustrates typical event data of two forklifts collected in a teleservice system from a warehouse. An example of the event, here, could be replacement of a part due to its failure. In the figure, we can see that the event repeatedly occurs for each forklift. The pattern of occurrence for two forklifts bears some similarity but is distinct. One of the challenges in event prediction is how to extract useful information from data collected from other units to improve the prediction for the unit under study. This setting is known as *multi-task learning*. The premise of this setting is that when multiple datasets from related outputs exist, their integrative analysis can be advantageous compared to learning multiple outputs independently. The goal of multi-task learning is to exploit commonalities between different units in order to improve the prediction and learning capabilities [1], [2]. The key feature of multi-task learning is to provide a shared representation between training and testing outputs to allow inductive transfer of knowledge. In this paper, this inductive transfer of knowledge is achieved through specifying a valid semi definite covariance function that models dependencies of all data points [3].

- *S. Jahani, S. Zhou (* Corresponding author) and D. Veeramani are with the Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI 53706, USA.*
  E-mail: jahani@wisc.edu; shiyuzhou@wisc.edu; raj.veeramani@wisc.edu
- *J. Schmidt is with the Raymond Corporation, Greene, NY 13778, USA*
  E-mail: Jeff.Schmidt@raymondcorp.com

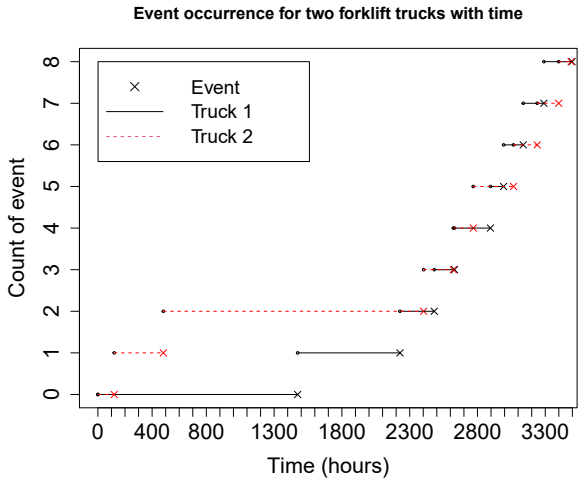

Fig. 2. Illustration of event data from material handling forklifts.

Events defined over a continuous domain arises in a variety of real-world applications including reliability analysis and event prediction for operational units/machines in connected manufacturing systems [4], disease prognosis in clinical trials [5] and events prediction using vital health signals from monitored patients at risk [6]. One thread of work in such point process data focuses on learning event intensity rates by imposing smoothness on a latent rate function [7], [8], [9]. Another consists of predicting future events as a direct function of past observations [10], [11]. Taking fleet based event prediction as a motivating example, we focus on the latter problem: given similar vehicles' history and the events history for the vehicle under-study up to time t^*, how many events will this vehicle have in $[t^*, t^* + L]$? Answering such a question provides a quantitative evaluation of the failure risk, which helps in making efficient maintenance plans and associated allocation of parts and resources.

Extensive research exists on event prediction, specially on failure prediction [12], [13]. The main avenue of research for event prediction using event data is focused on the data-driven statistical models. Data-driven statistical models typically estimate the probability of time-to-event distribution through parametric models (such as Weibull distribution) or non-parametric models [14], [15]. In this context, Cox PH regression has been widely used in clinical survival analysis and reliability engineering to investigate the effect of some covariates on the hazard rate/survival of a patient or a machine [16], [17]. Using the previously occurred events as covariates, Cox PH model has been used for event prediction using event stream data [18], [19]. The structure of the Cox PH model is very flexible such that various unit-specific factors can be incorporated as covariates in the regression. However, one limitation is that although the value of covariates can be unit specific, the model parameters are fixed and cannot reflect unit-to-unit variation. In other words, if the values of parameters for two units is the same, then the event prediction for these two units using the Cox PH model will always be the same. Also, the Cox PH regression model becomes inapplicable when good covariates are not available.

Another stream of research based on the event data uses frailty models as an extension of the conventional survival regression models, like Cox regression models, by incorporating a random effect term, typically called frailty term, to allow for unit-to-unit variation. Since the frailty term is random and follows a common distribution, the frailty model allows for unit-to-unit variation [20], [21], [22]. Frailty models have been used in reliability engineering and biomedical applications to describe the heterogeneity among the units in the population. For example, the use of frailty for modeling unobserved heterogeneity in reliability engineering using frailty model can be found in [23] and [24]. However, the majority of research on frailty modeling is predominantly focused on investigating the significance of covariates and the frailty term in the fitted model rather than prediction for the individual units. In addition, frailty model is a parametric model that often needs relatively strong assumptions.

A few works have explored the event prediction problem in point processes by learning a functional mapping from history features to the current event intensity rate [10], [11], [25]. In Gunawardana et al. [11] the intensity function is constrained to be piecewise-constant, learned using decision trees and used for events' prediction. This setting is not appropriate for the events observation where the event intensity rate (or average incidence rate) varies smoothly over time. Moreover, this modeling approach does not take into account the variation between individual units. In our proposed approach, we will consider data where the intensity of the event generating process is assumed to vary smoothly over the domain. A popular model for such data is the inhomogeneous Poisson process with a Gaussian Process (GP) prior for the smoothly varying intensity function. This form of point processes are typically known as *Cox processes* in literature [26], [27]. An example of such modeling approach is the Log Gaussian Cox Process (LGCP) where the log intensity function is driven by a GP prior [27]. The flexibility of the LGCP comes at the cost of incredibly hard inference challenges due to its doubly stochastic nature and the notorious scalability issues of GP models. Various approximations have been introduced to deal with this issue. The classic approach of Diggle [28] uses Parzen-type kernel densities to construct a nonparametric estimator, with the bandwidth chosen via the empirical Ripley's function [29]. Nonparametric Bayesian approaches have also been studied which introduced tractable finite-dimensional proxy distributions via discretization [27], [30]. There have also been nonparametric Bayesian approaches to inhomogeneous Poisson Process inference that do not use underlying Gaussian processes, e.g. Dirichlet process mixtures of Beta distributions [31] as well as approaches based on Markov Chain Monte Carlo (MCMC) [7]. Among these approximation techniques, variational inference approaches which give low-rank Gaussian process functions by augmenting a small number of inducing points has found great success in practice [9], [32], [33]. The reason is that the variational inference approach eliminates the requirement for discretization, protects against model overfitting, while simultaneously estimating the parameters of the joint Gaussian process-Poisson process model and

facilitating the scalability to large data sets.

Another difficulty with the Cox processes is that when available training data for each unit are scarce, building such predictive models for event occurrence processes is difficult. This happens because industrial equipment nowadays tend to be generally very reliable and not subject to frequent failures. To tackle this issue, here, we treat each individual event occurrence process as a task and follow a multi-task learning approach to share information from all tasks. This approach is in contrast to the general school of thought where a population level model is constructed [14], [18], [34]. Building a population model treats event prediction of different units' similarly. Such a population-level approach lacks the individualization capability where we need event predictions customized to an individual unit's history. The multi-task learning approach we propose here borrows information from the off-line historical event data and makes individualized predictions for a specific unit operating in the field.

Methods for learning GPs from multiple tasks have been proposed [35], but they typically involve a shared global mean function and require inference at all observed data points across all the tasks. In the context of Cox processes, the inference of such multi-task models become even more challenging as it is doubly stochastic in nature and involves multiple correlated tasks [36], [37], [38]. More details on double-stochasticity or double-interactability of Cox processes are given later in this study. One approximation method, variational inference, is often applied in such models leading to improvement in the scalablity. However, the inference in such predictive models still lacks individualization capability. In this paper we propose a multi-task modeling approach enabling inference at individual level while sharing information from the historical offline data set.

The main objective of this study is to provide a framework for analysis of event occurrence probability of individual units under study. One challenge is that the available training data for each unit is typically sparse. Here, we propose a multivariate Gaussian convolution process (MGCP) modulated Poisson process model which facilities sharing of information from all units through a shared latent function. The proposed framework borrows commonalities from different units and makes it possible to do inference and prediction at individual level. As mentioned before, a difficulty with building such a predictive model is that the inference is doubly-stochastic in nature and it scales poorly with the number of tasks and data points. Borrowing from the framework of the *inducing variables* or *pseudo inputs* in the GP literature [39], [40], we propose a variational inference framework to simultaneously estimate parameters in the resulting MGCP-Poisson Process (MGCP-PP) model. This facilitates the scalability and safeguards against model overfitting. Finally the advantageous feature of the proposed model is demonstrated through numerical studies and a case study with real-world data from forklift trucks' events.

The main contribution of the proposed method is to provide a non-parametric framework for individualized event prediction based on inhomogeneous Poisson processes with a MGCP prior on the intensity functions. The proposed framework models event occurrence at individual level and considers unit-to-unit variation which contrasts with the population wise modeling framework that provides event occurrence probability at a population level rather than for each individual unit. Moreover, the proposed modeling framework is highly flexible and non-parametric in nature which does not assume any specific functional form for event occurrence patterns. The MGCP prior on the individual intensity functions of inhomogeneous Poisson processes also maps data from similar historical units to the current unit under study which facilitates sharing of information and allows for analysis of flexible event patterns. Moreover, an inference framework is developed using the variational inference technique for learning and estimation of parameters that scales reasonably with the number of data points.

The remainder of this paper is organized as follows: In section 2, we provide an overview of the Cox processes. In section 3, we describe the problem formulation and inference scheme. In section 4 and section 5, we report the results of numerical studies and a real-world case study based on event data from a fleet of forklift trucks. Finally, our concluding remarks are given in section 6.

2 Gaussian Process Modulated Poisson Process

Assume data have been collected from N units and let $I = \{1, 2, \ldots, N\}$ denote the set of all units. For unit $i \in I$, its associated data is $\boldsymbol{\mathcal{D}}_i = \{t_i^{(p)}\}_{p=1}^{P_i}$ where $t_i^{(p)}$ is the time that event p occurred for unit i. Formally a Cox process –a particular type of inhomogeneous Poisson process– is defined via a stochastic intensity function $\lambda_i(t) : \mathcal{X} \to \mathbb{R}^+$ for unit $i \in I$. For a domain $\mathcal{X} = \mathbb{R}$ where $\mathbb{R}$ is the real coordinate space, the number of points, $N(\mathcal{T})$, found in a subregion $\mathcal{T} \subset \mathcal{X}$ of unit i is Poisson distributed with parameter $\lambda^i = \int_{\mathcal{T}} \lambda_i(t)dt$ and for disjoint subsets $\mathcal{T}_m$ of $\mathcal{X}$, the counts $N(\mathcal{T}_m)$ are independent. This independence is due to the completely independent nature of points in a Poisson process [26]. If we restrict our consideration to some bounded region $\mathcal{T}$, the probability density of a set of P_i observed points, $\boldsymbol{\mathcal{D}}_i$, conditioned on the rate function $\lambda_i(t)$ is

$$p(\boldsymbol{\mathcal{D}}_i|\lambda_i) = \exp\left\{-\int_{\mathcal{T}} \lambda_i(u)du\right\} \prod_{p=1}^{P_i} \lambda_i(t_i^{(p)}). \qquad (1)$$

The likelihood of observed data across all N units is $p(\boldsymbol{\mathcal{D}}|\boldsymbol{\lambda}) = \prod_{i=1}^{N} p(\boldsymbol{\mathcal{D}}_i|\lambda_i)$ where $\boldsymbol{\mathcal{D}} = \{\boldsymbol{\mathcal{D}}_i\}_{i=1}^{N}$ and $\boldsymbol{\lambda} = \{\lambda_i\}_{i=1}^{N}$ is the collection of intensity functions for all units. Using Bayes' rule, the posterior distribution of the rate functions conditioned on the data, $p_d(\boldsymbol{\lambda}|\boldsymbol{\mathcal{D}})$, is:

$$\frac{p_d(\boldsymbol{\lambda}) \prod_{i=1}^{N} \exp\{-\int_{\mathcal{T}} \lambda_i(u)du\} \prod_{p=1}^{P_i} \lambda_i(t_i^{(p)})}{\int p_d(\boldsymbol{\lambda}) \prod_{i=1}^{N} \exp\{-\int_{\mathcal{T}} \lambda_i(u)du\} \prod_{p=1}^{P_i} \lambda_i(t_i^{(p)}) d\boldsymbol{\lambda}}, \qquad (2)$$

which is often described as doubly-stochastic or doubly-intractable because of the nested integral in the denominator. Here we use the subscript d to indicate the probability density function.

To overcome the challenges posed by the doubly-intractable integral, Adams et al. [7] propose the Sigmoidal

Gaussian Cox Process (SGCP). In the proposed SGCP model, a Gaussian process prior [41] is used to construct an intensity function by passing a random function, $f \sim \mathcal{GP}$, through a sigmoid transformation and scaling it with a maximum intensity λ^*. The intensity function is therefore $\lambda(t) = \lambda^* \sigma\left(f(t)\right)$, where $\sigma(.)$ is the logistic function

$$\sigma(x) = \frac{1}{1+\exp(-x)}. \tag{3}$$

In the SGCP approach, the space is discretized and the variable set is augmented to include latent data such that the joint distribution of the latent and the observed data is uniform Poisson. However, this model scales poorly with the dimensionality of the domain and the maximum observed density of points. This is mainly due to the incorporation of latent or thinned data whose number grows exponentially with the dimensionality of the space. Moreover, it only considers the event data from the in-field unit and does not incorporate the off-line information from the historical units stored in the data repository. In order to tackle the issue of scalability, Lloyd et al. [9] propose to use a variational inference scheme. They assumed that the intensity is defined as $\lambda(t) = f^2(t)$ where $f \sim \mathcal{GP}$ is a GP distributed random function. This approach, also, falls short of considering the information that comes from the peer units in making inference and prediction, and only relies on the event data from the in-field unit. Moreover, it requires integrating the square of a Gaussian process over a definite region to achieve the model evidence. The integration of $f^2(t)$ can be done using numerical techniques but it can result in poor numerical stability. Moreover, the square transformation is not a one-to-one function where any rate function may have been generated by $f^2(t)$ or $\big(-f(t)\big)^2$.

In this study we use a multi-task modeling approach to model the intensity functions of different units. This approach takes advantage of the multi-output GPs to share information between units from offline historical data and the online in-field unit via a shared latent function. Specially, we model the individual latent log intensity function $\log \lambda(t) = f(t)$ with a multivariate GP prior. This approach requires no discretization of the space, makes sharing of information through multi-variate GP prior possible while giving closed form formula for inference and prediction.

A GP is formally defined as a collection of random variables, any finite number of which have consistent joint Gaussian distributions. For any input point $t \in \mathcal{X} \subset \mathbb{R}$, observations from a random dataset $\boldsymbol{f}(t) = \{f(t_1), f(t_2), ..., f(t_p)\}^T$ are considered as single sample from some multivariate Gaussian distribution. Thus, the GP can be expressed as $f(t) \sim \mathcal{GP}(0, \Sigma(t,t'))$, where $\Sigma(t,t')$ is a positive definite covariance function. An alternative approach for constructing a Gaussian process is to convolve the GP random variables with an arbitrary kernel. Thus, $f(t)$ can be expressed as the convolution between a smoothing kernel $G(t)$ and a latent function $X(t)$ as follows:

$$f(t) = \int_{\mathbb{R}} G(t-u)X(u)du. \tag{4}$$

The resulting covariance function for $f(t)$ is then derived as

$$\mathrm{cov}_f(t,t') = \int_{\mathbb{R}} G(t-u) \int_{\mathbb{R}} G(t'-u')\kappa(u,u')du'du \tag{5}$$

where $\mathrm{cov}[X(t), X(t')] = \kappa(t,t')$ is the covariance defining the latent function $X(t)$. We note that this construction is general in the sense that $X(t)$ can be any GP random variable [42]. Therefore, the covariance matrix can be directly parametrized through parameters in the smoothing kernel. In this article we employ Convolution Process (CP) to build covariance functions that model dependencies within and across units. The basic idea is to build multiple GPs where all outputs depends on some common latent processes. The proposed framework can provide each output with both shared and unique features and allows commonalities between different outputs to be automatically inferred. We introduce our multi-task modeling approach which takes advantage of the MGCPs in modeling the intensity of inhomogeneous Poisson processes in the next section. We also introduce a variational inference approach to make inference in the resulting MGCP modulated Poisson processes. The modeling framework introduced in this study makes inference and prediction for the individual in-field unit possible while tackling the sparsity in the observed event data.

3 Construction of Multi-output Gaussian Process Modulated Poisson Process

We construct our prior over the individual rate functions using GPs and assume that the resulting Cox process is driven by a latent log intensity function $\log \lambda_i := f_i$ with a GP prior:

$$f_i(t) \sim \mathcal{GP}\left(0, \Sigma_i(t,t')\right). \tag{6}$$

To obtain an accurate predictive result, we need to capture relatedness among all N units. Particularly, we use CP as mentioned in section 2 to model the latent log intensity functions $f_i(t)$ for each unit $i \in I$. We can consider a shared independent latent function $X(t)$ and N different smoothing kernels $G_i(t) : i = 1, ..., N$. The latent function is assumed a GP with covariance $\mathrm{cov}[X(t), X(t')] = \kappa(t,t')$. We set the kernels as

$$G_i(t) = \frac{a_i \pi^{-\frac{1}{4}}}{\sqrt{|\xi_i|}} \exp(-\frac{1}{2}\frac{t^2}{\xi_i^2}) := \alpha_i \mathcal{N}(t; 0, \xi_i^2), \tag{7}$$

to be scaled Gaussian kernels where $\mathcal{N}(t; 0, \xi_i^2)$ is the density function of a zero mean normal distribution with variance ξ_i^2. We also consider $\kappa(t,t')$ to be the squared exponential covariance function [42] as follows:

$$\begin{aligned} \kappa(t,t') &= \exp\left[-\frac{1}{2}\frac{(t-t')^2}{\lambda^2}\right] \\ &= \sqrt{2\pi\lambda^2}\mathcal{N}(d; 0, \lambda^2) := C\mathcal{N}(d; 0, \lambda^2), \end{aligned} \tag{8}$$

The GP $f_i(t)$ is then constructed by convolving the shared latent function with the smoothing kernel as follows:

$$f_i(t) = \int_{\mathbb{R}} G_i(t-u)X(u)du. \tag{9}$$

This is the underlying principle of MGCP, where the latent functions $X(t)$ is shared across different units through

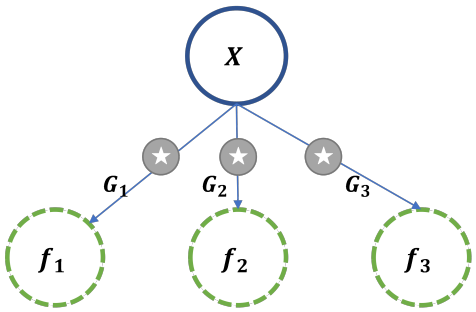

Fig. 3. A convolution process with one latent functions

the corresponding kernel $G_i(t)$. Since the model in Eq. (9) shares the latent function, a GP, across multiple units and since convolution is a linear operator, all outputs can be expressed as a jointly distributed GP. Figure 3 shows an illustration of such a convolution structure. As shown in figure 3, the key feature is that information is shared through parameters encoded in the kernels $G_i(t)$. Outputs then possess both unique and shared features; thus, accounting for heterogeneity in the intensity functions. It should be mentioned that the shared latent function here does not have a specific physical meaning and rather is a mathematical approach to induce correlation between different separately evolving processes [32], [39], [41].

Based on equation (9), the covariance function between f_i and f_j and the covariance function between f_i and X, can be calculated as follows:

$$\begin{aligned}\mathrm{cov}_{f_i,f_j}(t,t') &= \int_{\mathbb{R}} G_i(t-u)\int_{\mathbb{R}} G_i(t'-u')\kappa(u,u')du'du \\ &= \alpha_i\alpha_j\sqrt{\frac{\lambda^2}{\eta_{i,j}^2}}\exp(-\frac{1}{2}\frac{(t-t')^2}{\eta_{i,j}^2}), \\ \mathrm{cov}_{f_i,X}(t,u) &= \int_{\mathbb{R}} G_i(t-u')\kappa(u,u')du' \\ &= \alpha_i\sqrt{\frac{\lambda^2}{\eta_i^2}}\exp(-\frac{1}{2}\frac{(t-u)^2}{\eta_i^2}),\end{aligned} \tag{10}$$

where $\eta_{i,j}^2 = \xi_i^2 + \xi_j^2 + \lambda^2$ and $\eta_i^2 = \xi_i^2 + \lambda^2$. Now denote the underlying latent log intensity rates at the input data points as $\boldsymbol{f} = \{\boldsymbol{f}_1^T, ..., \boldsymbol{f}_N^T\}^T$, where $\boldsymbol{f}_i = \{f_i(t_i^{(1)}), ..., f_i(t_i^{(p_i)})\}^T$. The density function of $\boldsymbol{f}$ can be obtained as $p_d(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}; \boldsymbol{0}, \boldsymbol{K_{f,f}})$, where $\boldsymbol{K_{f,f}}$ sized $(\sum_{i=1}^N p_i) \times (\sum_{i=1}^N p_i)$ is the covariance function. More details about the properties of convolution process in (10) can be found in [41], [43], [44], [45], [46]

Exact inference in the proposed model entails optimizing the *model evidence* $p(\boldsymbol{\mathcal{D}}) = \mathbb{E}_{p(\boldsymbol{f})}[p(\boldsymbol{\mathcal{D}}|\boldsymbol{\lambda} = \exp(\boldsymbol{f}))]$ for which the marginal log-likelihood can be obtained as follows:

$$\log p(\boldsymbol{\mathcal{D}}) = \log \int p\left(\boldsymbol{\mathcal{D}}|\boldsymbol{\lambda} = \exp(\boldsymbol{f})\right) p_d(\boldsymbol{f}) d\boldsymbol{f}, \tag{11}$$

where as noted before $p_d(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}; \boldsymbol{0}, \boldsymbol{K_{f,f}})$. The likelihood of $\boldsymbol{f}$ in Eq. (11) involves inversion of the large matrix $\boldsymbol{K_{f,f}}$ which has a limiting cubic complexity $O\left(\left(\sum_{i=1}^N p_i\right)^3\right)$ and is in general intractable. Moreover, as mentioned in section 2, we see that the log-likelihood is doubly-stochastic as it also involves an integration over the latent log intensity functions (see Eq. (1) and Eq. (11)). This, in turn, makes the exact inference more challenging. To alleviate the computation burden of matrix inversion, low-rank Gaussian process functions can be constructed by *augmenting* the Gaussian process with a small number of M inducing points or pseudo-inputs from the shared latent function [44], [47], [48]. In next subsection, we introduce a variational inference framework based on the inducing points which tackles the double-stocasticity of Eq. (11) by obtaining a lower bound on the model evidence.

3.1 Variational Inference

We denote the inducing points by $\boldsymbol{\mathcal{Z}} = \{z_i\}_{i=1}^M$ and the value of shared latent function at the inducing points by $\boldsymbol{X} = [X(z_1), ..., X(z_M)]^T$. Since the latent function is GP, any sample $\boldsymbol{X}$ follows a multivariate Gaussian distribution. Therefore, the probability distribution of $\boldsymbol{X}$ can be expressed as $p_d(\boldsymbol{X}|\boldsymbol{\mathcal{Z}}) = \mathcal{N}(\boldsymbol{X}; \boldsymbol{0}, \boldsymbol{K_{X,X}})$, where $\boldsymbol{K_{X,X}}$ is constructed by the covariance function in equation (8). We now can sample from the conditional prior $p(X(u)|\boldsymbol{X}, \boldsymbol{\mathcal{Z}})$. In equation (9) where we construct latent intensity function $f_i(t)$, $X(u)$ can be well approximated by the expectation $\mathbb{E}(X(u)|\boldsymbol{X}, \boldsymbol{\mathcal{Z}})$ as long as the latent function is smooth [42]. By multivariate Gaussian identities [39], [41], [42], the probability distribution of $\boldsymbol{f}$ conditional on $\boldsymbol{X}$, $\boldsymbol{\mathcal{Z}}$ is:

$$\begin{aligned}p_d(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\mathcal{Z}}) =& \mathcal{N}(\boldsymbol{f}; \boldsymbol{K_{f,X}}\boldsymbol{K_{X,X}^{-1}}\boldsymbol{X}, \\ & \boldsymbol{K_{f,f}} - \boldsymbol{K_{f,X}}\boldsymbol{K_{X,X}^{-1}}\boldsymbol{K_{X,f}}),\end{aligned} \tag{12}$$

where $\boldsymbol{K_{X,X}}$ is the covariance matrix between the inducing variables and $\boldsymbol{K_{f,X}}$ is the covariance matrix between the latent log intensity values and the inducing variables. Therefore, $p_d(\boldsymbol{f})$ can be approximated by $p_d(\boldsymbol{f}|\boldsymbol{\mathcal{Z}})$, which is given as:

$$p_d(\boldsymbol{f}|\boldsymbol{\mathcal{Z}}) = \int p_d(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\mathcal{Z}}) p_d(\boldsymbol{X}|\boldsymbol{\mathcal{Z}}) d\boldsymbol{X}. \tag{13}$$

By equation (13), the marginal log-likelihood function can be approximated as follows:

$$\begin{aligned}&\log p(\boldsymbol{\mathcal{D}}) = \log \int p\left(\boldsymbol{\mathcal{D}}|\boldsymbol{\lambda} = \exp(\boldsymbol{f})\right) p_d(\boldsymbol{f}) d\boldsymbol{f} \\ &\approx \log \int\int p\left(\boldsymbol{\mathcal{D}}|\boldsymbol{\lambda} = \exp(\boldsymbol{f})\right) p_d(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\mathcal{Z}}) p_d(\boldsymbol{X}|\boldsymbol{\mathcal{Z}}) d\boldsymbol{X} d\boldsymbol{f}\end{aligned} \tag{14}$$

We next continue by integrating out the inducing variables $\boldsymbol{X}$, using a variational distribution $q_d(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{X}; \boldsymbol{m}, \boldsymbol{S})$ over the inducing points. We then multiply and divide the joint by $q_d(\boldsymbol{X})$ and lower bound using Jensen's inequality to obtain a lower bound on the model evidence:

$$\begin{aligned}\log p(\boldsymbol{\mathcal{D}}) &= \log\left[\int\int p(\boldsymbol{\mathcal{D}}|\boldsymbol{f})p_d(\boldsymbol{f}|\boldsymbol{X})p_d(\boldsymbol{X})\frac{q_d(\boldsymbol{X})}{q_d(\boldsymbol{X})}d\boldsymbol{X}d\boldsymbol{f}\right]\\ &\geq \int\int p_d(\boldsymbol{f}|\boldsymbol{X})q_d(\boldsymbol{X})d\boldsymbol{X}\log(p(\boldsymbol{\mathcal{D}}|\boldsymbol{f}))d\boldsymbol{f}\\ &+\int\int p_d(\boldsymbol{f}|\boldsymbol{X})q_d(\boldsymbol{X})d\boldsymbol{f}\log(\frac{p_d(\boldsymbol{X})}{q_d(\boldsymbol{X})})d\boldsymbol{X}\\ &= \mathbb{E}_{q_d(\boldsymbol{f})}\left[\log p(\boldsymbol{\mathcal{D}}|\boldsymbol{f})\right] - KL(q_d(\boldsymbol{X})\parallel p_d(\boldsymbol{X})) \triangleq \mathcal{L}\end{aligned} \tag{15}$$

Since $p_d(\boldsymbol{f}|\boldsymbol{X})$ is conjugate to $q_d(\boldsymbol{X})$, we can write down in closed form the resulting integral:

$$\begin{aligned}q_d(\boldsymbol{f}) &= \int p_d(\boldsymbol{f}|\boldsymbol{X})q_d(\boldsymbol{X})d\boldsymbol{X} := \mathcal{N}(\boldsymbol{f};\boldsymbol{\mu},\boldsymbol{\Sigma})\\ \boldsymbol{\mu} &= \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{X}}\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1}\boldsymbol{m}\\ \boldsymbol{\Sigma} &= \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{f}} - \boldsymbol{K}_{\boldsymbol{f},\boldsymbol{X}}\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1}(\boldsymbol{I}-\boldsymbol{S}\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1})\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{f}}.\end{aligned} \tag{16}$$

Here, $KL(q_d(\boldsymbol{X}) \parallel p_d(\boldsymbol{X}))$ is simply the KL-divergence between two Gaussians:

$$\begin{aligned}KL(q_d(\boldsymbol{X})\parallel p_d(\boldsymbol{X})) =& \frac{1}{2}\Big[\mathrm{Tr}(\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1}\boldsymbol{S}) - \log\frac{|\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}|}{|\boldsymbol{S}|}\\ &-M + (\vec{0}-\boldsymbol{m})^T\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1}(\vec{0}-\boldsymbol{m})\Big],\end{aligned} \tag{17}$$

where $\mathrm{Tr}(.)$ is a trace operator. We now take expectation of data log-likelihood under $q_d(\boldsymbol{f})$:

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_{q_d(\boldsymbol{f})}\left[\log p(\boldsymbol{\mathcal{D}}|\boldsymbol{f})\right] - KL(q_d(\boldsymbol{X})\parallel p_d(\boldsymbol{X}))\\ &= \mathbb{E}_{q_d(\boldsymbol{f})}\left[-\sum_{i=1}^{N}\int_{\mathcal{T}}\exp(f_i(u))du + \sum_{i=1}^{N}\sum_{p=1}^{P_i}f_i(t_i^{(p)})\right]\\ &- KL(q_d(\boldsymbol{X})\parallel p_d(\boldsymbol{X}))\end{aligned} \tag{18}$$

The first term in (18) can be estimated by the moment generating function (MGF) and the numerical integration:

$$\begin{aligned}&\mathbb{E}_{q_d(\boldsymbol{f})}\left[\sum_{i=1}^{N}\int_{\mathcal{T}}\exp(f_i(u))du\right]\\ &= \sum_{i=1}^{N}\int\int q_d(\boldsymbol{f})\exp(f_i(u))d\boldsymbol{f}du\\ &= \sum_{i=1}^{N}\int\exp(\mu_i(u)+\frac{1}{2}\sigma_i^2(u))du\end{aligned} \tag{19}$$

where $\mu_i(u) := \boldsymbol{K}_{f_i(u),\boldsymbol{X}}\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1}\boldsymbol{m}$ and $\sigma_i^2(u) := \boldsymbol{K}_{f_i(u),f_i(u)} - \boldsymbol{K}_{f_i(u),\boldsymbol{X}}\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1}(\boldsymbol{I}-\boldsymbol{S}\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1})\,\boldsymbol{K}_{\boldsymbol{X},f_i(u)}$. The second term in (18) can be calculated based on the definition of Gaussian processes as follows:

$$\mathbb{E}_{q(\boldsymbol{f})}\left[\sum_{i=1}^{N}\sum_{p=1}^{P_i}f_i(t_i^{(p)})\right] = \sum_{i=1}^{N}\sum_{p=1}^{P_i}\mu_i(t_i^{(p)}) \tag{20}$$

where $\mu_i(t_i^{(p)}) = \boldsymbol{K}_{t_i^{(p)},\boldsymbol{X}}\boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1}\boldsymbol{m}$. To perform inference one needs to take following two steps:

1) A small number of M inducing points, distributed equally, from the shared latent function X should be defined.
2) The lower bound $\mathcal{L}$ constructed through Eqs. (15)-(20) should be maximized to find the variational parameters $\boldsymbol{m}^*$, $\boldsymbol{S}^*$ and the model parameters $\boldsymbol{\theta}^* = (\lambda, \{\xi_i, \alpha_i\}_{i=1}^N)$.

To optimize these parameters simultaneously, we construct an augmented vector $\boldsymbol{\Theta} = \left[\boldsymbol{\theta}, \boldsymbol{m}^T, vech(\boldsymbol{L})^T\right]$ where $vech(\boldsymbol{L})$ is the vectorization of the lower triangular elements of $\boldsymbol{L}$, such that $\boldsymbol{S} = \boldsymbol{L}\boldsymbol{L}^T$. The $vech(.)$ operator is a linear transformation which converts a matrix into a column vector. Any multivariate optimization algorithm such as Nelder-Mead simplex, conjugate gardient, Adam optimization, and etc. can be used to optimize the lower bound $\mathcal{L}$. It should be noted that, in principal, any number of inducing points can be considered for a given set of parameters $\boldsymbol{\theta}$; however, the proper value can be tuned using cross-validation [9], [39], [40].

3.2 Predictive Distribution

In this section, we derive the predictive distribution for the test unit N based on the optimized $\boldsymbol{\Theta}^*$. Our training data (denoted as $\boldsymbol{\mathcal{D}}$) includes the observations from the offline units $i = 1, 2, ..., N-1$ as well as the partial observations from the online test unit N.

Suppose observations from the test unit N have been collected up to time t^*. We can next derive the predictive distribution for any new input time T of the test unit N. In order to form the predictive distribution we assume our optimised variational distribution $q_d^*(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{X};\boldsymbol{m}^*,\boldsymbol{S}^*)$ approximates the posterior $p_d(\boldsymbol{X}|\boldsymbol{\mathcal{D}})$. This assumption can be made because the variational inference appraoch minimizes the KL-divergence distance between the true distribution $p_d(\boldsymbol{X}|\boldsymbol{\mathcal{D}})$ and the variational distribution $q_d^*(\boldsymbol{X})$. Similar to Equation (16), we next compute $q_d^*(\boldsymbol{f}) \approx p_d(\boldsymbol{f}|\boldsymbol{\mathcal{D}})$. We can now derive a lower bound of the (approximate) predictive log-likelihood for unit N in any new input time T:

$$\begin{aligned}\log p(T|\boldsymbol{\mathcal{D}},\boldsymbol{\Theta}^*) &= \log\mathbb{E}_{p_d(\boldsymbol{f}|\boldsymbol{\mathcal{D}})}[p(T|\boldsymbol{f})]\\ &\approx \log\mathbb{E}_{q_d^*(\boldsymbol{f})}[p(T|\boldsymbol{f})]\\ &\geq \mathbb{E}_{q_d^*(\boldsymbol{f})}[\log p(T|\boldsymbol{f})] \triangleq \mathcal{L}_p\end{aligned} \tag{21}$$

The derivation of $\mathcal{L}_p$ follows Equations (18)-(20). The resulting bound is similar to $\mathcal{L}$ except that $\boldsymbol{m}$, $\boldsymbol{S}$ are replaced with $\boldsymbol{m}^*$ and $\boldsymbol{S}^*$, and there is no KL-divergence term. All the kernel matrices are computed using $\boldsymbol{\Theta}^*$. We use this bound to give results from approximate predictive likelihood when comparing against other approaches.

We can now answer the question posed at the beginning of this study using the derived predictive log-likelihood. The question involves estimating the distribution of event occurrence in $[t^*, t^*+L]$ for the test unit N. This event occurrence probability distribution for unit N depends on the predicted latent log intensity function $\log\lambda_N := f_N(u), u \in [t^*, t^*+L]$. Given the predicted intensity rate $\lambda^{N^*} = \int_{t^*}^{t^*+L}\lambda_N(t)dt$, the event occurrence has Poisson distribution. Given the estimated parameters, we are interested in:

$$p_d(N(t^*) = y) = \frac{e^{-\lambda^{N^*}}\lambda^{N^*y}}{y!},\ y = 0, 1, 2, ..., \tag{22}$$

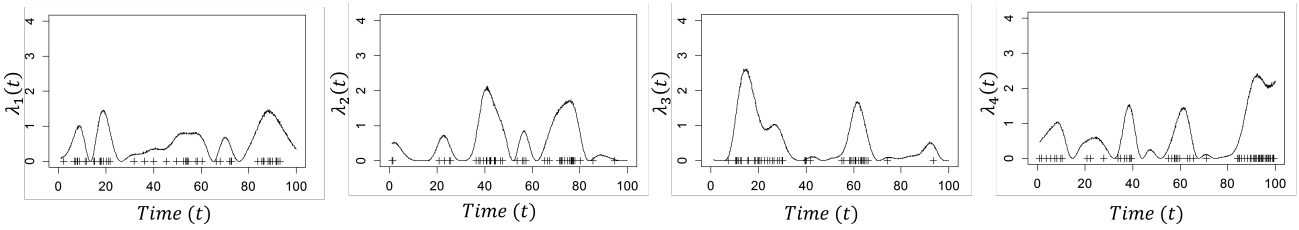

Fig. 4. A sample of intensity rates generated from MGCP and the sigmoid link function.

where y is the number of events. Based on (22), the accurate probability of event occurrence depends on the extrapolation of the intensity rate within L for the testing unit N. In the MGCP, the predictive distribution for any new input point T^* is given by:

$$\begin{aligned} p_d(f_N(T^*)|\boldsymbol{\mathcal{D}}) &= \int p_d(f_N(T^*)|\boldsymbol{X}) p_d(\boldsymbol{X}|\boldsymbol{\mathcal{D}}) d\boldsymbol{X} \\ &\approx \int p_d(f_N(T^*)|\boldsymbol{X}) q_d(\boldsymbol{X}) d\boldsymbol{X} \\ &= \mathcal{N}\Big(f_N(T^*); \boldsymbol{K}_{f_N(T^*),\boldsymbol{X}} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1} \boldsymbol{m}^*, \\ &\boldsymbol{K}_{f_N(T^*),f_N(T^*)} - \boldsymbol{K}_{f_N(T^*),\boldsymbol{X}} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1} \times \\ &(\boldsymbol{I} - \boldsymbol{S}^* \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1}) \boldsymbol{K}_{\boldsymbol{X},f_N(T^*)}\Big) \end{aligned} \tag{23}$$

where we assumed our optimized variational distribution $q^*(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{X}; \boldsymbol{m}^*, \boldsymbol{S}^*)$ approximates the posterior $p_d(\boldsymbol{X}|\boldsymbol{\mathcal{D}})$. We used $\boldsymbol{K}_{f_N(T^*),f_N(T^*)}$ as a notation when the covariance matrix is evaluated at T^*. Consequently, the predictions at the time point T^* for unit N is $\hat{f}_N(T^*) = \boldsymbol{K}_{f_N(T^*),\boldsymbol{X}} \boldsymbol{K}_{\boldsymbol{X},\boldsymbol{X}}^{-1} \boldsymbol{m}^*$. It should be noted that the results in Eq. (23) is a direct consequence of the conditional distribution from the joint distribution of $\{\boldsymbol{\mathcal{D}}, f_N(T^*)\}$. This, indeed, happens because we assume the observed data and the unseen data in the future have a joint Gaussian process distribution.

4 Experiments

In this section, the performance of our proposed methodology, denoted as MGCP-PP is investigated. We benchmark the prediction performance of our proposed framework using both synthetic and real-world data. Specifically, we benchmark the performance against Variational Bayes for Point Processes (VBPP) approach of Lloyd et al. [9] and the Sigmoidal Gaussian Cox Process (SGCP) of Adams et al. [7] which are based on considering a univariate Gaussian process prior for the intensity rate. Unlike our proposed approach, the methods discussed in [9] and [7] do not consider the cross-correlation that exists between different units. Regarding our MGCP-PP model we set the number of pseudo-inputs to $M = 10$. Throughout this section we consider $N = 10$ units and it is assumed that the observations are made in $t \in [0, 100]$.

4.1 Data Setting

For the synthetic dataset, we simulate the underlying latent functions $f_i(t), i = 1, ..., 10$, using MGCPs and generate the intensity rates of different units using a sigmoid link function. Then conditioned on this function, we draw training datasets and test datasets [7]. The number of units generated is $N = 10$ where we pick the Nth unit as the testing unit. This experiment is repeated for $Q = 1000$ times and we report the average prediction performance of the test dataset for each approach. Figure 4 visualizes a sample of intensity rates drawn from the defined MGCP model with four outputs passed through a sigmoid link function along with the simulated events based on the method discussed in [7]. We note that here we only have four outputs for the purpose of illustration while the actual simulation study is done using MGCP model with $N = 10$ outputs. We note that data generated using the MGCP prior has an inherent correlation. Moreover, the intensity functions generated in this case study are heterogeneous as also shown in figure 4; therefore, it can provide a measure for flexibility of different approaches.

In addition to generating the intensity rates using MGCP and sigmoid link function, we define two parametric functional forms to generate the intensity rates as follows:

1) A sum of an exponential and a Gaussian bump $\lambda_i(x) = a \exp(-\frac{x}{b}) + \exp(-(\frac{x-c}{15})^2)$ where $[a, b, c]^T \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ with $\boldsymbol{\mu}_1 = [3, 20, 65]^T$ and $\boldsymbol{\Sigma}_1 = \begin{bmatrix} 5e-1 & 4e-4 & -e-5 \\ 4e-4 & 2.5e-1 & 3e-7 \\ e-5 & 3e-7 & 1 \end{bmatrix}$.

2) A sinusoid with increasing frequency: $\lambda_i(x) = a' \sin(b' x^2) \exp(-\frac{x}{c'}) + 1$ where $[a', b', c']^T \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ with $\boldsymbol{\mu}_2 = [2, 2e - 3, 50]^T$ and $\boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & -e-7 & 2e-4 \\ -e-7 & e-2 & 3e-7 \\ e-5 & 3e-7 & 1 \end{bmatrix}$.

We, also, generate the training and testing data conditioned on these functional forms in the simulation study. It should be noted that the intensity rate and consequently the event data generated using the first parametric form vary smoothly over the time, while the second parametric form generates data that has more seasonality (See figure 5 and [7], [9] for reference).

4.2 Results

We compare different approaches in terms of predictive log-likelihood (LL) and root mean squared error (RMSE) between the predicted intensity rate and the true intensity rate of the testing unit N. It should be noted that models with higher predictive log-likelihood or lower root mean squared error have better prediction performance [7], [9]. Prediction performance at varying time points t^* for the partially observed unit N is reported. The time instant

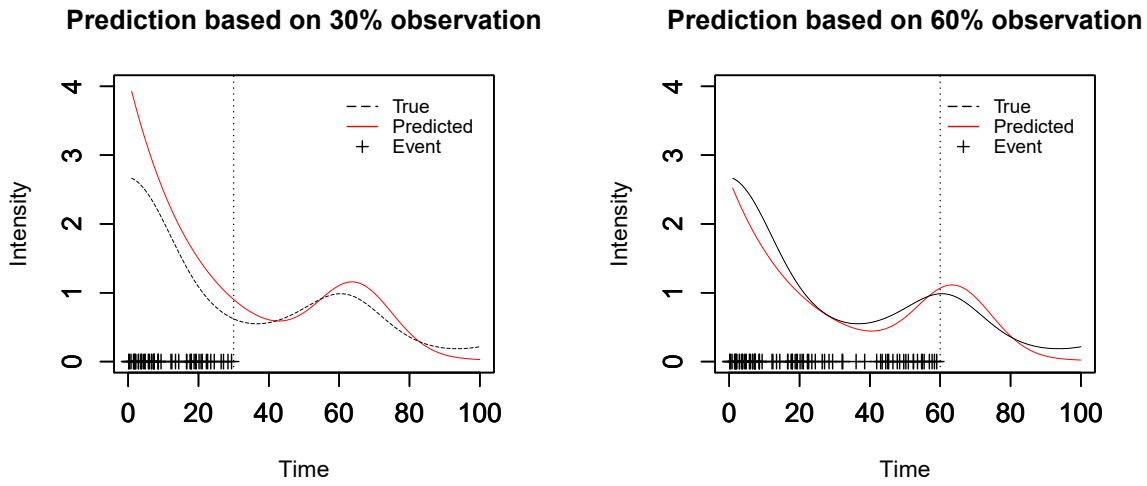

Fig. 5. Prediction performance for testing unit in different observation percentiles

$t^* = \alpha \times 100$ is defined as the α-observation percentile of the testing unit N. The values of α is specified as 30% and 60% in the simulations. Figure 5 illustrates an example of the unit observed up to different percentiles of its life. The intensity rates here are generated using the first parametric functional form explained above. The illustrative example in Figure 5 demonstrates the behavior of our method. As can be seen from the figure, our joint modeling framework can provide accurate prediction of the true intensity rate for the testing unit N. It is mainly because of the flexible convolution structure considered for the MGCP approach that makes sharing of information possible among different units. The unique smoothing kernel G_i for each individual allows flexibility in the prediction as it enables each training signal to have its own characteristics. This indeed substantiates the strength of the MGCP. Using the shared latent processes, the model can infer the similarities among all units and predict the intensity rate for the testing unit more accurately by borrowing strength from the training data.

The results in figures 6, 7 and 8 indicates that our MGCP-PP model clearly outperforms the benchmarked models. Based on the figures we can get some important insights. First, as expected, the prediction errors decrease as the lifetime percentile increase for the testing unit N. Thus, the prediction accuracy from the MGCP-PP becomes more accurate as t^* increases and more data are collected from the online monitoring unit. Second, we can observe that in the first simulation study, where the data are generated using MGCP and the sigmoid link function, the SGCP approach gives better predictive performance than the VBPP; however, our MGCP-PP model approach always remains superior as it takes advantage of the pool of historical offline units in making inference for the online unit under consideration. The reason that the SGCP approach performs better than the VBPP here can be attributed to the fact that the SGCP uses the same link function and the generative process which results in well-tuned hyper-parameters. Moreover the results in figure 6 show that when the intensity rates are heterogeneous,

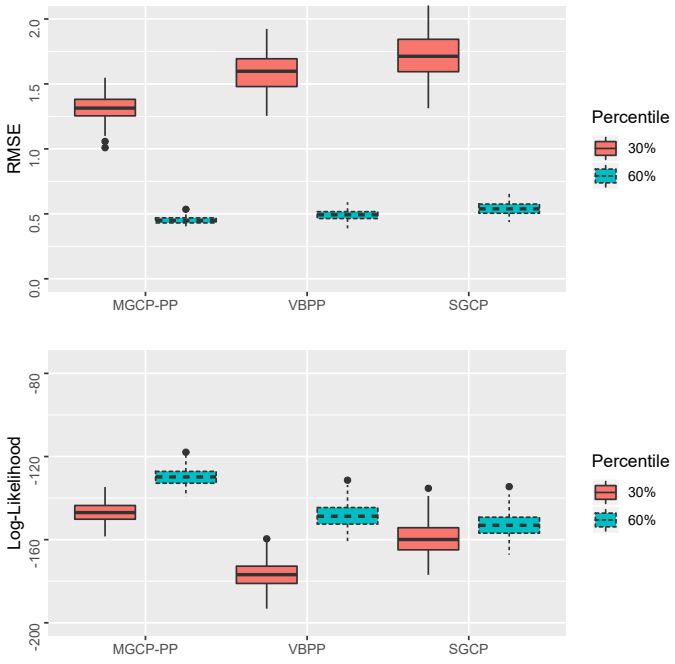

Fig. 6. Simulation study results with MGCP and the sigmoid link function

the MGCP-PP model that takes advantage of a flexible MGCP prior which does not assume any functional form on the intensity functions outperforms other competing approaches in terms of RMSE and predictive log-likelihood. The results in figure 7 and figure 8 demonstrate the performance of three approaches when the intensity functions vary smoothly (first functional form) or has seasonality (second functional form). It can be observed that the proposed model in this study which shares information between the testing unit and the units in historical dataset performs relatively better than other approaches that does not include sharing of information. Lastly, one striking feature shown in figures 6, 7 and 8, is that even with a small number of observations (30% observation percentile) from

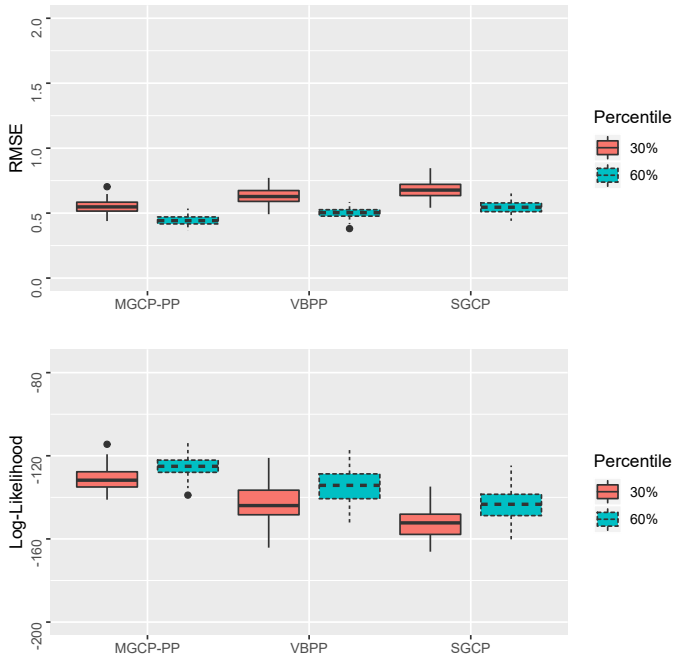

Fig. 7. Simulation study results with parametric functional form 1

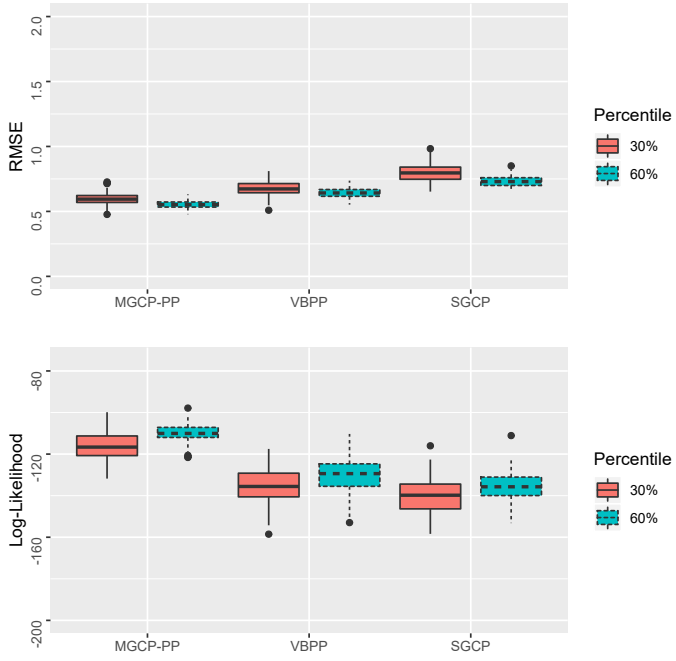

Fig. 8. Simulation study results with parametric functional form 2

the testing unit we are still able to get accurate prediction results. This is crucially important in many applications, specially when observed data are sparse, as it allows early prediction of an event occurrence such as part replacements.

5 Real-world case study

In this section, application of the proposed procedure on the real-world data for fleet based event prediction is demonstrated. The data is collected from material handling forklift trucks and events of interest are part replacements captured in real-time. Predicting the occurrence of recurrent part failures is of paramount importance in industry, specially when multiple units are operating simultaneously. With the advent of IoT teleservice systems, the event occurrence data are collected through sensors mounted on the equipment and are transmitted to a back office for analysis and inference. This huge amount of offline collected data provides an opportunity for accurate event prediction for on-line units operating in the field. The accuracy of event prediction plays a critical role in part procurement and maintenance planning. A critical factor in part replacement cost is the setup cost that is due to mobilizing repair crew, safety provision, special transportation, disassembling machines, and etc. These costs are shared by all the maintenance activities performed close to each other. Planning for part procurement and maintenance depends on the accurate prediction of event occurrence which, in turn, leads to considerable cost saving.

In the case study considered here, we have information regarding the occurrence of the event of interest from 20 trucks where the number of events for each trucks lies in the range of 6-23. The actual calendar time is adjusted for each unit, i.e. the starting time is made zero for all the units. Figure 2 illustrates the data collected from forklift trucks collected in a teleservice system for warehouse material handling equipment. Please note that the time axis is the lifetime of the forklifts, not the calendar time.

The models MGCP-PP, VBPP, and SGCP are fitted on the case study data. To perform a comprehensive performance evaluation we use a leave-one-out cross-validation approach. First, we exclude one of the 20 units as the testing unit N and the rest are used as the training units. Prediction for testing unit N is then performed at 50% lifetime percentile. The whole procedure is repeated 20 times and the predictive performance of different approaches as a function of prediction window L is illustrated in figure 9 .

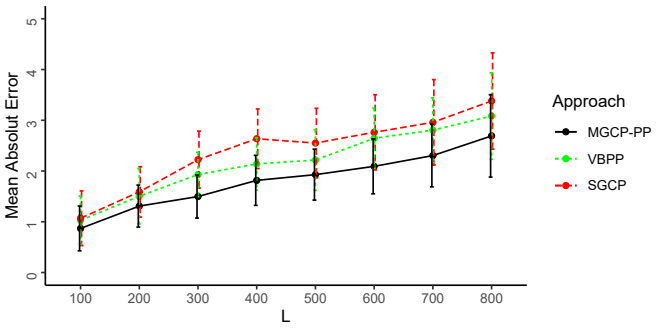

Fig. 9. MAE of event occurrence in $[t^*, t^* + L]$.

Figure 9 shows the Mean Absolute Error (MAE) of the event occurrence counts in $[t^*, t^* + L]$. The MAE of all methods increases monotonically as prediction window length increases. The MGCP-PP approach outperforms the SGCP and VBPP approaches that are based on the univariate Gaussian processes. On average, MGCP-PP improves prediction accuracy in terms of MAE by %16.34 compared to VBPP and by %23.92 compared to SGCP. This indeed highlights the importance of borrowing information from

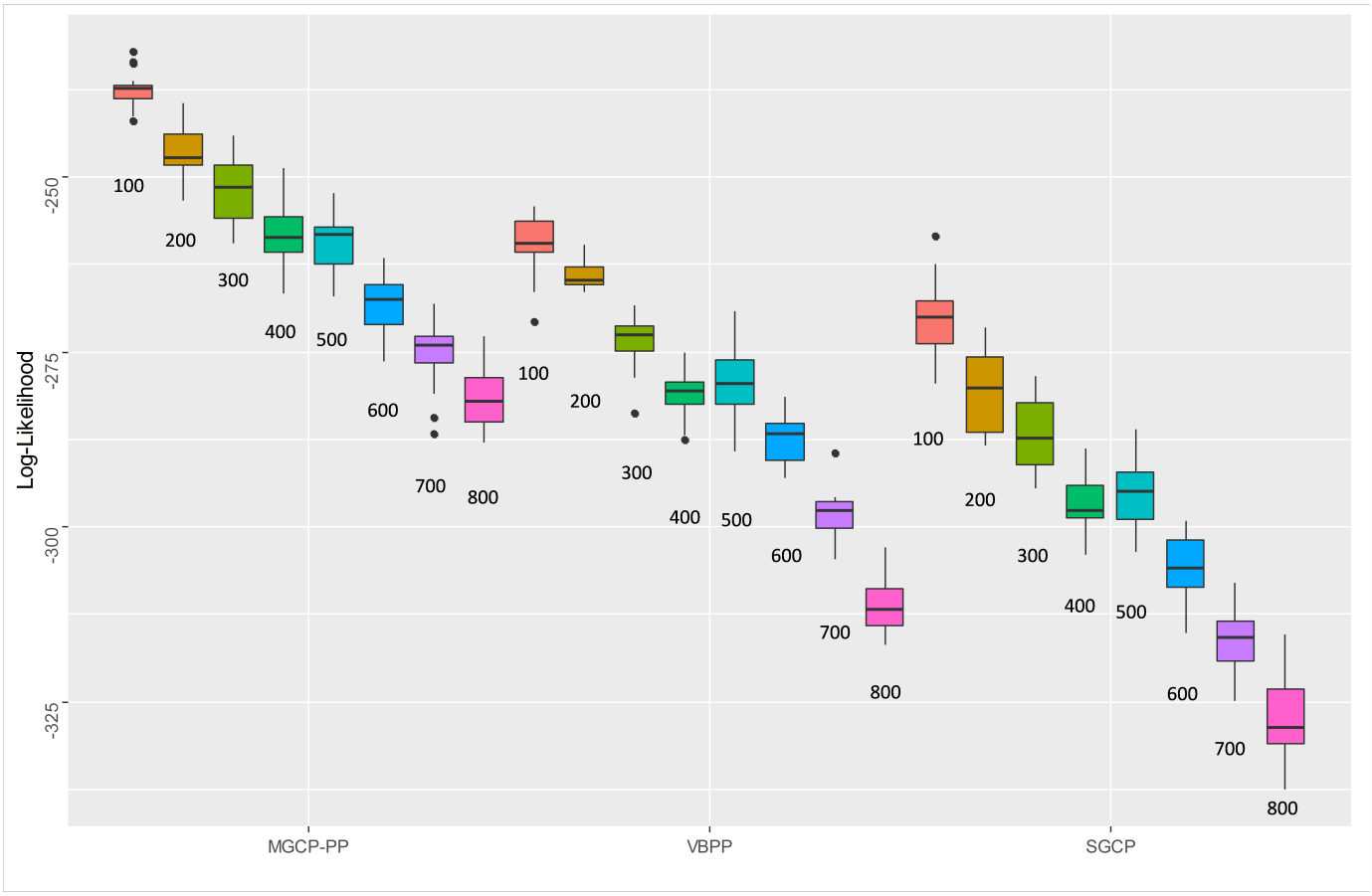

Fig. 10. Predictive log-likelihood based on cross-validation of case study dataset. (The number below each bar corresponds to prediction windows L)

the peer units. Our MGCP-Poisson model that facilitates sharing of information between the testing unit and the training units in the historical datatset clearly extrapolates the intensity rate more accurately which results in better prediction performance.

Moreover, we compare the predictive log-likelihood of different approaches based on the leave-one-out cross validation of case study dataset. The results of this analysis is shown in figure 10. We note that here, the true intensity rate is unknown for the case study dataset; thus, we cannot compare based on RMSE between the predicted intensity rate and true intensity rate. As it can be seen from figure 10, the MGCP-PP approach outperforms other three approaches in terms of predictive log-likelihood in different prediction windows L. More precisely, MGCP-PP improves prediction accuracy in terms of predictive log-likelihood by %7.73 compared to VBPP and by %12.61 compared to SGCP on average and in different prediction windows L.

6 Conclusion

In this study, a flexible and efficient non-parametric joint modeling framework for analyzing event data is presented. Specifically, we propose a multivariate Gaussian convolution process modulated Poisson process model that leverages information from all units via a shared latent function. A variational inference framework using inducing variables is further established to jointly estimate parameters from the MGCP-Poisson model accurately. The main advantage of the proposed framework is that it allows accurate individualized prediction for units in the field by sharing information obtained from the historical off-line units. The performance of proposed framework is benchmarked against two other competing approaches namely VBPP and SGCP which do not offer a sharing capability. The results of numerical studies and a case study based on real-world data for fleet based event prediction confirm that the proposed framework outperforms other competing approaches in terms of predictive log-likelihood and RMSE between the true intensity rate and the predicted intensity rate. More specifically, the case study results indicate that our proposed approach improves the predictive log-likelihood by %7.73 compared to VBPP and by %12.61 compared to SGCP, on average, for different prediction windows L. Moreover, an analysis based on the MAE of event occurrence count for the case study confirms a %16.34 and a %23.92 improvement compared to VBPP and SGCP, respectively, when using our proposed framework. Thus, the empirical studies highlight the advantageous features of our modeling framework to predict the intensity rates and provide reliable event prediction.

The model presented in this study can be readily extended to incorporate other observation covariates. Moreover, the convolution structure proposed in this study

is flexible. Here, we only shared one latent process across all the units. One can modify this structure by adding more independent latent processes for each unit to improve accuracy in modeling heterogeneity across units. Other structures that share a group of latent processes among selected group of units can be also extended from our model structure. Future work will be aimed towards developing these variational structured Poisson process frameworks.

ACKNOWLEDGMENTS

The financial support of this work is provided by The Raymond Corporation and National Science Foundation research grant #1824761.

REFERENCES

[1] X.-T. Yuan, X. Liu, and S. Yan, "Visual classification with multitask joint sparse representation," *IEEE Transactions on Image Processing*, vol. 21, no. 10, pp. 4349–4360, 2012.

[2] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[3] S. Conti, J. P. Gosling, J. E. Oakley, and A. O'Hagan, "Gaussian process emulation of dynamic computer codes," *Biometrika*, vol. 96, no. 3, pp. 663–676, 2009.

[4] H. Soleimani, J. Hensman, and S. Saria, "Scalable joint models for reliable uncertainty-aware event prediction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 8, pp. 1948–1963, 2017.

[5] D. Jarrett, J. Yoon, and M. van der Schaar, "Dynamic prediction in clinical survival analysis using temporal convolutional networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 424–436, 2019.

[6] W. Lian, R. Henao, V. Rao, J. Lucas, and L. Carin, "A multitask point process predictive model," in *International Conference on Machine Learning*, 2015, pp. 2030–2038.

[7] R. P. Adams, I. Murray, and D. J. MacKay, "Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 9–16.

[8] Y. W. Teh and V. Rao, "Gaussian process modulated renewal processes," in *Advances in Neural Information Processing Systems*, 2011, pp. 2474–2482.

[9] C. Lloyd, T. Gunter, M. Osborne, and S. Roberts, "Variational inference for gaussian process modulated poisson processes," in *International Conference on Machine Learning*, 2015, pp. 1814–1822.

[10] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. Chichilnisky, and E. P. Simoncelli, "Spatio-temporal correlations and visual signalling in a complete neuronal population," *Nature*, vol. 454, no. 7207, pp. 995–999, 2008.

[11] A. Gunawardana, C. Meek, and P. Xu, "A model for temporal dependencies in event streams," in *Advances in neural information processing systems*, 2011, pp. 1962–1970.

[12] A. K. Jardine, D. Lin, and D. Banjevic, "A review on machinery diagnostics and prognostics implementing condition-based maintenance," *Mechanical systems and signal processing*, vol. 20, no. 7, pp. 1483–1510, 2006.

[13] X.-S. Si, W. Wang, C.-H. Hu, and D.-H. Zhou, "Remaining useful life estimation–a review on the statistical data driven approaches," *European journal of operational research*, vol. 213, no. 1, pp. 1–14, 2011.

[14] W. Q. Meeker and L. A. Escobar, *Statistical methods for reliability data*. John Wiley & Sons, 2014.

[15] J. W. McPherson, J. McPherson, and Glaser, *Reliability physics and engineering*. Springer, 2010.

[16] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[17] J. P. Klein and M. L. Moeschberger, *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006.

[18] Z. Li, S. Zhou, S. Choubey, and C. Sievenpiper, "Failure event prediction using the cox proportional hazard model driven by frequent failure signatures," *IIE transactions*, vol. 39, no. 3, pp. 303–315, 2007.

[19] Y. Yuan, S. Zhou, C. Sievenpiper, K. Mannar, and Y. Zheng, "Event log modeling and analysis for system failure prediction," *IIE Transactions*, vol. 43, no. 9, pp. 647–660, 2011.

[20] P. Hougaard, *Analysis of multivariate survival data*. Springer Science & Business Media, 2012.

[21] L. Duchateau and P. Janssen, *The frailty model*. Springer Science & Business Media, 2007.

[22] A. Deep, D. Veeramani, and S. Zhou, "Event prediction for individual unit based on recurrent event data collected in teleservice systems," *IEEE Transactions on Reliability*, vol. 69, no. 1, pp. 216–227, 2019.

[23] B. H. Lindqvist *et al.*, "On the statistical modeling and analysis of repairable systems," *Statistical science*, vol. 21, no. 4, pp. 532–551, 2006.

[24] J. H. Cha and M. Finkelstein, "Some notes on unobserved parameters (frailties) in reliability modeling," *Reliability Engineering & System Safety*, vol. 123, pp. 99–103, 2014.

[25] M. Peng, Q. Xie, H. Wang, Y. Zhang, and G. Tian, "Bayesian sparse topical coding," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 6, pp. 1080–1093, 2018.

[26] J. Kingman, "Poisson processes," *Oxford Studies in Probability. Oxford: Oxford University Press*, 1993.

[27] J. Møller, A. R. Syversveen, and R. P. Waagepetersen, "Log gaussian cox processes," *Scandinavian journal of statistics*, vol. 25, no. 3, pp. 451–482, 1998.

[28] P. Diggle, "A kernel method for smoothing point process data," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 34, no. 2, pp. 138–147, 1985.

[29] B. D. Ripley, "Modelling spatial patterns," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 2, pp. 172–192, 1977.

[30] S. L. Rathbun and N. Cressie, "Asymptotic properties of estimators for the parameters of spatial inhomogeneous poisson point processes," *Advances in Applied Probability*, pp. 122–154, 1994.

[31] A. Kottas and B. Sansó, "Bayesian mixture modeling for spatial poisson process intensities, with applications to extreme value analysis," *Journal of Statistical Planning and Inference*, vol. 137, no. 10, pp. 3151–3163, 2007.

[32] M. A. Álvarez and N. D. Lawrence, "Computationally efficient convolved multiple output gaussian processes," *The Journal of Machine Learning Research*, vol. 12, pp. 1459–1500, 2011.

[33] X. Yue and R. A. Kontar, "Joint models for event prediction from time series and survival data," *Technometrics*, pp. 1–10, 2020.

[34] P. Vrignat, M. Avila, F. Duculty, and F. Kratz, "Failure event prediction using hidden markov model approaches," *IEEE Transactions on Reliability*, vol. 64, no. 3, pp. 1038–1048, 2015.

[35] K. Yu, V. Tresp, and A. Schwaighofer, "Learning gaussian processes from multiple tasks," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 1012–1019.

[36] P. J. Diggle, P. Moraga, B. Rowlingson, B. M. Taylor *et al.*, "Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm," *Statistical Science*, vol. 28, no. 4, pp. 542–563, 2013.

[37] S. Flaxman, A. Wilson, D. Neill, H. Nickisch, and A. Smola, "Fast kronecker inference in gaussian processes with non-gaussian likelihoods," in *International Conference on Machine Learning*, 2015, pp. 607–616.

[38] T. J. Leininger, A. E. Gelfand *et al.*, "Bayesian inference and model assessment for spatial point patterns using posterior predictive samples," *Bayesian Analysis*, vol. 12, no. 1, pp. 1–30, 2017.

[39] E. Snelson and Z. Ghahramani, "Sparse gaussian processes using pseudo-inputs," in *Advances in neural information processing systems*, 2006, pp. 1257–1264.

[40] M. Titsias, "Variational learning of inducing variables in sparse gaussian processes," in *Artificial Intelligence and Statistics*, 2009, pp. 567–574.

[41] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*, ser. Adaptive computation and machine learning. MIT Press, 2006.

[42] M. Alvarez and N. D. Lawrence, "Sparse convolved gaussian processes for multi-output regression," in *Advances in neural information processing systems*, 2009, pp. 57–64.

[43] J. Quinonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate gaussian process regression," *The Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, 2005.
[44] J. Quinonero-Candela, C. E. Rasmussen, and C. K. Williams, "Approximation methods for gaussian process regression," in *Large-scale kernel machines*. MIT Press, 2007, pp. 203–223.
[45] X. Yue and R. Kontar, "The rényi gaussian process," *arXiv preprint arXiv:1910.06990*, 2019.
[46] S. Chung, R. A. Kontar, and Z. Wu, "Weakly-supervised multi-output regression via correlated gaussian processes," *arXiv preprint arXiv:2002.08412*, 2020.
[47] E. Snelson and Z. Ghahramani, "Sparse gaussian processes using pseudo-inputs," *Advances in neural information processing systems*, vol. 18, pp. 1257–1264, 2005.
[48] M. Álvarez, D. Luengo, M. Titsias, and N. D. Lawrence, "Efficient multioutput gaussian processes through variational inducing kernels," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 25–32.

Salman Jahani received the B.S. degree in industrial engineering from University of Tehran, Tehran, Iran, in 2013, and the M.S. degree in industrial engineering from Sharif University of Technology, Tehran, Iran, in 2015. He received a Ph.D. in industrial engineering from University of Wisconsin-Madison, Madison, WI, USA, in 2021. His research interests revolve around data analytics, machine learning, optimization, and stochastic modeling with applications on system failure diagnosis, prognosis and maintenance decision making.

Shiyu Zhou received the B.S. and M.S. degrees in mechanical engineering from the University of Science and Technology of China, Hefei, China, in 1993 and 1996, respectively, and the master's degree in industrial engineering and the Ph.D. degree in mechanical engineering from the University of Michigan, Ann Arbor, MI, USA, both in 2000. He is the Vilas Distinguished Achievement Professor in the Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI, USA. His research interests include data-driven modeling, monitoring, diagnosis, and prognosis for engineering systems with particular emphasis on manufacturing and after-sales service systems. Dr. Zhou is a recipient of a CAREER Award from the National Science Foundation and the Best Application Paper Award from IIE Transactions. He is currently the Director of the IoT Systems Research Center at the University of Wisconsin-Madison and a Fellow of the IISE, the ASME, and the SME.

Dharmaraj Veeramani received the B.S. degree in mechanical engineering from the Indian Institute of Technology Madras, Chennai, India, in 1985, and the M.S. and Ph.D. degrees in industrial engineering from Purdue University, West Lafayette, IN, USA, in 1987 and 1991, respectively. He is the Robert Ratner Chair Professor with the Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI, USA. He has received numerous research grants from federal agencies and industry. His research focuses on emerging frontiers of digital business, Internet of Things technologies and applications, smart and connected systems, and supply chain management. Dr. Veeramani has received multiple honors and awards from organizations, such as the National Science Foundation, the SME, the SAE International, and the ASEE in recognition of his scholarly contributions.

Jeffrey Schmidt is a Data Scientist for the Raymond Corporation. He received his MS degree in systems science from Binghamton University in 2012. He is currently a Ph.D. candidate in systems science in the Collective Dynamics of Complex Systems research group at Binghamton University. His research focuses on adaptive networks and the dynamics of complex systems.