This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Gaussian Process Latent Variable Model-Based Multi-Output Modeling of Incomplete Data

Zhiyong Hu, Chao Wang, Jianguo Wu, *Member, IEEE*, and Dongping Du, *Member, IEEE*

***Abstract*— The rapid development of sensor technologies allows the acquisition of high dimensional sensing data. Multi-output modeling techniques have been developed to leverage the data for decision making. However, the data often contain segments of missing values, which cause great information loss and thus affect the modeling performance. This study explores the missing pattern and the correlation structure of missing segments and maximally exploits useful information in the data to improve multi-output modeling accuracy. Specifically, a new multi-output modeling method is developed based on Gaussian Process Latent Variable Model (GPLVM). A decision score is developed to seek an optimal modeling strategy and then a tailored Expectation-Maximization (EM) algorithm based on GPLVM is designed to estimate the missing segments while optimizing model parameters. The proposed method demonstrates superior performance in both a simulation study and a case study, which makes it a powerful tool to enable process automation.**

***Note to Practitioners*—In real-life applications, missing values are constantly present in multi-output sensing data, which greatly affects data-driven decision-making. Modeling of such data becomes more difficult when consecutive observations within or across different outputs are missing. Existing methods often discard the missing values and extract information only from the available observations. However, the pattern of missing values may contain important messages that can potentially boost the modeling performance. This research develops a new framework based on GPLVM to model multi-output data with segmented missing patterns. A tailored EM algorithm is developed to iteratively impute the missing values and optimize model parameters. In addition, a decision score that quantifies both the missing pattern and correlation is designed to determine an optimal modeling strategy. The proposed method can benefit many applications across different industries that require modeling of multi-output incomplete data, especially when the data have many segments of missing observations.**

***Index Terms*— Gaussian process latent variable model, multi-output modeling, missing data, expectation maximization.**

I. Introduction

THE rapid development of sensor technologies enables fast and convenient information acquisition, which generates an unprecedented amount of data. This provides great opportunities for effective monitoring, prognosis, and control of complex systems [1], [2], [3]. These data are usually multi-output in nature, i.e., multiple variables will be recorded at a single input. For example, in a machining process, the time-varying measurements of force and vibration provide two-output data, in which two observations are collected at each time. In addition, data are correlated within each output and/or across multiple outputs. The within-output correlation reflects the functional relationship between the input and each output. For example, the battery capacity (output) decreases over time (input) [4] and the stress (output) varies according to strain (input) [5] in mechanical property data. The cross-output correlation indicates the functional relationship among different outputs. For example, after operating for the same period of time, bearings under higher loads often degrade more severely [6]; another example is that, at a specific strain level, the stress of steel decreases as testing temperature increases [5]. To capture the within- and cross-output correlations in multi-output data, many modeling methods have been developed, such as functional principle component analysis [7], and multi-output Gaussian Process (MOGP) [8], [9]. These methods often require a complete data set for model training. However, incomplete data, i.e., data with missing values, are common in practice [7], and the missing values greatly challenge the existing techniques.

In real-life applications, missing data are often in different patterns, which can be generally categorized into sparse and segmented patterns [7]. In the sparse pattern, values in a subset of outputs are missing at scattered input locations. This pattern is often caused by the removal of outliers [10], e.g., measurements exceeding the sensor range. Many methods have been developed to deal with sparse missing data. For example, Rodrigues et al. [11] constructed a MOGP model based on a convolution process to impute the missing speed in road sections. Fang et al. [7] utilized functional principle component analysis to extract features from degradation signals and applied a kernel smoother to deal with missing values. Song et al. [12] assumed a parametric form for the

Manuscript received 2 May 2022; revised 3 December 2022; accepted 13 February 2023. This article was recommended for publication by Associate Editor R. Jin and Editor J. Li upon evaluation of the reviewers' comments. This work was supported in part by the National Science Foundation under Grant CMMI-1728338 and in part by the National Natural Science Foundation of China under Grant 71932006 and Grant 72171003. *(Corresponding author: Dongping Du.)*

Zhiyong Hu is with the Department of Precision Machinery and Precision Instrumentation, University of Science and Technology of China, Hefei, Anhui 230026, China (e-mail: hzyllwen@ustc.edu.cn).

Chao Wang is with the Department of Industrial and Systems Engineering, The University of Iowa, Iowa City, IA 52242 USA (e-mail: chao-wang-2@uiowa.edu).

Jianguo Wu is with the Department of Industrial Engineering and Management, College of Engineering, Peking University, Beijing 100080, China (e-mail: j.wu@pku.edu.cn).

Dongping Du is with the Department of Industrial, Manufacturing and Systems Engineering, Texas Tech University, Lubbock, TX 79409 USA (e-mail: dongping.du@ttu.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TASE.2023.3251386.

Digital Object Identifier 10.1109/TASE.2023.3251386

1545-5955 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Peking University. Downloaded on March 08,2023 at 08:00:25 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

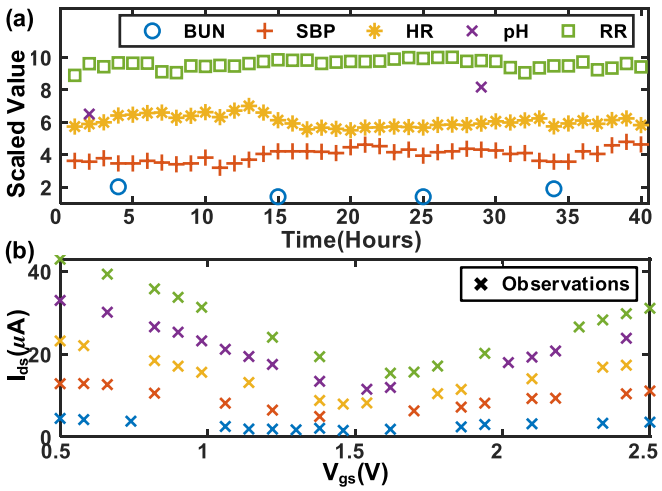


Fig. 1. Examples of segmented missing patterns. (a) Within-output segmented pattern. (b) Cross-output segmented pattern.

degradation signals and used observable values to estimate model parameters and missing values. In addition, matrix [13], [14] and tensor completion methods [15], [16], [17] have also been developed for missing value imputation. Although these methods showed good performance, they are designed for imputing data with sparse missing patterns; their performance for data with segmented missing patterns needs validation.

As for the segmented pattern, consecutive observations are missing, and there are two different types. The first type is the within-output segmented pattern, where one or more outputs lose their observations at consecutive inputs. A common reason for this type of missing pattern is sensor failure [10]. Another reason is that some outputs have higher sampling rates than others. An example of this pattern is shown in Fig. 1 (a), which has five different vital signs, i.e., blood urea nitrogen (BUN), respiratory rate (RR), heart rate (HR), pH, and systolic blood pressure (SBP), of a patient in the Intensive Care Unit (ICU). For demonstration purposes, these signals have been shifted and scaled. In the figure, consecutive recordings are missing in the BUN and pH signals due to their lower sampling rates, which causes segments of missing values within the outputs. The second type is the cross-output segmented pattern, where the observations across multiple outputs are lost at some inputs. This missing pattern can be caused by system breakdown [10] or limited sensing budget [18], an example of which is shown in Fig. 1 (b). The data show the change of drain-to-source currents I_{ds} according to the gate-to-source voltage V_{gs} under five different conditions [10]. In the figure, the observations of some neighboring outputs are missing at several voltages, causing the missing segments across outputs. Segmented missing patterns usually lead to greater information loss than the sparse patterns and are often more difficult to deal with. Therefore, we will focus on modeling the incomplete data with segmented missing patterns in this study.

Among existing multi-output modeling methods, Gaussian Process Latent Variable Model (GPLVM) has been reported with the capability to quantify the input-output relationship when the data contain segmented missing values [19]. In GPLVM, a latent variable is introduced at each input to link the input with different outputs, which can capture both within- and cross-output correlation. When modeling data with missing values, the GPLVM approach will delete the missing values in each output and use the available observations to optimize the model. This modeling process can work well when the available observations contain enough information for model optimization [19]. However, it cannot deal with data containing many missing segments because the missing segments cause significant information loss, and available observations alone are too sparse to provide good accuracy. To address this issue, a straightforward idea is to complete the data with imputed values before model training. However, inaccurate imputation might affect the data quality and further influence the performance of the GPLVM model.

Moreover, modeling of missing data with segmented patterns should exploit all useful information both within an output and across multiple outputs. When the information loss within outputs is notable and the within-output correlation is low, the modeling efforts should focus on learning the interactions across different outputs. On the other hand, when the data has more cross-output missing values and the cross-output correlation is weak, the modeling efforts should concentrate on exploring within-output correlations. However, GPLVM deals with different missing patterns following the same procedure. To the best of our knowledge, few studies have considered the missing patterns, e.g., the size and distribution of the missing segments, in multi-output modeling. Therefore, new techniques should be developed to model incomplete data with different segmented missing patterns.

In this paper, we propose a new modeling framework to fill the above-mentioned gaps. The framework is developed based on the GPLVM but has two novel designs to address the limitations of GPLVM. Given the data with segments of missing values, we propose to iteratively impute the missing values and optimize model parameters through a tailored Expectation-Maximization (EM) algorithm. Under this framework, GPLVM is built using both the observed data and the imputed missing values, and the imputation accuracy is improved across different iterations to obtain a better modeling result. In addition, two modeling options, i.e., within- and cross-output modeling, are considered. Both options can simultaneously capture the within- and cross-output correlations in the multi-output data, but each option has its focus on either the within- or cross-output correlation, depending on the proportion and distribution of the missing segments. To choose between the two options, a decision score is developed based on missing patterns and correlations. The intuition to incorporate missing pattern is that different missing patterns can lead to different information loss. For example, the missing pattern in Fig. 1 (a) causes significant information loss within each output and across different outputs, while the missing pattern in Fig. 1 (b) causes more information loss across different outputs. Further, correlation also has a significant impact on the modeling performance. For example, if outputs are independent, i.e., cross-output correlation is zero, then cross-output modeling cannot extract useful information

Authorized licensed use limited to: Peking University. Downloaded on March 08,2023 at 08:00:25 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

from the data. Therefore, we propose to incorporate both the missing pattern and the correlation information in the decision score. With the combination of the decision score and the EM algorithm, the proposed modeling framework can maximally exploit useful information and find an accurate representation of the multi-output data.

It is worth mentioning that other methods have been developed to deal with data containing segmented missing values. For example, Parra and Tobar [20] developed a spectral mixture kernel with a phase shift for MOGP to model the cross-covariance of different outputs. Zhao and Sun [21] developed a variationally dependent multi-output dynamic model for time series modeling. Although these methods achieved promising results, their performance on the data with many missing segments needs further validation. In addition, the focus of this study is different from the previous studies. We intend to build a new modeling framework that can learn the missing patterns and correlation structure of data and further leverage the learned information to obtain a better modeling performance. Thus, to summarize, the contributions of this paper are as follows:

1) A decision score is developed to characterize the missing pattern and correlations in the data, which informs the decision on within- or cross-output modeling.
2) A tailored EM algorithm is developed to iteratively update the imputation for missing values and optimize model parameters, which enables accurate modeling of incomplete data with segmented missing patterns.
3) The performance of the proposed modeling framework is verified with a simulation study and a case study using the transfer characteristics of a graphene field-effect transistor (GFET). The superiority of the proposed method makes it a powerful tool for extracting valuable information from the data for process automation.

The rest of the paper is organized as follows. In Section II, assumptions and the problem formulation are provided. Section III provides the details of the proposed method. Section IV presents the performance comparisons of the proposed framework to some existing methods. Conclusions and future works are given in Section V.

II. Gaussian Process Latent Variable Model Based Multi-Output Modeling

A. Assumptions

To facilitate the multi-output modeling based on GPLVM, two assumptions are specified below:

A1 The input space is the same for different outputs, i.e., the input space of different outputs is of same dimension.
A2 The underlying function corresponding to each output is smooth. This assumption is made due to the adoption of GP prior with a squared exponential covariance function for the mapping between latent variables and outputs.

These two assumptions are general assumptions for Gaussian Process (GP) based multi-output modeling methods. Therefore, our proposed method is applicable to all scenarios where standard GP applies. The input dimension of the data is assumed to be one in the following sections. A single dimensional input is general enough since the input of many multi-output modeling problems in real-world applications is in one dimension, e.g., the medical data described in Fig. 1 (a). In addition, to testify the performance in high dimensional input space, a simulation study has been done, and the results are provided in Appendix B.

B. Problem Formulation

In this section, we will first introduce the standard GPLVM and explain its limitations for modeling multi-output data with many missing segments. GPLVM is chosen because it can effectively capture complex correlation between multi-variate data using latent variables. Suppose there are M outputs and each with N observations at the input vector $\mathbf{s} = [s_1, \cdots, s_N]^T$. For the m^{th} output, $m = 1, \cdots, M$, we assume the observation vector is $\mathbf{y}_m = [y_m(s_1), \cdots, y_m(s_N)]^T$. By grouping different outputs together, we have a data matrix $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_M]^T$. In addition, we denote $\mathbf{x}(s_n) \in \mathbb{R}^L$ as the latent variable corresponding to the input s_n and $L < M$. With these notations, the observation at the n^{th} input of m^{th} output is assumed to be:

$$y_m(s_n) = f_m(\mathbf{x}(s_n)) + \varepsilon(s_n) \tag{1}$$

where $\varepsilon(\cdot)$ is the measurement noise with independent and identically distributed (i.i.d.) normal distribution $\mathcal{N}(0, \beta^{-1})$. Please note that the i.i.d. noise assumption can be relaxed to incorporate correlated noise structure in the model. To simplify the notation, we will use $\mathbf{x}_n$ to represent $\mathbf{x}(s_n)$ henceforth. In addition, $f_m(\cdot)$s are functions to be estimated and assumed to follow the same GP, that is, for $m = 1, \cdots, M$,

$$f_m(\mathbf{x}_n) \sim \mathcal{GP}\big(\mu(\mathbf{x}_n), k(\mathbf{x}_n, \mathbf{x}_{n'})\big) \tag{2}$$

where $n, n' = 1, \cdots, N$, $\mu(\cdot)$ is the mean function and $k(\cdot,\cdot)$ is the covariance function. It worth noting that, when $\mathbf{x}_n$ is unknown, different f_ms are correlated. Since any finite collection of random variables from a GP has a multivariate normal distribution ($\mathcal{MVN}$) [22], we have $\mathbf{f}_m|\mathbf{X}, \mathbf{s} \sim \mathcal{MVN}(\mathbf{f}_m|\boldsymbol{\mu}, \mathbf{K})$, where $\mathbf{f}_m = [f_m(\mathbf{x}_1), \cdots, f_m(\mathbf{x}_N)]^T$ is the function value vector, $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]^T$ collects all latent variables, $\boldsymbol{\mu} = [\mu(\mathbf{x}_1), \cdots, \mu(\mathbf{x}_N)]^T$ is the mean vector, $\mathbf{K}$ is the covariance matrix between the function values and the item in the i^{th} row and j^{th} column equals $k(\mathbf{x}_i, \mathbf{x}_j)$. Since the mean function is usually assumed to be zero, we have $\boldsymbol{\mu} = \mathbf{0}$. Due to the normal assumption for the observation noise, we further have $\mathbf{y}_m|\mathbf{X} \sim \mathcal{MVN}(\mathbf{y}_m|\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \mathbf{K} + \beta^{-1}\boldsymbol{I}_N$ as the covariance matrix of $\mathbf{y}_m$ and $\boldsymbol{I}_N$ an $N \times N$ identity matrix. Then, for all M outputs, we have $p(\mathbf{Y}|\mathbf{X}) = \prod_{m=1}^{M} p(\mathbf{y}_m|\mathbf{X}, \mathbf{s}) = \prod_{m=1}^{M} \mathcal{MVN}(\mathbf{y}_m|\mathbf{0}, \boldsymbol{\Sigma})$. With this likelihood function, a variational inference method [19] has been proposed by assuming a variational distribution for the latent variable, i.e., $q(\mathbf{X}; \boldsymbol{\theta}_{\mathbf{X}})$, which derives a lower bound for the log marginal likelihood $\log p(\mathbf{Y})$, that is,

$$\log p(\mathbf{Y}) \geq \mathcal{J}(\boldsymbol{\theta}_{\mathbf{X}}, \boldsymbol{\theta}_{\mathcal{M}}) = -\frac{1}{2}\sum_{m=1}^{M} \mathbf{y}_m^T \mathbf{V} \mathbf{y}_m + \mathcal{H} \tag{3}$$

where $\boldsymbol{\theta}_{\mathcal{M}}$ represents other model parameters, $\mathbf{V}$ and $\mathcal{H}$ are both functions of $\boldsymbol{\theta}_{\mathbf{X}}$ and $\boldsymbol{\theta}_{\mathcal{M}}$. Under the case with complete

Authorized licensed use limited to: Peking University. Downloaded on March 08,2023 at 08:00:25 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

observations, this bound is maximized to obtain an estimation of $\boldsymbol{\theta}_{\mathbf{X}}$ and $\boldsymbol{\theta}_{\mathcal{M}}$, which are then used to estimate $f_m(\cdot)$. It should be noted that within-output modeling is utilized in this section to demonstrate the modeling process of GPLVM. As for cross-output modeling, the procedure is the same but instead take each column of $\mathbf{Y}$ as an output. Specifically, the transpose of the original data matrix $\tilde{\mathbf{Y}} = \mathbf{Y}^T$ is taken as the new data matrix, which is of size $N \times M$. Under this scenario, each row of $\tilde{\mathbf{Y}}$, i.e., $\tilde{\mathbf{y}}_n = [y_1(s_n), \cdots, y_M(s_n)]^T, n = 1, \cdots, N$ is taken as an output and the latent variables are used to capture the correlation between different $\tilde{\mathbf{y}}_n$s.

As seen in Eq. 3, the lower bound is decomposed across different outputs $\mathbf{y}_m$, and $\mathcal{H}$ is independent of the observations. Thus, missing values will only affect the calculation of the term $\sum_{m=1}^{M} \mathbf{y}_m^T \mathbf{V} \mathbf{y}_m$. Standard GPLVM discards the missing values in $\mathbf{y}_m$ and the corresponding rows and columns in $\mathbf{V}$. It can be observed that, when many $\mathbf{y}_m$s have segments of missing values, i.e., within-output segmented patterns, the remaining components in $\sum_{m=1}^{M} \mathbf{y}_m^T \mathbf{V} \mathbf{y}_m$ cannot provide enough information for model optimization. However, for the data with significant within-output information loss, we might be able to extract more useful information across different outputs. Thus, for an arbitrary data set with missing values, a decision should be made on whether to perform within-output modeling or cross-output modeling. Standard GPLVM does not have such decision process, which limits its applicability in handling data with significant information loss.

III. Proposed Framework for Modeling of Multi-Output Data With Segmented Missing Patterns

This section provides details of the proposed modeling framework for multi-output data with different missing patterns. Section III-A presents the tailored EM algorithm, which optimizes the model in a sequential manner with a continually improved imputation. Section III-B presents the design of the decision score, which quantifies both the size and distribution of missing segments and the within- and cross-output correlation. The general scheme of the proposed method is shown in Fig. 2. Specifically, given a set of multi-output data with segments of missing values, a decision score is first calculated based on the missing pattern and correlation; then the decision score is used to determine whether within- or cross-output modeling should be conducted; lastly, a tailored EM algorithm is implemented to optimize the model.

A. Customized Expectation-Maximization Algorithm

In this subsection, within-output modeling is used to demonstrate the proposed EM algorithm. Notations for observed and missing values are provided as follows. We denote $\mathbf{y}_m^{\mathcal{O}} = [y_m(s_{m1}^{\mathcal{O}}), \cdots, y_m(s_{mN_m^{\mathcal{O}}}^{\mathcal{O}})]^T$ and $\mathbf{s}_m^{\mathcal{O}} = [s_{m1}^{\mathcal{O}}, \cdots, s_{mN_m^{\mathcal{O}}}^{\mathcal{O}}]^T$ as the observed values and the corresponding input vector for the m^{th} output, respectively. Similarly, we use $\mathbf{y}_m^{\mathcal{U}} = [y_m(s_{m1}^{\mathcal{U}}), \cdots, y_m(s_{mN_m^{\mathcal{U}}}^{\mathcal{U}})]^T$ and $\mathbf{s}_m^{\mathcal{U}} = [s_{m1}^{\mathcal{U}}, \cdots, s_{mN_m^{\mathcal{U}}}^{\mathcal{U}}]^T$ to denote the components corresponding to the missing section of the m^{th} output. It should be noted that, for $m = 1, \cdots, M$, we have $\mathbf{y}_m = \mathbf{y}_m^{\mathcal{O}} \bigcup \mathbf{y}_m^{\mathcal{U}}$, $\mathbf{s} = \mathbf{s}_m^{\mathcal{O}} \bigcup \mathbf{s}_m^{\mathcal{U}}$ and $N = N_m^{\mathcal{O}} + N_m^{\mathcal{U}}$. By grouping M outputs together, $\mathbf{Y}^{\mathcal{O}} = \{\mathbf{y}_1^{\mathcal{O}}, \cdots, \mathbf{y}_M^{\mathcal{O}}\}$ and $\mathbf{Y}^{\mathcal{U}} = \{\mathbf{y}_1^{\mathcal{U}}, \cdots, \mathbf{y}_M^{\mathcal{U}}\}$ are used to correspondingly denote the observed and missing values in the data. It should be noted that the missing mechanism is missing at random in this study, which indicates that the missingness depends on the observed values [23].

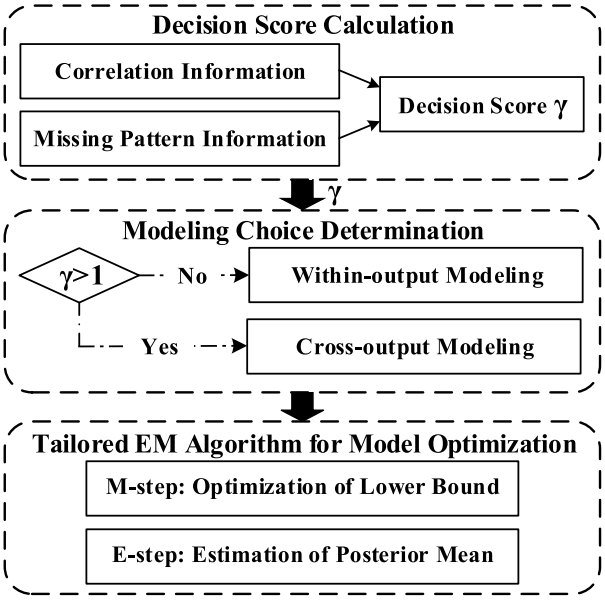


Fig. 2. Scheme of proposed method.

As illustrated in Fig. 3, the algorithm starts with an initial imputation. With the imputed data, the model parameters are then optimized in the M-step which maximizes a lower bound of the marginal log-likelihood of the available observations. Further, the estimation of the missing observations will be updated with the optimized model from its posterior distribution in the E-step. Under this framework, the model is optimized from a continually improved imputation, which alleviates the drawback of GPLVM and avoids the negative effect from inaccurate imputations.

We will first introduce the M-step, which optimizes model parameters given the estimation of missing segments $\mathbf{Y}^{\mathcal{U}}$. Let $\boldsymbol{\lambda}_m = \mathbb{E}(\mathbf{f}_m|\mathbf{Y}^{\mathcal{O}}) = [\lambda_m(\mathbf{x}_1), \cdots, \lambda_m(\mathbf{x}_N)]^T$ be the posterior mean for the m^{th} output ($m = 1, \cdots, M$) and $\widehat{\boldsymbol{\lambda}}_m$ be its estimation. Let $\widehat{\boldsymbol{\lambda}}_m^{\mathcal{O}}$ and $\widehat{\boldsymbol{\lambda}}_m^{\mathcal{U}}$ be the observed part and missing part, respectively. Then, the following lemma is proposed to optimize model parameters.

Lemma 1: If the modeling assumption of Eqs. 1 and 2 are satisfied and $p(\mathbf{Y}^{\mathcal{U}}|\mathbf{Y}^{\mathcal{O}}, \mathbf{s}) = \prod_{m=1}^{M} p(\mathbf{y}_m^{\mathcal{U}}|\mathbf{Y}^{\mathcal{O}}, \mathbf{s}) = \prod_{m=1}^{M} \delta(\mathbf{y}_m^{\mathcal{U}} - \widehat{\boldsymbol{\lambda}}_m^{\mathcal{U}})$ is the posterior density of the unobserved segments of the data, then maximizing the log marginal likelihood of data, i.e., $\log p(\mathbf{Y}^{\mathcal{O}}; \boldsymbol{\theta})$, is equivalent to the maximization of the following lower bound:

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{\beta}{2} \sum_{m=1}^{M} \widetilde{\mathbf{y}}_m^T \big(\mathbf{I}_N - \mathbf{W}(\boldsymbol{\theta})\big) \widetilde{\mathbf{y}}_m + \mathcal{F}(\boldsymbol{\theta}) \tag{4}$$

where $\delta(\cdot)$ is the Dirac delta function, $\boldsymbol{\theta}$ is the vector containing all unknown parameters including β. $\widetilde{\mathbf{y}}_m$ is the pseudo

Authorized licensed use limited to: Peking University. Downloaded on March 08,2023 at 08:00:25 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

observation obtained by replacing the unobserved segments in $\mathbf{y}_m$ with $\widehat{\boldsymbol{\lambda}}_m^{\mathcal{U}}$. We can similarly denote the segments corresponding to observed and missing values as $\widetilde{\mathbf{y}}_m^{\mathcal{O}}$ and $\widetilde{\mathbf{y}}_m^{\mathcal{U}}$, then $\widetilde{\mathbf{y}}_m^{\mathcal{O}} = \mathbf{y}_m^{\mathcal{O}}$ and $\widetilde{\mathbf{y}}_m^{\mathcal{O}} = \widehat{\boldsymbol{\lambda}}_m^{\mathcal{O}}$. The detailed expression for $\mathbf{W}(\boldsymbol{\theta})$ and $\mathcal{F}(\boldsymbol{\theta})$ as well as the proof of Lemma 1 are provided in Appendix A.

Remark 1: The assumption $p\left(\mathbf{Y}^{\mathcal{U}}|\mathbf{Y}^{\mathcal{O}},\mathbf{s}\right) = \prod_{m=1}^{M} p\left(\mathbf{y}_m^{\mathcal{U}}|\mathbf{Y}^{\mathcal{O}},\mathbf{s}\right)$ in Lemma 1 means that missing values can be recovered using observed values and the input sequence. Please note that this assumption can be violated at the input location where observations from all outputs are missing, i.e., at the location $\tilde{\mathbf{s}} = \bigcap_{m=1}^{M} \mathbf{s}_m^{\mathcal{U}}$.

Remark 2: Due to the nonlinear mapping $f_m(\cdot)$, there is no closed form expression for $p\left(\mathbf{y}_m^{\mathcal{U}}|\mathbf{Y}^{\mathcal{O}},\mathbf{s}\right)$ and some approximations are needed to derive the lower bound. In this paper, we choose to use a simple approximation which assumes a point mass on the estimated posterior mean. The justification is that it offers a straightforward bound for optimization and, at the same time, achieves an accurate estimation. However, if the research aim is different, other approximations can be easily incorporated. For example, some research may require the uncertainty estimation of $\mathbf{Y}^{\mathcal{U}}$ for the construction of control charts [24], then a normal distribution with the estimated posterior mean and variance can be adopted.

We can now move to the E-step, in which the estimation of the missing segments is updated. After optimizing the lower bound in Eq. 4, an estimation of the model parameters $\widehat{\boldsymbol{\theta}}$ can be obtained. With the optimized latent variables, the GP prediction with uncertain inputs [25] can be used to estimate the posterior mean, i.e., $\widehat{\boldsymbol{\lambda}}_m = \mathbf{W}(\boldsymbol{\theta})\widetilde{\mathbf{y}}_m$. By doing so, the estimation of $\boldsymbol{\lambda}_m$ and $\boldsymbol{\theta}$ are coupled with each other, which completes the design of the customized EM algorithm.

To better formulate the EM algorithm, a superscript denoting the sequentially updated estimations and parameters is introduced to the current notation system. For example, $\widehat{\boldsymbol{\lambda}}_m^{(i)}$ and $\widehat{\boldsymbol{\theta}}^{(i)}$ respectively denote the estimation of posterior mean and model parameters at iteration i. We can similarly denote $\widetilde{\mathbf{y}}_m^{(i)}$ as the pseudo observation vector of the m^{th} output at iteration i. With these notations, implementation of the proposed algorithm is summarized in Fig. 3, which starts by assuming the mean of the missing values as zero, i.e., $\widetilde{\mathbf{y}}_m^{\mathcal{U}(0)} = \mathbf{0}$ and $\widetilde{\mathbf{y}}_m^{\mathcal{O}(0)} = \mathbf{y}_m^{\mathcal{O}}$. At each iteration, we optimize the lower bound in Eq. 4 to obtain $\widehat{\boldsymbol{\theta}}^{(i)}$ during the M-step and then calculate the posterior mean using $\widehat{\boldsymbol{\lambda}}_m^{(i)} = \mathbf{W}(\boldsymbol{\theta}^{(i)})\widetilde{\mathbf{y}}_m^{(i-1)}$ in the E-step for the next iteration. The algorithm ends when either the maximum number of iterations is reached or the difference of the estimated posterior mean between consecutive iterations decreases to a predefined threshold. After convergence, the posterior mean will be calculated and compared to the true values to evaluate the modeling performance.

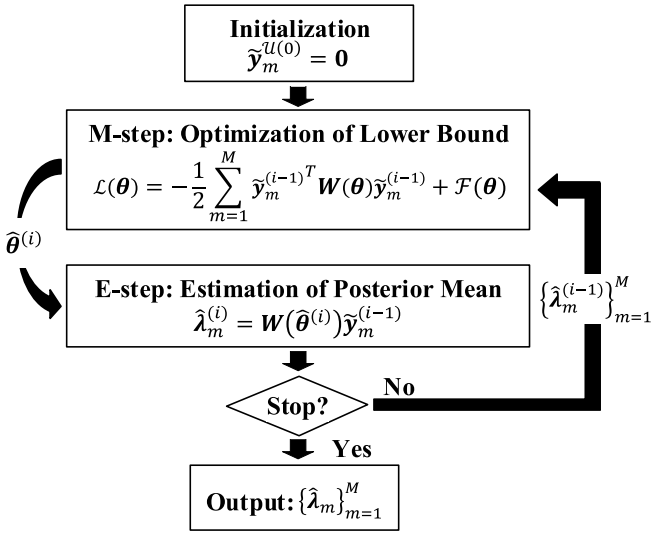


Fig. 3. Expectation-maximization algorithm.

B. Decision Score for Information Evaluation

In real-world applications, it is difficult to determine the missing pattern and correlation structure of the data from visual inspection. It is important to design a quantitative index to describe such information and further incorporate them in the modeling process. In this subsection, a decision score is developed to suggest performing within- or cross-output modeling. In addition, since the data are grouped into a 2D data matrix with each row corresponding to an output, within-output and cross-output modeling are equivalent to row-wise and column-wise modeling, respectively. That is, within-output (row-wise) modeling treats each row as an output, while cross-output (column-wise) modeling considers each column as an output. It should be noted that both modeling options can simultaneously capture the within- and cross-output correlations in the multi-output data, but each option has its emphasis on either the within- or cross-output correlation to achieve the best modeling performance.

Two indices, τ_r and τ_c, are defined to quantify the missing pattern when within-output (row-wise) and cross-output (column-wise) modeling are respectively conducted. To calculate τ_r and τ_c, two matrices $\mathbf{B}_r$ and $\mathbf{B}_c$ in the same size as the data matrix are first generated to label the missing segments. The entries of $\mathbf{B}_r$ are determined in a row-wise manner by taking each row of the data matrix as a unity. Specifically, if there is a missing segment in a row, the corresponding terms in $\mathbf{B}_r$ will be labeled by the length of the segment divided by the number of columns/inputs (i.e., N); otherwise, the terms are labeled as zeros. An example of calculating $\mathbf{B}_r$ is shown in Fig. 4. In the figure, the missing pattern of a 4×5 data matrix is shown, where white and black blocks respectively indicate the position of missing and observed values. The first three rows have missing values, where the 1^{st} row contains a missing segment of length one, the 2^{nd} row contains two missing segments of length two and one, and the 3^{rd} row contains two missing segments of length one. Thus, the corresponding terms in the three rows in $\mathbf{B}_r$ are set to the length of the missing segments divided by the column number 5. Alternatively, the entries of $\mathbf{B}_c$ are determined in a column-wise manner by taking each column as a unity. With the missing pattern in Fig. 4, an example of $\mathbf{B}_c$ is also computed. There are three columns contain missing values, i.e, the 2^{nd}, 4^{th} and 5^{th} column, and each contains a missing segment of length two.

Authorized licensed use limited to: Peking University. Downloaded on March 08,2023 at 08:00:25 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

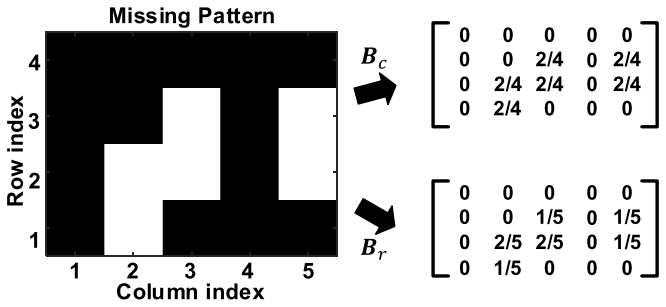


Fig. 4. Demonstration for the calculation of $\mathbf{B}_c$ and $\mathbf{B}_r$.

Thus, the terms in $\mathbf{B}_c$ corresponding to the missing values are all $2/4$. The defined matrices provide information on the size and the distribution of missing segments as well as the proportion of information loss in each row or column, which will serve as a good indicator of missing patterns.

Given $\mathbf{B}_r$ and $\mathbf{B}_c$, two new matrices are calculated as $\mathbf{R}_r = (\mathbf{1}_{NM} - \mathbf{B}_r)(\mathbf{1}_{NM} - \mathbf{B}_r)^T/N$ and $\mathbf{R}_c = (\mathbf{1}_{NM} - \mathbf{B}_c)^T(\mathbf{1}_{NM} - \mathbf{B}_c)/M$, where $\mathbf{1}_{NM}$ is an $N \times M$ matrix with all ones. $\mathbf{R}_r$ (or $\mathbf{R}_c$) describes the information preserved in each row (or column) and the correlation of the missing patterns across different rows (or columns). Lastly, $\tau_r = \|\mathbf{R}_r\|_F$ and $\tau_c = \|\mathbf{R}_c\|_F$, where $\|\cdot\|_F$ represents the Frobenius norm. Based on the calculation procedure, we can find that a larger value of τ_r encourages within-output modeling while a larger value of τ_c prefers cross-output modeling.

In addition, another two quantities ρ_c and ρ_r are introduced to determine the correlations among columns and rows, correspondingly. As for within-output modeling, each row is considered as an integrity, and the Pearson's correlation matrix $\mathbf{G}_r$ with dimension $M \times M$ is calculated. Then, ρ_r is calculated as $\rho_r = 1/(1 + \exp(-10(\|\mathbf{G}_r\|_F - 0.5)))$. ρ_c can be similarly calculated by considering each column as an integrity.

Based on the calculation, we can see that a larger value of τ_c (or τ_r) indicates less influence of missing segments on the cross-output (or within-output) modeling. On the other hand, a smaller value of ρ_c (or ρ_r) denotes less information could be extracted from the data using the within-output (or cross-output) modeling. Consequently, we propose to use τ_c/ρ_c and τ_r/ρ_r to respectively indicate the preference for the cross-output and within-output modeling. In summary, the following decision score is developed to determine the modeling choice:

$$\gamma = \frac{\tau_c/\rho_c}{\tau_r/\rho_r} \tag{5}$$

Thus, if $\gamma > 1$, cross-output modeling should be conducted, and, if $\gamma \leq 1$, within-output modeling is preferred.

IV. Model Performance

In this section, a simulation study and a case study with one dimensional input space are conducted to investigate the performance of the proposed method. Please note we have also conducted a simulation study with two dimensional input space. Due to space limit, we present the results in Appendix B. As for the simulation study, outputs with strong or weak correlations as well as different missing patterns are generated to evaluate our method. Then the transfer characteristics of GFET will be adopted in the case study to further test the modeling performance. Two multi-output modeling methods, i.e., the standard GPLVM and the MOGP based on the linear model of coregionalization (LMC) [8], and a matrix completion (MC) method [26] are used as benchmarks. As for the LMC model, the number of latent functions is set as 2 in all the experiments. In addition, within- and cross-output modeling using the proposed EM algorithm are represented as 'EM-Within' and 'EM-Cross', respectively. Squared exponential covariance function is adopted for the GPs. To evaluate the modeling performance, the root mean squared error (RMSE) between the true and estimated function values is computed and all experiments are repeated for at least 20 times for each γ to report the results. All the experiments are implemented on an Intel core Xeon CPU (@2.3GHz) and 64 GB RAM Windows PC with MATLAB 2020a. We also test the performance under alternative settings, and the results are consistent with those in this paper.

A. Simulation Study

To validate the performance of the proposed method and the effectiveness of the decision score, multi-output data with different missing patterns and varying correlations are simulated in this section. The complete data includes 12 signals and each has 30 observations generated using the following equation:

$$y_m(z) = \exp(\zeta_m)\Big(\frac{m}{2} + sin\big((z + 2\phi_m)\pi\big)\Big) + \epsilon_m(z) \tag{6}$$

where $m = 1, \cdots, 12$ is the index of the outputs, $z \in (-\pi/2, \pi/2)$, $\epsilon_m(z)$ is i.i.d. observation noise with $\mathcal{N}(0, 0.1^2)$, ζ_m and ϕ_m controls the modulation level and the phase shift, respectively. We will vary the value of ϕ_m to control the correlation between different signals. In addition, the missing data are generated as follows: First, 6 out of the 12 signals are randomly selected; then a segment containing 30% to 60% of all observations in a signal is removed.

We applied the proposed method and the standard GPLVM to model the simulated data, and the results are shown in Fig.5. Fig. 5 (a) presents the estimations of six outputs by GPLVM, EM-Cross and EM-within in one experiment with γ equals 1.1. The observations used to develop the multi-output models and the true values are respectively marked by blue cross symbols and solid black lines. From the figure, we can find that the EM-Cross model provides the best estimation for all signals in the entire input domain. The EM-within model does not perform as good as the EM-Cross model for signal y_8 and signal y_9. The reason is that the two signals have more within-output missing segments; cross-output modeling can learn the correlations across different outputs to compensate the information loss, while within-output model can not. This observation is consistent with the decision score since $\gamma > 1$ indicates the EM-Cross model is preferred over the EM-Within model. However, the EM-Within model can still provide better results than the GPLVM. The estimated missing values of the GPLVM differ significantly from the true values,

Authorized licensed use limited to: Peking University. Downloaded on March 08,2023 at 08:00:25 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

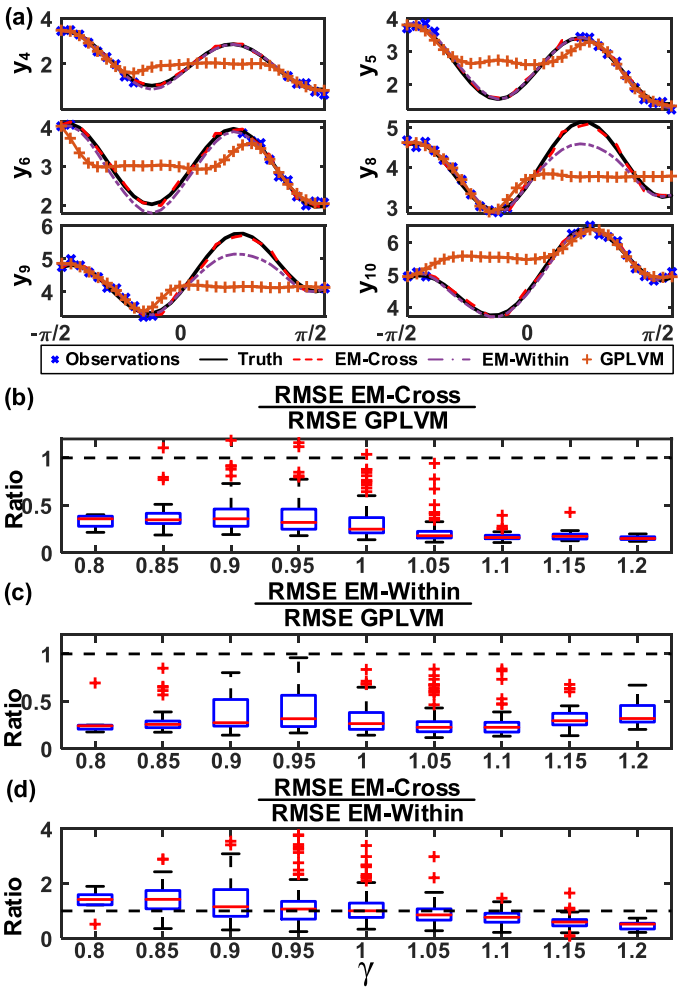


Fig. 5. Results for the simulation study. (a) Function prediction for one replication with $\gamma = 1.1$. RMSE ratio between (b) EM-Cross and GPLVM, (c) EM-Within and GPLVM, (d) EM-Cross and EM-Within.

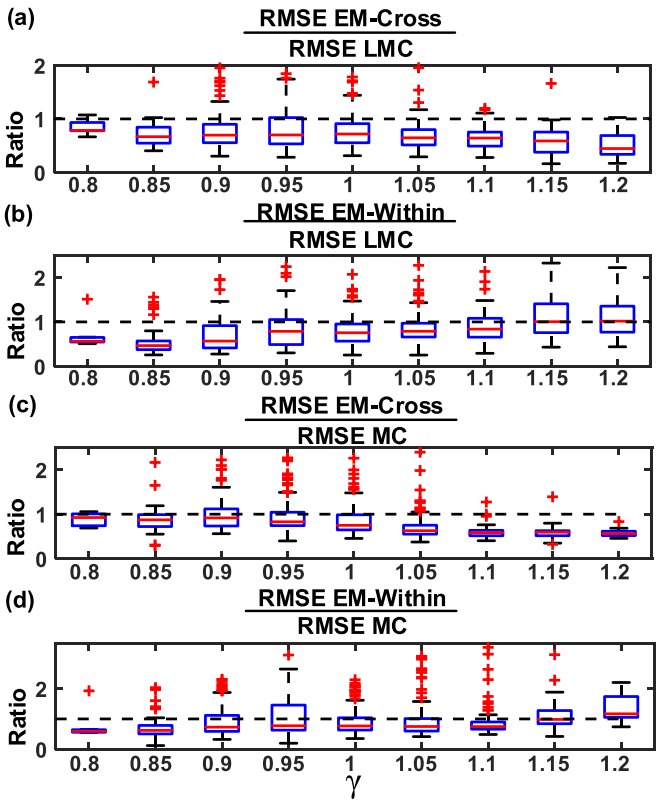


Fig. 6. Results for the simulation study. RMSE ratio between (a) EM-Cross and LMC, (b) EM-Within and LMC, (c) EM-Cross and MC, (d) EM-Within and MC.

which verifies that it cannot perform well when there are many missing segments in the data.

Moreover, additional experiments were done to compare the performance of the proposed method to GPLVM using data with different missing patterns. We computed the RMSE ratios of EM-Cross to GPLVM, EM-Within to GPLVM, and EM-Cross to EM-Within for different values of γ and the box plots of the ratios are shown in Fig. 5 (b), (c) and (d) respectively. From Fig. 5 (b) and (c), we can find that both the EM-Cross model and the EM-Within model have smaller RMSEs than the GPLVM model for different values of γ (i.e., the medians of RMSE ratios < 1). It is worth noting that, when the value of decision score is around 1, the boxes in 5 (b) and (c) are wider and/or with long whiskers. It is because, when the decision score approaches 1, the missing patterns in rows and in columns become similar, and the data contains more uncertainty. Subsequently, the model accuracy shows large variations across different replications. Despite the higher variation, the proposed method performs consistently better than GPLVM at all scenarios. Fig. 5 (d) shows that the median of the RMSE ratios between the EM-Cross model and the EM-Within model deceases as γ increases and the median becomes smaller than 1 when γ exceeds 1. This verifies the effectiveness of the proposed decision score in selecting between the within-output and cross-output modeling.

To further verify the accuracy of the proposed method, we compared the RMSEs of the proposed method with the RMSEs of the LMC model and MC method. The RMSE ratios of EM-Cross to LMC and EM-Within to LMC at different values of γ were computed and shown in Fig. 6 (a) and (b), while the RMSE ratios of EM-Cross to MC and EM-Within to MC are respectively shown in Fig. 6 (c) and (d). From the figure, we can find both the EM-Cross and EM-Within model achieved better results than the LMC model as well as the MC method for all values of γ. These results demonstrate that modeling the correlation structure in the original input space is not effective when the data have many segments of missing values, and thus justifies our selection of a latent variable model in this study. Moreover, from Fig. 6, we can find the EM-Cross model performs better when $\gamma > 1$ while the EM-Within model provides better results when $\gamma < 1$, which also justifies the effectiveness of the proposed decision score in choosing between the EM-Cross model and the EM-Within model.

B. Case Study

In this case study, the transfer characteristic curves of a GFET are used to further validate the proposed method. GFET is a type of popular transistors and the high carrier mobility of graphene has made it a competitive product for high electron

Authorized licensed use limited to: Peking University. Downloaded on March 08,2023 at 08:00:25 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

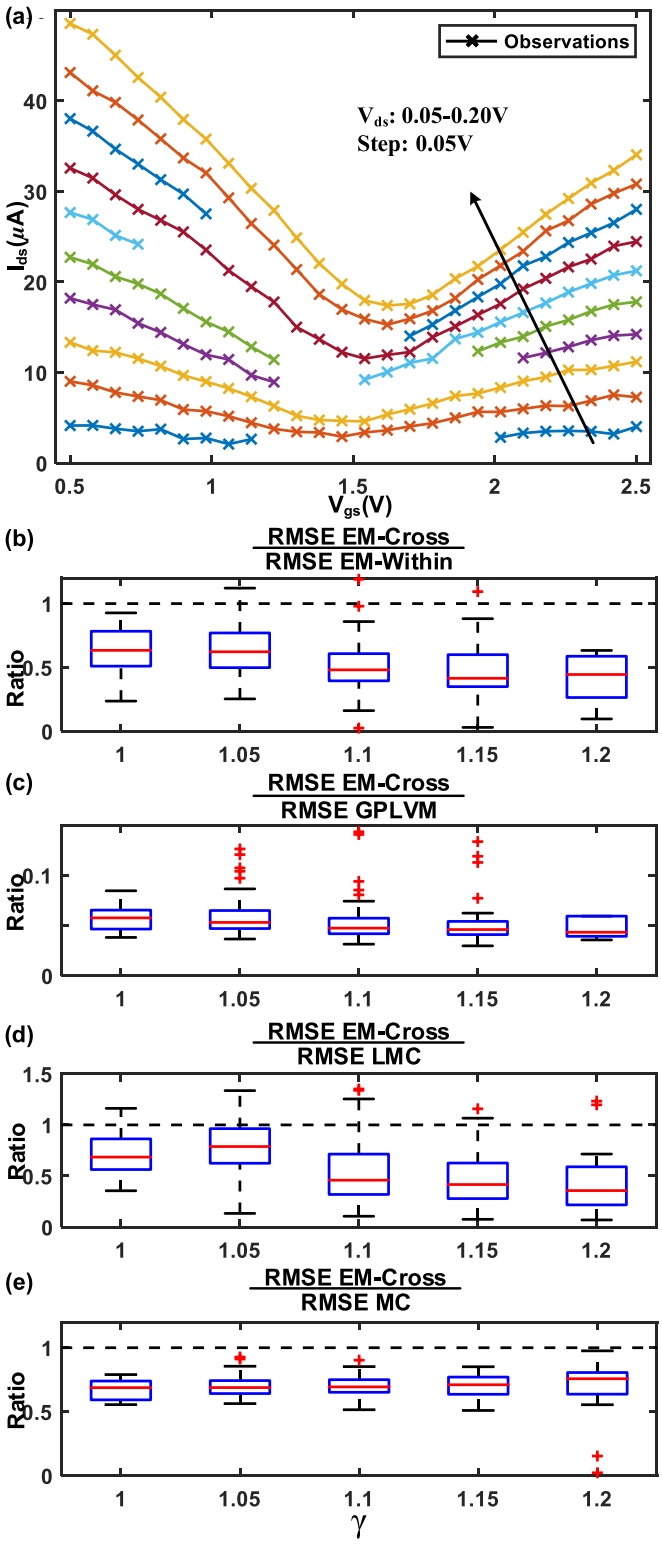


Fig. 7. Results for the case study. (a) Data with $\gamma = 1.1$. RMSE ratio between (b) EM-Cross and EM-Within, (c) EM-Cross and GPLVM, (d) EM-Cross and LMC, (e) EM-Cross and MC.

mobility transistors. With a more sophisticated analysis of the carrier mobility, Tian et al. [27] has come up with a model that can accurately predict the measurements from different GFETs. Therefore, this model is used to generate the data for the case study.

The transfer characteristic curve of an GFET describes the evolution of the drain-to-source current I_{ds} with respect to the gate-to-source voltage V_{gs} at different drain-to-source voltages V_{ds}. In this study, the characteristic curves of a 5-μm GFET are simulated with the parameters presented in [27]. In total, 10 different curves are produced by varying the drain to source voltage between 0.05V to 0.5V. The observation noise with a distribution $\mathcal{N}(0, 0.2^2)$ are added to the curves to generate the complete data. To generate the missing data, we randomly choose 5 different curves and a segment with 30% to 60% of the whole observations is set as missing within each curve. Fig. 7 (a) shows an example of the data with $\gamma = 1.1$.

The modeling results are displayed in Fig. 7 (b)−(e), which show the box plots of RMSE ratios for different values of γ. For the data in this case study, the rows and columns are all highly correlated with each other, i.e., $\rho_c/\rho_r \approx 1$, and thereby the decision score is mainly determined by the missing pattern, i.e., $\gamma = (\tau_c/\rho_c)/(\tau_r/\rho_r) \approx \tau_c/\tau_r$. In addition, $\tau_c \geq \tau_r$ in all experiments, which indicates the data contains more within-output information loss. Therefore, the values of γ are greater than or equal to one in all the experiments. Fig. 7 (b) shows the EM-Cross model has smaller RMSEs than the EM-Within model in all the experiments. This is expected as $\gamma \geq 1$ suggests that the EM-Cross model will provide better estimations, which further justifies the effectiveness of the proposed decision score. Fig. 7 (c)−(e) exhibit that the EM-Cross model generates consistently better results than other methods, which also validates the advantage of the proposed method in dealing multi-output data with segmented missing values.

V. Conclusion

This paper develops a new method based on the GPLVM theory to model multi-output data with missing data in segmented patterns. The method has a tailored EM algorithm to sequentially impute missing values while optimizing the model parameters. In addition, a decision score that quantifies the missing pattern and correlation is designed to choose between within- and cross-output modeling. With a simulation study and a case study, the proposed method was tested to perform better than the standard GPLVM, the LMC-based MOGP, and the MC method. The decision score was verified as a good indicator of the best modeling strategies (within- or cross-output modeling). The customized EM algorithm was shown to significantly boost the accuracy of standard GPLVM and attained good performance even when the outputs have lost 60% of the observations. This makes the proposed method a useful tool for different industries that require multi-output modeling with missing data.

Despite the improved performance of our method in dealing with incomplete data, there are some questions worth further investigation. In this paper, we assume the input space of different outputs are of same dimension in Assumption A1, which can be relaxed when modeling multi-output data with different input space, e.g., when the input space of one output is time and the input space for the other output is time and location. Also, the noise in our model is assumed to be

Authorized licensed use limited to: Peking University. Downloaded on March 08,2023 at 08:00:25 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

i.i.d., which can be violated in some applications, such as astronomical datasets. Thus, further studies can be done to investigate the extension to applications with correlated noise. Moreover, another important category of missing data is that with missing not at random, which requires the modeling of missing pattern. In the future, an extension for the investigation of such missing data will also be valuable.

APPENDIX A
PROOF OF LEMMA 1

To prove the lemma, we will firstly obtain a lower bound for the log marginal likelihood of complete data $\mathbf{Y}$. A bunch of inducing points are introduced, which are the function values at a group of auxiliary latent inputs $\mathbf{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_J]^T$. We denote $\mathbf{u}_m = [f_m(\mathbf{z}_1), \cdots, f_m(\mathbf{z}_J)]^T$ as the inducing points for the m^{th} output, and $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_M]^T$ includes all of the inducing points. Based on the i.i.d. GP assumption, the conditional density of $\mathbf{F}$ given $\mathbf{U}$ and $\mathbf{X}$ is:

$$p(\mathbf{F}|\mathbf{U},\mathbf{X}) = \prod_{m=1}^{M} \mathcal{MVN}(\mathbf{f}_m|\boldsymbol{\alpha}_m, \boldsymbol{\Omega}_{mm}) \tag{7}$$

where $\boldsymbol{\alpha}_m = \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{u}_m$ and $\boldsymbol{\Omega}_{mm} = \mathbf{K}_{ff} - \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}\mathbf{K}_{uf}$ are respectively the mean and covariance. The matrices $\mathbf{K}_{\omega\nu}(\omega, \nu = f, u)$ are the covariance matrix between different components. For example, $\mathbf{K}_{fu}$ is the covariance between $\mathbf{f}_m$ and $\mathbf{u}_m$ $(m, n = 1, \cdots, M)$ with $k(\mathbf{x}_i, \mathbf{z}_j)$ as the element at the i^{th} row and j^{th} column. Then a special variational density approximating the posterior density $p(\mathbf{F}, \mathbf{U}, \mathbf{X}|\mathbf{Y})$ is introduced, that is, $p(\mathbf{F}, \mathbf{U}, \mathbf{X}|\mathbf{Y}) \approx q(\mathbf{F}, \mathbf{U}, \mathbf{X}) = p(\mathbf{F}|\mathbf{X}, \mathbf{Y})q(\mathbf{U})q(\mathbf{X})$, wherer $q(\mathbf{U}$ and $q(\mathbf{X})$ are the variational approximation for $p(\mathbf{U}|\mathbf{Y})$ and $p(\mathbf{X}|\mathbf{Y})$, respectively. Based on this, we have

$$\begin{aligned}
&\log p(\mathbf{Y}) \\
&= \log\Big\{\int p(\mathbf{Y}|\mathbf{F})p(\mathbf{F}|\mathbf{U},\mathbf{X})p(\mathbf{U})p(\mathbf{X})d\mathbf{F}d\mathbf{U}d\mathbf{X}\Big\} \\
&= \log\Big\{\int p(\mathbf{F}|\mathbf{U},\mathbf{X})q(\mathbf{U})q(\mathbf{X})\frac{p(\mathbf{Y}|\mathbf{F})p(\mathbf{U})p(\mathbf{X})}{q(\mathbf{U})q(\mathbf{X})}d\mathbf{F}d\mathbf{U}d\mathbf{X}\Big\} \\
&\geq \int p(\mathbf{F}|\mathbf{U},\mathbf{X})q(\mathbf{U})q(\mathbf{X})\log\Big\{\frac{p(\mathbf{Y}|\mathbf{F})p(\mathbf{U})p(\mathbf{X})}{q(\mathbf{U})q(\mathbf{X})}d\mathbf{F}d\mathbf{U}d\mathbf{X}\Big\} \\
&= \int q(\mathbf{U})q(\mathbf{X})p(\mathbf{F}|\mathbf{U},\mathbf{X})\log\Big\{\frac{p(\mathbf{Y}|\mathbf{F})p(\mathbf{U})}{q(\mathbf{U})}\Big\}d\mathbf{F}d\mathbf{U}d\mathbf{X} \\
&\quad - \mathcal{KL}\big(q(\mathbf{X})||p(\mathbf{X})\big)
\end{aligned} \tag{8}$$

where $\mathcal{KL}\big(q(\mathbf{X})||p(\mathbf{X})\big)$ is the Kullback-Leibler divergence between $q(\mathbf{X})$ and $p(\mathbf{X})$. It should be noted that model parameters have been omitted for notational simplicity. To calculate the first term, we have:

$$\begin{aligned}
h(\mathbf{U},\mathbf{X}) &= \int p(\mathbf{F}|\mathbf{U},\mathbf{X})\log p(\mathbf{Y}|\mathbf{F})d\mathbf{F} \\
&= \sum_{m=1}^{M}\int p(\mathbf{f}_m|\mathbf{u}_m,\mathbf{X})\log p(\mathbf{y}_m|\mathbf{f}_m)d\mathbf{f}_m \\
&= \sum_{m=1}^{M}[\log\{\mathcal{N}(\mathbf{y}_m|\boldsymbol{\alpha}_m, \beta^{-1}\boldsymbol{I}_N)\} - \frac{\beta}{2}tr(\boldsymbol{\Omega}_{mm})]
\end{aligned} \tag{9}$$

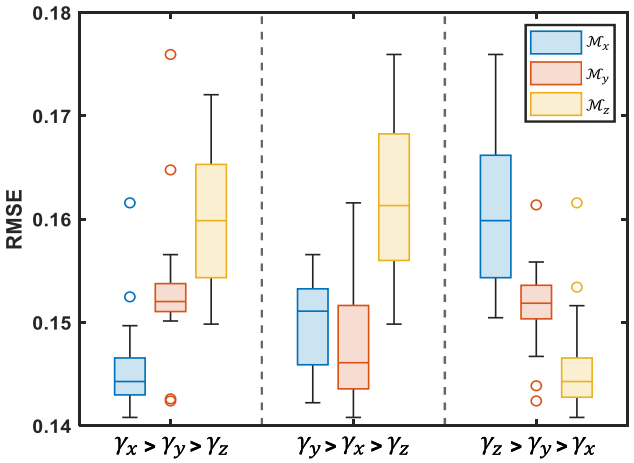


Fig. 8. Results for the simulation study with 2D input case.

where $tr(\cdot)$ represents the trace. Then, we further have

$$\begin{aligned}
\widetilde{h}(\mathbf{U}) &= \int q(\mathbf{X})h(\mathbf{U},\mathbf{X})d\mathbf{X} \\
&= \sum_{m=1}^{M}\Big\{\frac{N}{2}\log\frac{\beta}{2\pi} - \frac{\beta}{2}\mathbf{y}_m^T\mathbf{y}_m + \beta\mathbf{y}_m^T\mathbb{E}_{q(\mathbf{X})}[\boldsymbol{\alpha}_m] \\
&\quad - \frac{\beta}{2}\mathbb{E}_{q(\mathbf{X})}[\boldsymbol{\alpha}_m^T\boldsymbol{\alpha}_m] - tr\big(\mathbb{E}_{q(\mathbf{X})}[\boldsymbol{\Omega}_{mm}]\big)\Big\}
\end{aligned} \tag{10}$$

where $\mathbb{E}_{q(\mathbf{X})}[\cdot]$ represents the expectation with respect to the density $q(\mathbf{X})$. Then using Jensen's inequality, we have

$$\begin{aligned}
&\int q(\mathbf{U})\Big\{\widetilde{h}(\mathbf{U}) + \log\frac{p(\mathbf{U})}{q(\mathbf{U})}\Big\}d\mathbf{U} \leq \log\Big\{\int q(\mathbf{U})\frac{e^{\widetilde{h}(\mathbf{U})}p(\mathbf{U})}{q(\mathbf{U})}d\mathbf{U}\Big\} \\
&= \log\Big\{\int e^{\widetilde{h}(\mathbf{U})}p(\mathbf{U})d\mathbf{U}\Big\} = \sum_{m=1}^{M}\Big\{\frac{1}{2}\log\frac{(\beta/2)^N|\mathbf{K}_{uu}|}{|\beta\boldsymbol{\Psi}_2 + \mathbf{K}_{uu}|} \\
&\quad - \frac{\beta}{2}\Big[\mathbf{y}_m^T\big(\boldsymbol{I}_N - \beta\boldsymbol{\Psi}_1(\beta\boldsymbol{\Psi}_2 + \mathbf{K}_{uu})^{-1}\boldsymbol{\Psi}_1^T\big)\mathbf{y}_m \\
&\quad + \psi_0 - tr(\mathbf{K}_{uu}^{-1}\boldsymbol{\Psi}_2)\Big]\Big\}
\end{aligned} \tag{11}$$

where $\psi_0 = tr(\mathbb{E}_{q(\mathbf{X})}[\mathbf{K}_{ff}])$, $\boldsymbol{\Psi}_1 = \mathbb{E}_{q(\mathbf{X})}[\mathbf{K}_{fu}]$ and $\boldsymbol{\Psi}_2 = \mathbb{E}_{q(\mathbf{X})}[\mathbf{K}_{uf}\mathbf{K}_{fu}]$. Plugging this result into Eq. 8, we can obtain the lower bound as follows:

$$\log p(\mathbf{Y}) \geq -\frac{\beta}{2}\sum_{m=1}^{M}\mathbf{y}_m^T\big(\boldsymbol{I}_N - \mathbf{W}(\boldsymbol{\theta})\big)\mathbf{y}_m + \mathcal{F}(\boldsymbol{\theta}) \tag{12}$$

where $\mathbf{W}(\boldsymbol{\theta}) = \beta\boldsymbol{\Psi}_1(\beta\boldsymbol{\Psi}_2 + \mathbf{K}_{uu})^{-1}\boldsymbol{\Psi}_1^T$ and $\mathcal{F}(\boldsymbol{\theta}) = \frac{1}{2}\log\big(\frac{\beta^N|\mathbf{K}_{uu}|}{(2\pi)^N|\beta\boldsymbol{\Psi}_2+\mathbf{K}_{uu}|} - \frac{\beta}{2}(\psi_0 - tr(\mathbf{K}_{uu}^{-1}\boldsymbol{\Psi}_2))$. To simplify the notation, we will drop the $\boldsymbol{\theta}$ in $\mathbf{W}(\boldsymbol{\theta})$ and $\mathcal{F}(\boldsymbol{\theta})$.

Under the scenario of missing values, the data we got is $\mathbf{Y}^{\mathcal{O}}$. Using the Bayesian theorem, the log marginal likelihood of $\mathbf{Y}^{\mathcal{O}}$ can be expanded as follows:

$$\log p(\mathbf{Y}^{\mathcal{O}};\boldsymbol{\theta}) = \log p(\mathbf{Y}^{\mathcal{O}},\mathbf{Y}^{\mathcal{U}};\boldsymbol{\theta}) - \log p(\mathbf{Y}^{\mathcal{U}}|\mathbf{Y}^{\mathcal{O}};\boldsymbol{\theta}) \tag{13}$$

where $\boldsymbol{\theta}$ is introduced to specify the dependence on some unknown model parameters. By taking integration with respect

Authorized licensed use limited to: Peking University. Downloaded on March 08,2023 at 08:00:25 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

to $p(\mathbf{Y}^{\mathcal{U}}|\mathbf{Y}^{\mathcal{O}};\boldsymbol{\theta}^*)$ on both sides of Eq. 13, we have:

$$\int p(\mathbf{Y}^{\mathcal{U}}|\mathbf{Y}^{\mathcal{O}};\boldsymbol{\theta}^*)\log p(\mathbf{Y}^{\mathcal{O}};\boldsymbol{\theta})d\mathbf{Y}^{\mathcal{U}} = \Lambda(\boldsymbol{\theta},\boldsymbol{\theta}^*) - \Gamma(\boldsymbol{\theta},\boldsymbol{\theta}^*) \tag{14}$$

where $\Lambda(\boldsymbol{\theta},\boldsymbol{\theta}^*) = \int p(\mathbf{Y}^{\mathcal{U}}|\mathbf{Y}^{\mathcal{O}};\boldsymbol{\theta}^*)\log p(\mathbf{Y}^{\mathcal{U}},\mathbf{Y}^{\mathcal{O}};\boldsymbol{\theta})d\mathbf{Y}^{\mathcal{U}}$ and $\Gamma(\boldsymbol{\theta},\boldsymbol{\theta}^*) = \int p(\mathbf{Y}^{\mathcal{U}}|\mathbf{Y}^{\mathcal{O}};\boldsymbol{\theta}^*)\log p(\mathbf{Y}^{\mathcal{U}}|\mathbf{Y}^{\mathcal{O}};\boldsymbol{\theta})d\mathbf{Y}^{\mathcal{U}}$. The integration on the left hand side is just $\log p(\mathbf{Y}^{\mathcal{O}};\boldsymbol{\theta})$. By Jensen's inequality, $\Gamma(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ is maximized as a function of $\boldsymbol{\theta}$ when $\boldsymbol{\theta}=\boldsymbol{\theta}^*$ [28]. Thus, if $\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}$ maximizes $\Lambda(\boldsymbol{\theta},\boldsymbol{\theta}^*)$, we have:

$$\begin{aligned}&\log p(\mathbf{Y}^{\mathcal{O}};\widehat{\boldsymbol{\theta}}) - \log p(\mathbf{Y}^{\mathcal{O}};\boldsymbol{\theta}^*)\\ &\quad= \big[\Lambda(\widehat{\boldsymbol{\theta}},\boldsymbol{\theta}^*) - \Lambda(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)\big] - \big[\Gamma(\widehat{\boldsymbol{\theta}},\boldsymbol{\theta}^*) - \Gamma(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*)\big]\\ &\quad\geq \Lambda(\widehat{\boldsymbol{\theta}},\boldsymbol{\theta}^*) - \Lambda(\boldsymbol{\theta}^*,\boldsymbol{\theta}^*) \geq 0\end{aligned} \tag{15}$$

This means that the maximization of $\log p(\mathbf{Y}^{\mathcal{O}};\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is equivalent to the maximization of $\Lambda(\boldsymbol{\theta},\boldsymbol{\theta}^*)$. Then, from Eq. 8, we have:

$$\begin{aligned}&\Lambda(\boldsymbol{\theta},\boldsymbol{\theta}^*)\\ &= \int p(\mathbf{Y}^{\mathcal{U}}|\mathbf{Y}^{\mathcal{O}};\boldsymbol{\theta}^*)\log p(\mathbf{Y}^{\mathcal{O}},\mathbf{Y}^{\mathcal{U}};\boldsymbol{\theta})d\mathbf{Y}^{\mathcal{U}}\\ &\geq \int p(\mathbf{Y}^{\mathcal{U}}|\mathbf{Y}^{\mathcal{O}};\boldsymbol{\theta}^*)\Big\{-\frac{\beta}{2}\sum_{m=1}^{M}\Big[\mathbf{y}_m^T\big(\boldsymbol{I}_N - \mathbf{W}(\boldsymbol{\theta})\big)\mathbf{y}_m\Big]\\ &\quad+\mathcal{F}(\boldsymbol{\theta})\Big\}d\mathbf{Y}^{\mathcal{U}}\\ &= -\frac{\beta}{2}\sum_{m=1}^{M}\int\delta\big(\mathbf{y}_m^{\mathcal{U}} - \widehat{\boldsymbol{\lambda}}_m^{\mathcal{U}}\big)\Big[\mathbf{y}_m^T\big(\boldsymbol{I}_N - \mathbf{W}(\boldsymbol{\theta})\big)\mathbf{y}_m\Big]d\mathbf{y}_m^{\mathcal{U}} + \mathcal{F}(\boldsymbol{\theta})\\ &= -\frac{\beta}{2}\sum_{m=1}^{M}\widetilde{\mathbf{y}}_m^T\big(\boldsymbol{I}_N - \mathbf{W}(\boldsymbol{\theta})\big)\widetilde{\mathbf{y}}_m + \mathcal{F}(\boldsymbol{\theta})\end{aligned} \tag{16}$$

where $\widetilde{\mathbf{y}}_m$ is obtained by replacing the unobserved part in $\mathbf{y}_m$ with $\widehat{\boldsymbol{\lambda}}_m^{\mathcal{U}}$, i.e., $\widetilde{\mathbf{y}}_m^{\mathcal{O}} = \mathbf{y}_m^{\mathcal{O}}$ and $\widetilde{\mathbf{y}}_m^{\mathcal{U}} = \widehat{\boldsymbol{\lambda}}_m^{\mathcal{U}}$. It should be noted that $\widehat{\boldsymbol{\lambda}}_m^{\mathcal{U}}$ is a function of $\boldsymbol{\theta}^*$. Therefore, the maximization of $\log p(\mathbf{Y}^{\mathcal{O}};\boldsymbol{\theta})$ is equivalent to the maximization of the following lower bound:

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{\beta}{2}\sum_{m=1}^{M}\widetilde{\mathbf{y}}_m^T\big(\mathbf{I}_N - \mathbf{W}(\boldsymbol{\theta})\big)\widetilde{\mathbf{y}}_m + \mathcal{F}(\boldsymbol{\theta}) \tag{17}$$

This completes the proof of Lemma 1.

APPENDIX B
SIMULATION STUDY FOR HIGH-DIMENSIONAL INPUT

In order to demonstrate the performance of the proposed method in high-dimensional input scenarios, a simulation study with two-dimensional (2D) input is conducted here. To generate complete data, the following model is adopted:

$$\begin{aligned}v_z(x,y) = sin\big(\omega_x(x+\phi_x) + \omega_y(y+\phi_y) + \omega_z(z+\phi_z)\big)\\ +\epsilon_z(x,y)\end{aligned} \tag{18}$$

where $\epsilon_z(x,y)$ is observation noise with i.i.d Gaussian distribution $\mathcal{N}(0,\sigma^2)$. The correlation strength across different dimensions is controlled through the variation of angular frequencies ω_x, ω_y, ω_z and phases ϕ_x, ϕ_y, ϕ_z. In this simulation, the phase shift in different dimension is generated by uniformly sampling from the interval $[0,\ 0.2\pi]$. Under the scenario of 2D input case, our modeling structure provides three different options, i.e., x-direction, y-direction and z-direction. To choose from these options, decision scores are calculated following the procedure described in Section III-B. We denote the decision score for each direction as γ_x, γ_y and γ_z. Similarly, the model constructed are respectively denoted as $\mathcal{M}_x$, $\mathcal{M}_y$ and $\mathcal{M}_z$.

The results for the simulation study are shown in Fig. 8, which are separated into three different groups based on the relative value of decision scores. Based on the results, we can find that the larger the decision score the smaller the RMSE value. This validates the effectiveness of the proposed decision score in determining the best modeling choice under high dimensional input scenarios. In addition, under high dimensional input case, the complex missing pattern and correlation structure inhibit decision making based on visual check, which justifies the necessity of the proposed decision score. Therefore, the proposed modeling structure is flexible even in high-dimensional input scenarios.

REFERENCES

[1] M. S. Tootooni, P. K. Rao, C.-A. Chou, and Z. J. Kong, "A spectral graph theoretic approach for monitoring multivariate time series data from complex dynamical processes," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 1, pp. 127–144, Jan. 2018.

[2] C. Xiao, M. Yu, B. Zhang, H. Wang, and C. Jiang, "Discrete component prognosis for hybrid systems under intermittent faults," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 4, pp. 1766–1777, Oct. 2021.

[3] Z. Hu and C. Wang, "Nonlinear online multioutput Gaussian process for multistream data informatics," *IEEE Trans. Ind. Informat.*, vol. 18, no. 6, pp. 3885–3893, Jun. 2022.

[4] K. Liu, Y. Li, X. Hu, M. Lucu, and W. D. Widanage, "Gaussian process regression with automatic relevance determination kernel for calendar aging prediction of lithium-ion batteries," *IEEE Trans. Ind. Informat.*, vol. 16, no. 6, pp. 3767–3777, Jun. 2020.

[5] K. W. Poh, "Stress-strain-temperature relationship for structural steel," *J. Mater. Civil Eng.*, vol. 13, no. 5, pp. 371–379, Oct. 2001.

[6] J. Zhu, N. Chen, and C. Shen, "A new multiple source domain adaptation fault diagnosis method between different rotating machines," *IEEE Trans. Ind. Informat.*, vol. 17, no. 7, pp. 4788–4797, Jul. 2021.

[7] X. Fang, R. Zhou, and N. Gebraeel, "An adaptive functional regression-based prognostic model for applications with missing data," *Rel. Eng. Syst. Saf.*, vol. 133, pp. 266–274, Jan. 2015.

[8] M. A. Alvarez, L. Rosasco, and N. D. Lawrence, "Kernels for vector-valued functions: A review," 2011, *arXiv:1106.6251*.

[9] A. Fallahdizcheh and C. Wang, "Profile monitoring based on transfer learning of multiple profiles with incomplete samples," *IISE Trans.*, vol. 54, no. 7, pp. 643–658, 2022.

[10] S. A. Imtiaz and S. L. Shah, "Treatment of missing values in process data analysis," *Can. J. Chem. Eng.*, vol. 86, no. 5, pp. 838–858, Oct. 2008.

[11] F. Rodrigues, K. Henrickson, and F. C. Pereira, "Multi-output Gaussian processes for crowdsourced traffic data imputation," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 2, pp. 594–603, Feb. 2019.

[12] C. Song, K. Liu, and X. Zhang, "A generic framework for multisensor degradation modeling based on supervised classification and failure surface," *IISE Trans.*, vol. 51, no. 11, pp. 1288–1302, Nov. 2019.

[13] N. Srebro, J. Rennie, and T. Jaakkola, "Maximum-margin matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 17, 2004, pp. 1–8.

[14] B. Recht, "A simpler approach to matrix completion," *J. Mach. Learn. Res.*, vol. 12, no. 12, pp. 1–18, 2011.

[15] X. Chen, Z. He, and L. Sun, "A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation," *Transp. Res. C, Emerg. Technol.*, vol. 98, pp. 73–84, Jan. 2019.

[16] Q. Song, H. Ge, J. Caverlee, and X. Hu, "Tensor completion algorithms in big data analytics," *ACM Trans. Knowl. Discovery From Data*, vol. 13, no. 1, pp. 1–48, Feb. 2019.

Authorized licensed use limited to: Peking University. Downloaded on March 08,2023 at 08:00:25 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[17] J. Xue, Y. Zhao, Y. Bu, J. C.-W. Chan, and S. G. Kong, "When Laplacian scale mixture meets three-layer transform: A parametric tensor sparsity for tensor completion," *IEEE Trans. Cybern.*, vol. 52, no. 12, pp. 13887–13901, Dec. 2022.

[18] J. Honaker and G. King, "What to do about missing values in time-series cross-section data," *Amer. J. Political Sci.*, vol. 54, no. 2, pp. 561–581, Apr. 2010.

[19] A. C. Damianou, M. K. Titsias, and N. Lawrence, "Variational inference for latent variables and uncertain inputs in Gaussian processes," *J. Mach. Learn. Res.*, vol. 17, 2016.

[20] G. Parra and F. Tobar, "Spectral mixture kernels for multi-output Gaussian processes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6684–6693.

[21] J. Zhao and S. Sun, "Variational dependent multi-output Gaussian process dynamical systems," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 4134–4169, 2016.

[22] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning* (Adaptive Computation and Machine Learning). Cambridge, MA, USA: MIT Press, Jan. 2006.

[23] R. J. Little and D. B. Rubin, *Statistical Analysis With Missing Data*, vol. 793. Hoboken, NJ, USA: Wiley, 2019.

[24] A. Fallahdizcheh and C. Wang, "Profile monitoring based on transfer learning of multiple profiles with incomplete samples," *IISE Trans.*, vol. 54, no. 7, pp. 1–41, 2021.

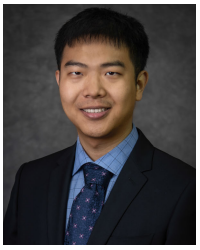
[25] A. Girard, C. E. Rasmussen, J. Quinonero-Candela, and R. Murray-Smith, "Gaussian process priors with uncertain inputs: Application to multiple-step ahead time series forecasting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 15, 2022.

[26] A. Majumdar and R. K. Ward, "Some empirical advances in matrix completion," *Signal Process.*, vol. 91, no. 5, pp. 1334–1338, 2011.

[27] J. Tian, A. Katsounaros, D. Smith, and Y. Hao, "Graphene field-effect transistor model with improved carrier mobility analysis," *IEEE Trans. Electron Devices*, vol. 62, no. 10, pp. 3433–3440, Oct. 2015.

[28] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Cham, Switzerland: Springer, 2009.

Zhiyong Hu received the B.E. degree in mechanical engineering from the Anhui University of Technology, Maanshan, China, in 2011, the M.E. degree in precision instrument and machinery from the University of Science and Technology of China in 2016, and the Ph.D. degree from the Department of Industrial, Manufacturing and Systems Engineering, Texas Tech University, in 2021. He is currently a Post-Doctoral Researcher with the Department of Precision Machinery and Precision Instrumentation, University of Science and Technology of China. His research interests include risk modeling and survival analysis of medical datasets, and the development of efficient statistical models for complex spatial–temporal process to understand the mechanism of cardiac diseases as well as assist clinical decision making.



Chao Wang received the B.S. degree in mechanical engineering from the Hefei University of Technology in 2012, the M.S. degree in mechanical engineering from the University of Science and Technology of China in 2015, and the M.S. degree in statistics and the Ph.D. degree in industrial and systems engineering from the University of Wisconsin–Madison in 2018 and 2019, respectively. He is currently an Assistant Professor with the Department of Industrial and Systems Engineering, The University of Iowa, Iowa City, IA, USA. His research interests include statistical modeling, analysis, monitoring, and control for complex systems. He is a member of INFORMS, IISE, and SME.

Jianguo Wu (Member, IEEE) received the B.S. degree in mechanical engineering from Tsinghua University, Beijing, China, in 2009, the M.S. degree in mechanical engineering from Purdue University in 2011, and the M.S. degree in statistics and the Ph.D. degree in industrial and systems engineering from the University of Wisconsin–Madison in 2014 and 2015, respectively. He was an Assistant Professor with the Department of Industrial and Manufacturing Systems Engineering (IMSE), The University of Texas at El Paso (UTEP), El Paso, TX, USA, from 2015 to 2017. He is currently an Assistant Professor with the Department of Industrial Engineering and Management, Peking University, Beijing. His research interests include data-driven modeling, monitoring and analysis of advanced manufacturing processes and complex systems for quality control and reliability improvement, and quality control and reliability engineering of intelligent manufacturing and complex systems through engineering informed machine learning and advanced data analytics. He is a member of INFORMS, IISE, and SME. He was a recipient of the STARS Award from the University of Texas Systems, the Overseas Distinguished Young Scholars from China, the P&G Faculty Fellowship, the BOSS Award from MSEC, and several best paper award/finalists from INFORMS/IISE Annual Meeting. He is an Associate Editor of the *Journal of Intelligent Manufacturing*.

Dongping Du (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from the China University of Mining and Technology, Beijing, China, and the Ph.D. degree in industrial engineering from the University of South Florida, Tampa, FL, USA. She is currently an Associate Professor with the Department of Industrial, Manufacturing, and Systems Engineering, Texas Tech University, Lubbock, TX, USA. Her research interests include computer modeling and simulation, applied statistics and machine learning, survival analysis, and risk assessment with applications in health care and manufacturing. She is a member of INFORMS and IISE.

Authorized licensed use limited to: Peking University. Downloaded on March 08,2023 at 08:00:25 UTC from IEEE Xplore. Restrictions apply.