Transfer Learning of Stochastic Kriging for Individualized Prediction

Jinwei Yao, Jianguo Wu, Yongxiang Li, and Chao Wang

Abstract—Stochastic Kriging (SK) is a generalized variant of Gaussian process regression, and it is developed for dealing with non-i.i.d. noise in functional responses. Although SK has achieved substantial success in various engineering applications, its intrinsic modeling strategy by focusing on the sample mean limits its flexibility and capability of predicting individual functional samples. Moreover, the performance of SK can be impaired under scarce data scenarios, which are commonly encountered in engineering applications, especially for start-up or just deployed systems. In this paper, we propose a novel transfer learning framework to address the challenges of individualization and data scarcity in traditional SK. The proposed framework features a within-process model to facilitate individualized prediction and a between-process model to leverage information from related processes for resolving the issue of data scarcity. The within- and between-process models are integrated through a tailored convolution process, which quantifies interactions within and between processes using a specially designed covariance matrix and corresponding kernel parameters. Statistical properties are investigated on the parameter estimation of the proposed framework, which provide theoretical guarantees for the performance of transfer learning. The proposed method is compared with benchmark methods through various numerical and real case studies, and the results demonstrate the superiority of the proposed method in dealing with individualized prediction of functional responses, especially when limited data are available in the process of interest.

Index Terms—Gaussian process, Stochastic Kriging, Transfer learning, Non-i.i.d. noise

NOMENCLATURE

I	Number of processes
I_t	Index of a replication in the Ith process
I_r	Index for the replications in the Ith process except for I_t
S	The set of source processes
m_i	Number of replications for the ith process
$\boldsymbol{x}_i$	Total input locations in the ith process
$x_{i,j}$	The jth input location in $\boldsymbol{x}_i$
n_i	Number of total input locations in the ith process
$\boldsymbol{x}_i^{(m)}$	Input locations of the mth replication in the ith process
$\boldsymbol{y}_i^{(m)}(\boldsymbol{x}_i^{(m)})$	Data sample vector of the mth replication in the ith process
$\bar{\boldsymbol{y}}_i$	Sample mean of $\boldsymbol{y}_i^{(m)}(\boldsymbol{x}_i^{(m)})$ across all ms in the ith process
$\bar{\boldsymbol{y}}$	Sample mean in every process $\bar{\boldsymbol{y}} = [\bar{\boldsymbol{y}}_1^T, \bar{\boldsymbol{y}}_2^T, \cdots, \bar{\boldsymbol{y}}_{I_r}^T, \bar{\boldsymbol{y}}_{I_t}^T]^T$
$\boldsymbol{\epsilon}_i^{(m)}(\boldsymbol{x}_i^{(m)})$	Noise vector of the mth replication in the ith process
$R_{i,j}$	The set of indices of replications that have data point(s) at $x_{i,j}$.
$\boldsymbol{f}_i(\boldsymbol{x}_i)$	Mean trend of the ith process on $\boldsymbol{x}_i$
$\boldsymbol{\mathcal{K}}_{i,i'}$	Covariance matrix for the mean trend between the ith and i'th processes
$\bar{\boldsymbol{\Sigma}}_{i,i'}$	Covariance matrix for the averaged noise between the ith and i'th processes
$\boldsymbol{\mathcal{K}}_{i,i}^*$	Covariance matrix for the mean trend of the ith process on new inputs $\boldsymbol{x}_i^*$
$\dot{\boldsymbol{\mathcal{K}}}_{i,i}$	Covariance matrix for the mean trend of the ith process between $\boldsymbol{x}_i$ and $\boldsymbol{x}_i^*$
$\boldsymbol{\Omega}$	Covariance matrix for $\bar{\boldsymbol{y}}$
$\boldsymbol{\Omega}_{S,I}$	Covariance matrix between sources and the target
$Z_e(\cdot)$	The eth Gaussian white noise process
$g_{e,i}(\cdot)$	The kernel convolved with $Z_e(\cdot)$ for the ith process
$\alpha_{i,i'}$	The scaling parameter in $g_{i,i'}(\cdot)$
$\beta_{i,i'}$	The length-scale parameter in $g_{i,i'}(\cdot)$
$\boldsymbol{\theta}$	The parameter set for $\boldsymbol{\Omega}$
$\boldsymbol{\theta_0}$	The penalty parameter set
$L(\boldsymbol{\theta}\|\bar{\boldsymbol{y}})$	The log-likelihood function for $\bar{\boldsymbol{y}}$
$\mathbb{P}_\gamma(\boldsymbol{\theta_0})$	The penalization function

1 INTRODUCTION

Gaussian process regression (GPR) is a probabilistic machine learning method that models functional relationships as realizations of random processes [1]. Due to its flexible modeling capability and elegant mathematical properties, the GPR has been widely used in various applications such as surrogate modeling, geostatistics, and spatial modeling [2, 3]. Nevertheless, the conventional GPR is built on the assumption that the noise term is independent and identically distributed (i.i.d.), which limits its broader applications [4].

- *Jinwei Yao and Chao Wang are with Department of Industrial and Systems Engineering, University of Iowa, Iowa City, IA, 52242 USA.*
- *Jianguo Wu is with Department of Industrial Engineering and Management, Peking University, Beijing, 100871 China.*
- *Yongxiang Li is with the Department of Industrial Engineering and Management, Shanhai Jiao Tong University, Shanghai, 200240 China.*

Corresponding author: Chao Wang, email: chao-wang-2@uiowa.edu.

In practice, the sources of noise vary significantly due to data generation mechanisms and data collection techniques/environments, leading to the violation of the i.i.d. noise assumption. A real-world example that illustrates this problem is shown in Fig. 1, where there are three testing processes for battery impedance. In each process, multiple batteries are tested at various frequencies, and each battery's testing results are plotted as blue dots along its unobservable function curve (i.e., one solid function curve corresponds to one battery). The (unobservable) mean function in each process is also marked as the dashed line. We denote the data from one specific battery as one functional replication from its process. Note that each battery (and its functional replication) in the process is distinct from the others because each battery's charging and recharging cycles are different. The data features in Fig. 1 can be summarized as follows:

- *Heterogeneous replications*: Each battery has not only distinct functional behaviors but also heterogeneous input locations due to different charging-recharging cycles and conditions.
- *Within-and between-process correlations*: The replications within-and between-process both have strong correlations, i.e., non-i.i.d. features. However, it is important to notice the sources of these two correlations are different: The within-correlation stems from the same testing process conditions, e.g., temperature; while the between-correlation is governed by the general frequency vs. impedance physics for the same type of batteries.
- *Heterogeneous availability*: The number of batteries (replications) in each testing process can be different. Typically, the new process has a much smaller number of replications.

These features are commonly observed in various applications, see examples in [5, 6]. In practice, it is desirable to predict each individual functional curve with as few experiments as possible, especially in a new process. Consequently, the research problem is how to leverage the within- and between-process correlations so that information from data rich processes can help predict each individual function in the data-scarce process.

In the literature, Stochastic Kriging (SK) was proposed to model the functional relationship with heterogeneous replications and non-i.i.d. features [7]. The basic idea is to represent the heterogeneous replications as a shared mean function with additive non-i.i.d. noise. In this case, the shared mean function can still be modeled by a conventional GPR, and the deviations from the mean can be modeled as a weighted correlation structure for characterizing non-i.i.d. features. As a result, the SK achieved great success and wide applications in modeling non-i.i.d. functional relationships with heterogeneous replications [8].

Nevertheless, the way of constructing SK by focusing on the sample mean poses challenges and limitations to its broader applications. For example, the SK suffers from the individualization issue, which means its prediction is used for the functional mean of a process instead of an individual replication. In many engineering practices, however, the function mean can barely provide an accurate characterization of all individual units because the data from an individual unit usually deviates from the population mean. This is evidenced by the battery-to-battery variation in Fig. 1 and many data profiles in other engineering applications, e.g., unit degradation [5] and chemical reactions [6] etc. Although it is feasible to use the estimated correlation structure to provide a conditional prediction for each replication, the prediction accuracy highly depends on the estimation quality of the mean and the correlation structure. Unfortunately, it is well documented that the estimation performance of SK will be significantly impaired when there is a limited number of replications [9], which makes SK unsuitable for individualized prediction in a data-scarce process.

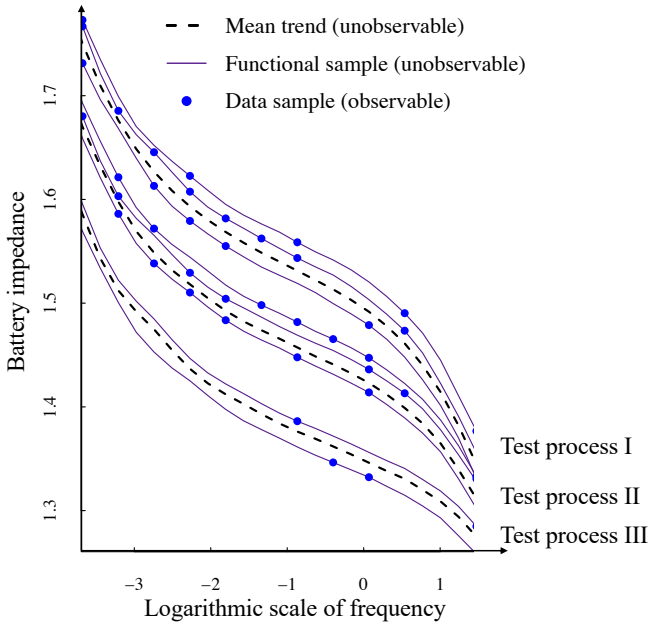


Fig. 1: The EIS test data in three processes.

An intuitive idea to remedy the individualization issue of SK in the data-scarce case is to resort to transfer learning, where the rich information from source processes can be leveraged to benefit the learning of the data-scarce target process. This is in fact attainable in many real applications, including the example shown in Fig. 1. There are indeed many existing works using transfer learning to deal with data-scarce problems in the context of GPR [10, 11], and we provide a review for these works in Section 2.1. Nevertheless, when dealing with processes with heterogeneous replications, these works often fail to comprehensively consider both individual-to-individual (within) correlations and process-to-process (between) correlations. For example, as illustrated in Fig. 1, current transfer learning GP techniques can facilitate information sharing among either multiple means (represented by dashed curves) across different processes or multiple individual replications within a single process, but they cannot model means and individuals in different processes simultaneously. In other words, these methods cannot fully leverage the within-and between-process correlations to benefit the prediction of individual replications. The root cause of this research gap stems from the lack of modeling components that differentiate between

individual-to-individual correlations and process-to-process correlations. In Fig. 1, this gap is illustrated: if the process-to-process correlation is modeled (by aggregating individuals to the mean curve in each process), then the model lacks the capability to further capture the correlation among individual replications within any given process, and vice versa. As a result, a straightforward extension of existing GPR transfer learning techniques to the SK cannot resolve the individualization and data-scarce issues.

In this paper, we propose a novel transfer learning framework for explicitly modeling the within-and between-process correlations in SK. This framework provides a comprehensive and scalable solution to both individualization and data scarce issues in the context of SK. Specifically, to facilitate individualization in the target process, we split heterogeneous replications into two groups: the replication of the individual unit of interest and the remaining replications. One of the key innovations of our work is that we use the convolution process [1] to build a tailored covariance structure between the individual unit of interest and the mean of the remaining replications. Based on their covariance, the mean of the remaining replications can serve as prior knowledge about the functional relationship of the individual unit of interest, and the available data in the individual unit of interest provides an update about this prior. For the between-process correlation, we again use the convolution process to build covariance connections. Specifically, we build links between each source and the two groups in the target so that the covariance matrix for the target process is embedded in the covariance matrix characterizing the process-to-process (between) correlation. We also add penalization terms when constructing the between-process covariance matrix. In this case, the individual unit of interest in the target receives two streams of information: the penalized information from each source and the unpenalized information from the remaining of the heterogeneous replications in the target.

Theoretical analysis shows that i) the constructed covariance matrix of all sources and the target is positive definite, and ii) it is guaranteed that we can select useful sources and remove negative transfer sources asymptotically. These theoretical properties thus provide evidence for resolving the issues of individualization and data scarce in the context of SK. As a result, the major contributions of the work include:

- A novel transfer learning framework is proposed to model the within-and between-process interactions in SK, where the accurate prediction of any individual unit in SK can be realized with limited data in the target process.
- Theoretical analysis provides guarantees that the proposed framework can resolve the issues of individualization and negative transfer asymptotically.
- The proposed framework is compared with benchmark methods in various numerical and case studies, and the results demonstrate the superiority of our method.

The remainder of the article is organized as follows. The formulation of the problem will be presented in Section 2, where related works in transfer learning using GP and the challenges in exiting SK will also be provided. In Section 3. the proposed transfer learning framework will be introduced, together with implementation details and theoretical analysis. The numerical and case studies will be in Section 4 and Section 5, respectively. Finally, we draw conclusion remarks in Section 6.

2 Preliminaries and Problem formulation

2.1 Related works in transfer learning of GP

The key to the GP-based transfer learning is to identify and capture the within-and between-process correlations. One effective way to model such correlations is to construct a positive semi-definite covariance matrix through a transfer kernel, which was studied in [12] for one target and one source domain.

When applying the transfer kernels to multiple source scenarios, there are two main categories of models: separable and non-separable models. The overall idea of the separable models is to use Kronecker products to form the covariance matrix among multiple GPs, where the positive semi-definite property of the joint matrix is guaranteed by the operation of the Kronecker product. Classical separable models include the intrinsic coregionalization model (ICM) [13] and the linear model of coregionalization (LMC) [14]. Extensions on ICM and LMC have also been made to facilitate scalability and handle non-Gaussian likelihoods and variance inflation [15–17]. Non-separable models, on the other hand, use a convolution process to generate GPs, allowing each GP to have its own set of hyperparameters that dominates the effectiveness of transfer learning [18]. Due to their modeling flexibility, the non-separable models attract prevalent attentions in the area of transfer learning [19]. For example, works in [19, 20] construct the multi-output Gaussian convolution process (MGCP) to facilitate transfer learning among multiple GPs. Some extensions are also made for dealing with incomplete samples [21] and computational issues [22].

Nevertheless, existing GP-based transfer learning approaches can model either multiple individual replications (within-process correlation) or multiple mean trends among processes (cross-process correlation), but they cannot model both within a single framework. In other words, these methods cannot fully leverage the process-to-process correlation to benefit individual replications. For example, the works in [10, 20, 23] all use non-separable models with different convolution structures, but they only focus on dealing with the within-process correlation among multiple individual replications. Works in [21, 22] also apply the non-separable models, but they can only model between-process correlation by aggregating replications in each process as a mean function.

Technically speaking, this research gap arises from the lack of modeling components that differentiate between individual-to-individual correlations and process-to-process correlations. As a result, it is imperative to develop a comprehensive framework for modeling and differentiating both correlations in the context of GP transfer learning.

2.2 Problem formulation for SK

Without loss of generality, we assume there are I processes and treat the Ith process as the target process. For the ith process, $i = 1, \cdots, I$, there will be m_i heterogeneous replications. We denote the mth replication in the ith process as $\boldsymbol{y}_i^{(m)}(\boldsymbol{x}_i^{(m)})$, where the input locations $\boldsymbol{x}_i^{(m)}$ are allowed to be different across different replications. We denote the union of all replications' inputs in the ith process as $\boldsymbol{x}_i = [x_{i,1}, x_{i,2}, \cdots, x_{i,n_i}]^T$, where n_i is the total number of *different* input locations across all replications in the ith process. As a result, we can express the mth replication in the ith process as $\boldsymbol{y}_i^{(m)}(\boldsymbol{x}_i^{(m)}) = \{y_i^{(m)}(x_{i,j}) | x_{i,j} \in \boldsymbol{x}_i^{(m)} \text{ and } \boldsymbol{x}_i^{(m)} \subseteq \boldsymbol{x}_i\}$. We further require the input elements in $\boldsymbol{x}_i^{(m)}$ are ordered in ascending order, i.e., $x_{i,j} < x_{i,j'}$ if $j < j'$ for $\forall x_{i,j} \in \boldsymbol{x}_i^{(m)}$, to facilitate the description of vectored $\boldsymbol{y}_i^{(m)}(\boldsymbol{x}_i^{(m)})$. Furthermore, for any $x_{i,j} \in \boldsymbol{x}_i$, there will be at least one replication having observation at this location. We denote the set of replication indices at $x_{i,j} \in \boldsymbol{x}_i$ as $R_{i,j}$, where $R_{i,j} \subseteq \{1, \cdots, m_i\}$ and $|R_{i,j}| \geq 1$. Figure 2 provides an example for our notation system in the ith process with $m_i = 3$ and $n_i = 4$, where each replication $\boldsymbol{y}_i^{(m)}(\boldsymbol{x}_i^{(m)})$ consists of data points indexed by different subset input locations of $\boldsymbol{x}_i$, and the elements in sets $R_{i,j}$ are the indices of replications that have data point(s) at $x_{i,j}$.

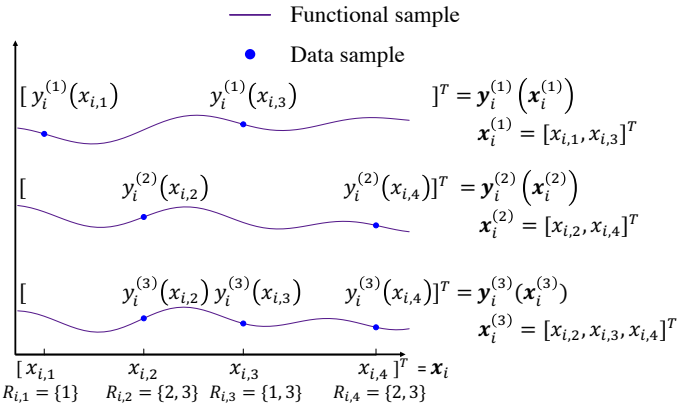


Fig. 2: Notation example for replications and input locations

To model the heterogeneous replications in Fig. 2, the SK formulates the data at input location $x_{i,j}$ from the mth replication in the ith process as:

$$y_i^{(m)}(x_{i,j}) = f_i(x_{i,j}) + \epsilon_i^{(m)}(x_{i,j}), x_{i,j} \in \boldsymbol{x}_i^{(m)}. \quad (1)$$

where the $f_i(\cdot)$ is the mean function shared across different replications in the ith process, and the $\epsilon_i^{(m)}(x_{i,j})$ is the noise term (usually non-i.i.d. across $x_{i,j}$) independent with $f_i(\cdot)$. The data generation procedure starts with $\boldsymbol{f}_i(\boldsymbol{x}_i) = [f_i(x_{i,1}), f_i(x_{i,2}), \cdots, f_i(x_{i,n_i})]^T$, which is a functional sample from $N(\boldsymbol{0}, \boldsymbol{\mathcal{K}}_{i,i})$. Note the $\boldsymbol{f}_i(\boldsymbol{x}_i)$ is sampled at all possible input locations in the ith process, i.e., $\boldsymbol{x}_i$. Then, the sampled $\boldsymbol{f}_i(\boldsymbol{x}_i)$ serves as the mean for all heterogeneous replications, and m_i different noise functional samples are added. All m_i different noise functional samples $\boldsymbol{\epsilon}_i^{(m)}(\boldsymbol{x}_i)$ $N(\boldsymbol{0}, \boldsymbol{\Sigma}_{i,i}), m = 1, \cdots, m_i$, are added to each individual replication. Finally, the observed data $y_i^{(m)}(x_{i,j})$ are those with $x_{i,j}$ available in the mth replication's input location set, i.e., $\boldsymbol{x}_i^{(m)}$, which constructs m_i heterogeneous replications in the ith process. A detailed description of the data generation procedure is available in Section A of supplementary materials.

To quantify the mean trend and non-i.i.d. noise (parameterized by $\boldsymbol{\mathcal{K}}_{i,i}$ and $\boldsymbol{\Sigma}_{i,i}$, respectively), SK uses m_i heterogeneous replications to compute the sample mean $\bar{\boldsymbol{y}}_i = [\bar{y}_i(x_{i,1}), \bar{y}_i(x_{i,2}), \cdots, \bar{y}_i(x_{i,n_i})]^T$ at each input $x_{i,j} \in \boldsymbol{x}_i$, where $\bar{y}_i(x_{i,j})$ is computed by

$$\bar{y}_i(x_{i,j}) = f_i(x_{i,j}) + \frac{1}{|R_{i,j}|} \sum_{m \in R_{i,j}} \epsilon_i^{(m)}(x_{i,j}) \quad (2)$$

which results in a new co-variance representation for $\bar{\boldsymbol{y}}_i$:

$$Cov(\bar{\boldsymbol{y}}_i, \bar{\boldsymbol{y}}_i) = \boldsymbol{\mathcal{K}}_{i,i} + \bar{\boldsymbol{\Sigma}}_{i,i} \quad (3)$$

where the (j, j')th element of $\bar{\boldsymbol{\Sigma}}_{i,i}$ is:

$$\bar{\Sigma}_{i,i}(x_{i,j}, x_{i,j'}) = Cov\Big(\frac{1}{|R_{i,j}|} \sum_{m \in R_{i,j}} \epsilon_i^{(m)}(x_{i,j}), \frac{1}{|R_{i,j'}|} \sum_{m' \in R_{i,j'}} \epsilon_i^{(m')}(x_{i,j'})\Big) \quad (4)$$

The hyperparameters in $\boldsymbol{\mathcal{K}}_{i,i}$ and $\bar{\boldsymbol{\Sigma}}_{i,i}$ are usually estimated by data $\bar{\boldsymbol{y}}_i$ with maximum likelihood estimation.

2.3 Challenges in transfer learning of SK

Although SK has achieved successful application in many engineering fields due to its strategy in dealing with heterogeneous replications, it suffers from intrinsic challenges for predicting individual replications. This can be evidenced by its prediction strategy: When having a set of new input locations $\boldsymbol{x}_i^*$ on the ith process and $|\boldsymbol{x}_i^*| = n_i^*$, SK can only predict the mean trend and uncertainty on these new inputs:

$$\boldsymbol{f}_i(\boldsymbol{x}_i^*) | \bar{\boldsymbol{y}}_i \sim N(\dot{\boldsymbol{\mathcal{K}}}_{i,i}^T [\boldsymbol{\mathcal{K}}_{i,i} + \bar{\boldsymbol{\Sigma}}_{i,i}]^{-1} \bar{\boldsymbol{y}}_i, \boldsymbol{\mathcal{K}}_{i,i}^* - \dot{\boldsymbol{\mathcal{K}}}_{i,i}^T [\boldsymbol{\mathcal{K}}_{i,i} + \bar{\boldsymbol{\Sigma}}_{i,i}]^{-1} \dot{\boldsymbol{\mathcal{K}}}_{i,i}) \quad (5)$$

where $\dot{\boldsymbol{\mathcal{K}}}_{i,i}$ is the covariance matrix between $\boldsymbol{f}_i(\boldsymbol{x}_i)$ and $\boldsymbol{f}_i(\boldsymbol{x}_i^*)$, and $\boldsymbol{\mathcal{K}}_{i,i}^*$ is the covariance matrix within $\boldsymbol{f}_i(\boldsymbol{x}_i^*)$.

The prediction in Eq. 5 shows a clear limitation of SK that it can only predict the mean tread $f_i(\cdot)$ and fails to predict the individual replication $y_i^{(m)}(\cdot)$, i.e., the individualization issue. The root cause of the individualization issue is the model deficiency in representing individual replications in the existing SK. This can be confirmed in Eq. 1 that all m_i replications are characterized by the same $\boldsymbol{\Sigma}_{i,i}$. Thus, the estimation of the deviation term can only be made for the parameters in $\boldsymbol{\Sigma}_{i,i}$ instead of the exact deviation values in each replication. In other words, there lacks a set of parameters or a representation tailored for modeling values in the individual deviation/replication.

To solve the individualization issue, an intuitive idea is to use the current SK framework to estimate the mean and $\boldsymbol{\Sigma}_{i,i}$, then the estimated mean can be subtracted from the data to obtain the deviation in each replication, and $\boldsymbol{\Sigma}_{i,i}$ can be used to provide a conditional prediction for the deviation at any input locations. However, this idea highly depends on the accurate modeling of the estimated mean, which is especially challenging when having a small amount

of data. An alternative solution is to resort to an effective transfer learning framework. However, this is a non-trivial task since it requires i) a holistic consideration of the within- and between-process correlation with heterogeneous replications in each process; and ii) an efficient transfer learning structure for alleviating negative transfer.

3 Model Development

To address these challenges and facilitate our proposed transfer learning framework, we list some assumptions as follows:

- A1 The mean trend of heterogeneous replications in each process is smooth and is from a stationary Gaussian process.
- A2 The noise in each heterogeneous replication is non-i.i.d. and stationary.
- A3 There exist similarities between the target process and source processes, and the similarity can be modeled by kernel functions and their parameters.
- A4 We only consider the information transfer from sources to the target, so the interactions among sources will be ignored to reduce computational complexity.

We want to mention the first two assumptions are commonly used in regular SK [7]. The last two assumptions are made specifically for transfer learning, and it is also validated in various literature studying multi-task and/or transfer learning of GP [21, 23].

3.1 Transfer Learning for Individualized SK

Our proposed transfer learning framework provides a general and explainable solution to the individualization issue by devising a holistic within-and between-process correlation structure. More specifically, the structure of within-process correlation is inspired by providing parameters/representation for the individual replication (as discussed in Section 2.3). This is mathematically equivalent to splitting and modeling all replications in a process as two groups: the individual replication of interest and the remaining replications.

Here we slightly abuse the notation of processes to denote the set of source processes as $S = \{1, \cdots, I-1\}$ and split the target process I into two new processes $I = \{I_r, I_t\}$, where I_t is for the individual replication of interest and I_r is for the remaining of the replications. These two new processes have the following formulations:

$$\begin{aligned} y^{(1)}_{I_t}(x_{I_t,j}) &= f_{I_t}(x_{I_t,j}) + \epsilon^{(1)}_{I_t}(x_{I_t,j}), x_{I_t,j} \in \boldsymbol{x}^{(1)}_{I_t}, \boldsymbol{x}^{(1)}_{I_t} = \boldsymbol{x}_{I_t} \\ y^{(m)}_{I_r}(x_{I_r,j}) &= f_{I_r}(x_{I_r,j}) + \epsilon^{(m)}_{I_r}(x_{I_r,j}), x_{I_r,j} \in \boldsymbol{x}^{(m)}_{I_r} \end{aligned} \tag{6}$$

where all notations are consistent with the notation system defined in Section 2, i.e., the subscript denotes the index of process and the superscript denotes the index of replication. Specifically, the $\boldsymbol{y}^{(1)}_{I_t}(\boldsymbol{x}^{(1)}_{I_t}) = \{y^{(1)}_{I_t}(x_{I_t,j}) | x_{I_t,j} \in \boldsymbol{x}^{(1)}_{I_t}\}$ is the individual replication of interest, which can be any replication in the target process. The $\boldsymbol{y}^{(m)}_{I_r}(\boldsymbol{x}^{(m)}_{I_r}) = \{y^{(m)}_{I_r}(x_{I_r,j}) | x_{I_r,j} \in \boldsymbol{x}^{(m)}_{I_r}\}$ is the mth replication in the remaining group. As a result, there are $m_{I_r} = (m_I - 1)$ and $m_{I_t} = 1$ replications in the I_rth process and the I_t process, respectively, and we have $\left(\bigcup_{m=1}^{m_{I_r}} \boldsymbol{x}^{(m)}_{I_r}\right) \bigcup \boldsymbol{x}^{(1)}_{I_t} = \boldsymbol{x}_I$.

The importance of the split of replications in the target process in Eq. 6 is that it provides additional parameterization space for modeling the individual replication and within-process correlation. For example, the $f_{I_r}(\cdot)$ and $f_{I_t}(\cdot)$ can be parameterized by different co-variance matrices, i.e., $\boldsymbol{\mathcal{K}}_{I_r,I_r}$ and $\boldsymbol{\mathcal{K}}_{I_t,I_t}$, where the $\boldsymbol{\mathcal{K}}_{I_t,I_t}$ is specifically for the estimation and prediction of the individual replication. The $\epsilon^{(m)}_{I_r}(\cdot)$ and $\epsilon^{(1)}_{I_t}(\cdot)$ can also be parameterized by $\boldsymbol{\Sigma}_{I_r,I_r}$ and $\boldsymbol{\Sigma}_{I_t,I_t}$, respectively. The $f_{I_r}(\cdot)$ and $f_{I_t}(\cdot)$ should also be dependent since they are both in the process I. Meanwhile, the $f_{I_r}(\cdot)$ and $f_{I_t}(\cdot)$ should be correlated with $f_i(\cdot), i \in S$, to facilitate the transfer learning.

To model the within-and between-correlation among $f_{I_r}(\cdot)$, $f_{I_t}(\cdot)$, and $f_i(\cdot), i \in S$, we resort to the convolution process [24], which provides a flexible framework for constructing Gaussian processes. More specifically, convolution process can represent any $f_i(\cdot)$ as a convolution between Gaussian white noise processes $Z_e(\cdot), e = 1, \cdots, l$ and smoothing kernels $g_{i,e}(\cdot)$ as follows [1]:

$$\begin{aligned} f_i(x_{i,j}) &= \sum_{e=1}^{l} g_{e,i}(x_{i,j}) * Z_e(x_{i,j}) \\ &= \sum_{e=1}^{l} \int_{-\infty}^{\infty} g_{e,i}(x_{i,j} - u) Z_e(u) du. \end{aligned} \tag{7}$$

where $*$ denotes the convolution operator, the $g_{e,i}(\cdot)$ is the kernel convolved with the eth Gaussian white noise process $Z_e(\cdot)$ for the $i$the process, and the Gaussian white noise process $Z_e(\cdot), e = 1, \cdots, l$, are assumed to be mutually independent. Although the Eq. 7 only works for a single $f_i(\cdot)$, i.e., within-process correlation, in the literature, we notice it is powerful and flexible in constructing different $f_i(\cdot)$ by assigning different combinations of convolution operations. We leverage this unique feature and specifically design the allocations of $g_{e,i}(\cdot)$ and $Z_e(\cdot)$ to extend it to model the within-and between-process correlation among $f_{I_r}(\cdot)$, $f_{I_t}(\cdot)$, and $f_i(\cdot)$. Our proposed transfer learning framework is presented in Fig. 3, which features some unique designs tailored for solving individualization issues:

- *Within-process correlation for individualization*. We use solid arrows to facilitate the convolution operation of within-process correlation. It is clear that the $f_{I_t}(\cdot)$ has its unique kernel g_{I_t,I_t} to characterize its within-process correlation. This corresponds to the requirement of "additional parameterization space" for modeling the individual replication. Meanwhile, the kernel g_{I_r,I_t} represents the intrinsic within-process dependence between $f_{I_t}(\cdot)$ and $f_{I_r}(\cdot)$. The g_{I_r,I_t} thus differentiates the relationship between $f_{I_t}(\cdot)$ and $f_{I_r}(\cdot)$ from the relationship between $f_{I_t}(\cdot)$ and $f_i(\cdot), i \in S$.
- *Between-process correlation for transfer learning*. We use dashed arrows to represent the convolution operation of between-process correlation. More specifically, the $f_{I_t}(\cdot)$ and $f_{I_r}(\cdot)$ in the target process both receive between-process correlation from all source processes. It is worth noting that the between-process

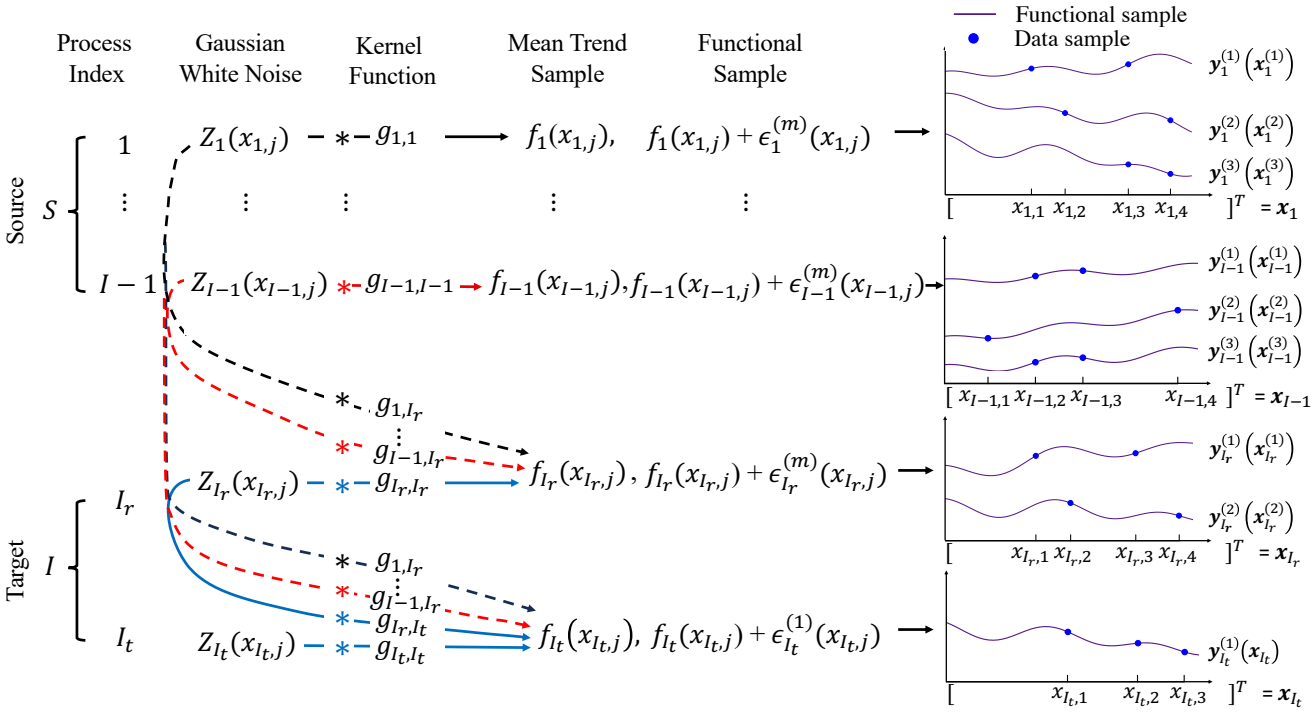


Fig. 3: The transfer learning structure of the proposed method.

correlation from the ith source process shares the same kernel function g_{i,I_t} for $f_{I_t}(\cdot)$ and $f_{I_r}(\cdot)$, or equivalently $g_{i,I_r} = g_{i,I_t}$ for $i \in S$. This is indeed expected since the $f_{I_t}(\cdot)$ and $f_{I_r}(\cdot)$ are both in the target process. The knowledge transferred through the between-process correlation from source processes will be critical and useful when the replications in the target process is limited, i.e., $m_I \ll m_i, i \in S$.

- *Flexible and explainable representation of data in heterogeneous replications.* The right-hand side column in Fig. 3 demonstrates the data in heterogeneous replications in each process. It is clear that given the constructed $f_i(\cdot)$, the data generation procedures for each process are the same as those in Fig. 2 for single SK. This indicates that the strategy of taking sample means from heterogeneous replications can also be used for parameter estimation in the proposed transfer learning framework. As a result, the proposed transfer learning framework does not pose any restrictions on data generation compared to the single process SK, which successfully inherits the flexible modeling capability of SK while solving the individualization issue in the context of transfer learning.

The mentioned features in Fig. 3 can also be formulated as mathematical representations to facilitate parameter estimation and prediction of the proposed transfer learning framework. Specifically, the $f_i(x_{i,j}), i \in S \cup I$, can be represented as follows:

$$f_i(x_{i,j}) = \begin{cases} g_{i,i}(x_{i,j}) * Z_i(x_{i,j}) & \text{if } i \in S \\ \sum_{e \in S \cup I_r} g_{e,I_r}(x_{i,j}) * Z_e(x_{i,j}) & \text{if } i = I_r \\ \sum_{e \in S \cup I} g_{e,I_t}(x_{i,j}) * Z_e(x_{i,j}) & \text{if } i = I_t. \end{cases} \quad (8)$$

To formulate the within-and between-process correlation, we first apply Eq. 2 to each process to obtain the sample mean $\bar{\boldsymbol{y}}_i$. As a result, we have the inputs and sample means from all processes as $\boldsymbol{x} = [\boldsymbol{x}_1{}^T, \boldsymbol{x}_2{}^T, \cdots, \boldsymbol{x}_{I_r}^T, \boldsymbol{x}_{I_t}^T]^T$ and $\bar{\boldsymbol{y}} = [\bar{\boldsymbol{y}}_1^T, \bar{\boldsymbol{y}}_2^T, \cdots, \bar{\boldsymbol{y}}_{I_r}^T, \bar{\boldsymbol{y}}_{I_t}^T]^T$. Note the $\bar{\boldsymbol{y}}_{I_t}$ is equivalent to $\boldsymbol{y}_{I_t}^{(1)}(\boldsymbol{x}_{I_t})$ since the replication number in the I_tth process is always 1, i.e., the individual replication of interest in the target process. We provide Lemma 1 to formulate the covariance matrix among sample means of all processes.

Lemma 1 *Suppose the assumptions A1 to A4 are satisfied, and the kernels in Fig. 3 are square-integrable, then* $\bar{\boldsymbol{y}} \sim N(\boldsymbol{0}, \boldsymbol{\Omega})$, *where* $\boldsymbol{\Omega}$ *is a positive definite matrix represented as follows:*

$$\boldsymbol{\Omega} = \boldsymbol{\mathcal{K}} + \bar{\boldsymbol{\Sigma}} = \left[\begin{array}{c|c} \boldsymbol{\Omega}_{S,S} & \boldsymbol{\Omega}_{S,I} \\ \hline \boldsymbol{\Omega}_{S,I}^T & \boldsymbol{\Omega}_{I,I} \end{array}\right]$$

$$\boldsymbol{\Omega}_{S,S} = \begin{bmatrix} \boldsymbol{\mathcal{K}}_{1,1} + \bar{\boldsymbol{\Sigma}}_{1,1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\mathcal{K}}_{2,2} + \bar{\boldsymbol{\Sigma}}_{2,2} & \cdots & \boldsymbol{0} \\ \vdots & \cdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{\mathcal{K}}_{I-1,I-1} + \bar{\boldsymbol{\Sigma}}_{I-1,I-1} \end{bmatrix}$$

$$\boldsymbol{\Omega}_{S,I} = \begin{bmatrix} \boldsymbol{\mathcal{K}}_{1,I_r} & \boldsymbol{\mathcal{K}}_{1,I_t} \\ \boldsymbol{\mathcal{K}}_{2,I_r} & \boldsymbol{\mathcal{K}}_{2,I_t} \\ \vdots & \vdots \\ \boldsymbol{\mathcal{K}}_{I-1,I_r} & \boldsymbol{\mathcal{K}}_{I-1,I_t} \end{bmatrix}, \boldsymbol{\Omega}_{I,I} = \begin{bmatrix} \boldsymbol{\mathcal{K}}_{I_r,I_r} + \bar{\boldsymbol{\Sigma}}_{I_r,I_r} & \boldsymbol{\mathcal{K}}_{I_r,I_t} \\ \boldsymbol{\mathcal{K}}_{I_r,I_t}^T & \boldsymbol{\mathcal{K}}_{I_t,I_t} + \bar{\boldsymbol{\Sigma}}_{I_t,I_t} \end{bmatrix} \quad (9)$$

where the $\boldsymbol{\mathcal{K}}$ *is constructed by the* $n_i \times n_{i'}$ *block matrix* $\boldsymbol{\mathcal{K}}_{i,i'}$, *and* $\boldsymbol{\mathcal{K}}_{i,i'} = Cov(\boldsymbol{f}_i(\boldsymbol{x}_i), \boldsymbol{f}_{i'}(\boldsymbol{x}_{i'}))$. *The* $\bar{\boldsymbol{\Sigma}}$ *is a diagonal block matrix with* $\bar{\boldsymbol{\Sigma}}_{i,i}$ *on the diagonals, and the* $\bar{\boldsymbol{\Sigma}}_{i,i}, \forall i \in S \cup I$, *is the* $n_i \times n_i$ *matrix defined in Eqs. 3 and 4.*

The $\boldsymbol{\Omega}$ can further be partitioned into four blocks as demonstrated in Eq. 9, where the $\boldsymbol{\Omega}_{S,S}$, $\boldsymbol{\Omega}_{S,I}$, and $\boldsymbol{\Omega}_{I,I}$ are for covariance within sources, between sources and target, and within target, respectively. The proof of Lemma 1 with the details of each element in $\boldsymbol{\mathcal{K}}$ and $\bar{\boldsymbol{\Sigma}}$ are in Section B of supplementary materials.

Remarks on Lemma 1: The 2-by-2 block representation of $\boldsymbol{\Omega}$ in Eq. 9 clearly interprets the within-and between-process correlation structure designed in Fig. 3. The $\boldsymbol{\Omega}_{S,S}$ represents the covariance among $I-1$ source processes, where only the within-process correlation in the main diagonal is non-zero. This is because all source processes in Fig. 3 are designed to only have solid arrows (within-process correlation). The $\boldsymbol{\Omega}_{S,I}$ represents the between-process correlation between each of the source processes and the target process, where the target process is further split into two processes. This block corresponds to the dashed arrows in Fig. 3. Finally, the $\boldsymbol{\Omega}_{I,I}$ is the within target process correlation, which consists of the within-process correlations of processes I_t and I_r (on the main diagonal) and the dependence between processes I_t and I_r (represented as $\boldsymbol{\mathcal{K}}_{I_r,I_t}$). These matrices refer to the solid arrows in the target process in Fig. 3.

The direct use of Lemma 1 results in the prediction of the individual replication at arbitrary inputs $\boldsymbol{x}^*_{I_t}$ in the process of interest:

$$\begin{bmatrix} \bar{\boldsymbol{y}} \\ \boldsymbol{y}^{(1)}_{I_t}(\boldsymbol{x}^*_{I_t}) \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\mathcal{K}}+\bar{\boldsymbol{\Sigma}} & \dot{\boldsymbol{\mathcal{K}}} \\ \dot{\boldsymbol{\mathcal{K}}}^T & \boldsymbol{\mathcal{K}}^*_{I_t,I_t}+\boldsymbol{\Sigma}^*_{I_t,I_t} \end{bmatrix} \right)$$
$$\boldsymbol{y}^{(1)}_{I_t}(\boldsymbol{x}^*_{I_t})|\bar{\boldsymbol{y}} \sim N(\dot{\boldsymbol{\mathcal{K}}}^T\boldsymbol{\Omega}^{-1}\bar{\boldsymbol{y}},\ \boldsymbol{\mathcal{K}}^*_{I_t,I_t}+\boldsymbol{\Sigma}^*_{I_t,I_t}-\dot{\boldsymbol{\mathcal{K}}}^T\boldsymbol{\Omega}^{-1}\dot{\boldsymbol{\mathcal{K}}}) \tag{10}$$

where $\dot{\boldsymbol{\mathcal{K}}} = [\dot{\boldsymbol{\mathcal{K}}}^T_{1,I_t}, \dot{\boldsymbol{\mathcal{K}}}^T_{2,I_t}, \cdots, \dot{\boldsymbol{\mathcal{K}}}^T_{I-1,I_t}, \dot{\boldsymbol{\mathcal{K}}}^T_{I_r,I_t}, \dot{\boldsymbol{\mathcal{K}}}^T_{I_t,I_t}]^T$, $\dot{\boldsymbol{\mathcal{K}}}_{i,I_t} = Cov(\boldsymbol{f}_i(\boldsymbol{x}_i), \boldsymbol{f}_{I_t}(\boldsymbol{x}^*_{I_t})), i \in S \cup I$, and $\boldsymbol{\mathcal{K}}^*_{I_t,I_t} = Cov(\boldsymbol{f}_{I_t}(\boldsymbol{x}^*_{I_t}), \boldsymbol{f}_{I_t}(\boldsymbol{x}^*_{I_t}))$. It is worth noting that $\dot{\boldsymbol{\mathcal{K}}}_{i,I_t}$ in Eq. 10 and the $\boldsymbol{\mathcal{K}}_{i,I_t}$ in Eq. 9 use the same kernel (parameters) to construct the covariance matrix, but they differ in terms of the input locations. The $\dot{\boldsymbol{\mathcal{K}}}_{i,I_t}$ uses the inputs $\boldsymbol{x}^*_{I_t}$ for prediction, while the $\boldsymbol{\mathcal{K}}_{i,I_t}$ uses the inputs $\boldsymbol{x}_{I_t}$ from training data. The same difference applies to $\boldsymbol{\mathcal{K}}^*_{I_t,I_t}$ vs. $\boldsymbol{\mathcal{K}}_{I_t,I_t}$ and $\boldsymbol{\Sigma}^*_{I_t,I_t}$ vs. $\boldsymbol{\Sigma}_{I_t,I_t}$.

3.2 Implementation Details and Properties for Dealing with Negative Transfer

The Eqs. 9 and 10 provide a novel and general transfer learning framework for dealing with the individualization issue in SK. As we mentioned in Section 2.3, however, the effectiveness of any transfer learning method depends on the successful identification of useful sources and the exclusion of inappropriate sources. This is especially critical in the era of "Big Data", where extensive data or processes are involved in transfer learning as potential sources. In this section, we will first parameterize the Eqs. 9 and 10 with the widely adopted kernels and non-i.i.d. noise matrix. Then, a parameter estimation method and its statistical properties are proposed to deal with the potential negative transfer and guarantee the performance of the proposed transfer learning framework.

Specifically, we use the Gaussian kernel for $g_{i,i'}(x)$ in Fig. 3:

$$g_{i,i'}(x) = \frac{\alpha_{i,i'}}{2\sqrt{\pi}\sqrt{|\beta_{i,i'}|}} \exp(\frac{-x^2}{2\beta^2_{i,i'}}) \tag{11}$$

where $\alpha_{i,i'}$ is the scaling parameter and $\beta_{i,i'}$ is the length-scale parameter.

For the non-i.i.d noise in $\boldsymbol{\Sigma}_{i,i}$, we parameterize it using the equicorrelation matrix (a special scenario of Toeplitz matrix) [25], which is widely used in characterizing auto-correlated noise [26]:

$$\boldsymbol{\Sigma}_{i,i} = \sigma^2_i \begin{bmatrix} 1 & \rho_i & \cdots & \rho_i \\ \rho_i & 1 & \cdots & \rho_i \\ \vdots & \vdots & \ddots & \vdots \\ \rho_i & \rho_i & \cdots & 1 \end{bmatrix} \tag{12}$$

where σ^2_i and ρ_i are the intensity of noise and auto-correlation for the non-i.i.d. noise in the ith process. To guarantee the positive definiteness of $\bar{\boldsymbol{\Sigma}}$, it requires $-\frac{1}{n_i} < \rho_i < 1$. The detailed structure of $\boldsymbol{\mathcal{K}}_{i,i'}$ and $\bar{\boldsymbol{\Sigma}}$ by using Eq. 11 and Eq. 12 is provided in Section B of supplementary materials. As a result, the $\boldsymbol{\Omega}$ can be fully parameterized by the parameter set $\boldsymbol{\theta} = \{\alpha_{i,i}, \alpha_{i,I_r}, \beta_{i,i}, \beta_{i,I_r} | i \in S\} \cup \{\alpha_{I_r,I_r}, \alpha_{I_t,I_t}, \alpha_{I_r,I_t}, \beta_{I_r,I_r}, \beta_{I_t,I_t}, \beta_{I_r,I_t},\} \cup \{\sigma_i, \rho_i | i \in S \cup I\}$.

Before we introduce the parameter estimation method, we want to point out an interesting observation that the $\boldsymbol{\mathcal{K}}_{i,I_r}$ in Eq. 9 will be a zero block matrix if $\alpha_{i,I_r} = 0$ for $i \in S$. This means the parameters α_{i,I_r} dominate the effectiveness of transfer learning from source processes. In fact, the parameters α_{i,I_r} determine the inclusion or exclusion of observed data in source processes, and we extend this observation into a general case in Section C of supplementary materials.

The importance of this observation is that it reveals a feasible way to select source processes and data, i.e., the α_{i,I_r} should be close to 0 for non-informative source processes. To implement this idea, we add a penalty to the parameter set $\boldsymbol{\theta_0} = \{\alpha_{1,I_r}, \cdots, \alpha_{I-1,I_r}\}$ during the parameter estimation procedure:

$$\max_{\boldsymbol{\theta}} L_{\mathbb{P}}(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - \mathbb{P}_\gamma(\boldsymbol{\theta_0})$$
$$L(\boldsymbol{\theta}) = -\frac{1}{2}\bar{\boldsymbol{y}}^T[\boldsymbol{\mathcal{K}}+\bar{\boldsymbol{\Sigma}}]^{-1}\bar{\boldsymbol{y}} - \frac{1}{2}\log|\boldsymbol{\mathcal{K}}+\bar{\boldsymbol{\Sigma}}| - \frac{N}{2} * log(2\pi)$$
$$\mathbb{P}_\gamma(\boldsymbol{\theta_0}) = \begin{cases} \gamma \sum_{i\in S} \left(\frac{1}{2\eta}\alpha^2_{i,I_r}\right) & \text{if } |\alpha_{i,I}| \le \eta \\ \gamma \sum_{i\in S} \left(|\alpha_{i,I_r}| - \frac{\eta}{2}\right) & \text{if } |\alpha_{i,I}| > \eta \end{cases} \tag{13}$$

where $L(\boldsymbol{\theta})$ is the log-likelihood function for $\bar{\boldsymbol{y}}$, $\mathbb{P}_\gamma(\boldsymbol{\theta_0})$ is the penalization term, the $L_{\mathbb{P}}(\boldsymbol{\theta})$ is the penalized log-likelihood function, and $N = \sum_i n_i$. We apply the Huber smooth approximation [27] of a L_1 norm to construct the $\mathbb{P}_\gamma(\boldsymbol{\theta_0})$, where the γ is a tuning parameter that can be determined by cross-validation and η is a small constant.

The formulation of the parameter estimation in Eq. 13 provides a general platform for penalizing those α_{i,I_r} that do not contribute to maximizing the $L(\boldsymbol{\theta})$. In this way, the corresponding data and processes are also penalized and given less attention in the parameter estimation procedure. Moreover, the formulation in Eq. 13 enjoys some theoretical properties that provide conditions and guarantees for excluding the non-informative sources.

Theorem 1 *(Parameter estimation consistency) Given that the MLE $\hat{\boldsymbol{\theta}}_u$ for the unpenalized likelihood function $L(\boldsymbol{\theta})$ converges to the true parameters $\boldsymbol{\theta}^*$ with rate r_N, if $\max\{|\mathbb{P}''_\gamma(\alpha^*_{i,I_r})| : \alpha^*_{i,I_r} \neq 0\} \to 0$, under some regularity conditions, there exists $\hat{\boldsymbol{\theta}}$ that attains the local maximum of $L_{\mathbb{P}}(\boldsymbol{\theta})$ such that $||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*|| = O_p(r_N^{-1} + r_0)$, where $r_0 = \max\{|\mathbb{P}'_\gamma(\alpha^*_{i,I_r})| : \alpha^*_{i,I_r} \neq 0\}$, α^*_{i,I_r} are the true parameters in $\boldsymbol{\theta}^*$, and $||\cdot||$ is the L_2 norm.*

The proof of Theorem 1 is detailed in Section D of supplementary materials. This theorem states the estimation results from Eq. 13 are consistent under regularity conditions, which provides theoretical justifications for the accuracy of parameter estimation and prediction of the proposed framework. More importantly, the negative transfer can be asymptotically resolved based on the result of Theorem 1:

Theorem 2 *(Resolve negative transfer) Let* $\boldsymbol{\theta}_{10}^* = \{\alpha_{i,I_r}^* | \alpha_{i,I_r}^* = 0, i \in S\}$ *be the parameter set of zeros in true parameters* $\boldsymbol{\theta}_0^*$ *for* $\boldsymbol{\theta}_0$*, and let* $\hat{\boldsymbol{\theta}}_{10}$ *be the corresponding local optimal based on* $L_{\mathbb{P}}(\boldsymbol{\theta})$*. If the Theorem 1 holds,* $\liminf_{N\to\infty}\liminf_{\theta\to 0^+}\gamma^{-1}\mathbb{P}_\gamma(\alpha_{i,I_r}^*) > 0$ *and* ${r_N}^{-1}\gamma^{-1} \to 0$ *with* $r_N \to \infty$*, then*

$$\lim_{N\to\infty} P(\hat{\boldsymbol{\theta}}_{10} = \mathbf{0}) = 1. \tag{14}$$

The proof of Theorem 2 is detailed in Section E of supplementary materials. This theorem shows the capability of resolving negative transfer in the proposed framework, where the non-informative sources (with $\alpha_{i,I_r} = 0$) will be eventually identified and excluded from the learning procedure. As a result, the proposed framework comprehensively resolves the issues of individualization and negative transfer, thus providing a flexible and robust solution to predicting non-i.i.d. functional data.

It is worth noting that the performance of alleviating negative transfer depends on many factors, including the designed covariance structure, the penalization form, and the type of kernels. Fortunately, given the proposed covariance structure (in Fig. 3) and the penalization form (in Eq. 13), the performance of alleviating negative transfer will not be significantly affected by most, if not all, kernels. This is because many widely used kernels, e.g., rational quadratic, Matérn, and periodic, can be represented in the form $\alpha \cdot f(\Delta x)$ [28, 29]. In our proposed framework, the target process receives distinct kernels, i.e., g_{i,I_r}, from each source, and the penalization part applies to the scaling parameters of these kernels, i.e., α_{i,I_r}. In this case, as long as the kernels used for g_{i,I_r} have the form $\alpha \cdot f(\Delta x)$, our theorems hold. As a result, our proposed framework can accommodate different types of kernels with guaranteed performance in alleviating negative transfer, and we use Gaussian kernels in the paper to illustrate the implementation details.

We also want to point out the scalability of the proposed framework in terms of the number of observations and the dimension of the input space. For the number of observations, thanks to the sparse transfer learning structure in Fig. 3, the computational complexity (dominated by the inverse of $\boldsymbol{\Omega}$) has improved to $O(\sum_i n_i^3)$, compared with $O\big((\sum_i n_i)^3\big)$ in commonly used multivariate Gaussian process [15]. Nevertheless, it still scales exponentially with the number of data. For the dimension of input, it also requires the number of data scales exponentially with the dimension [30, 31], which is a challenge widely known as the curse of dimensionality [30]. In practice, the variational inference [32] and the Gaussian Process Latent Variable Model (GPLVM) [33] are commonly used tools to further reduce the computational complexity and improve the scalability of GP. However, the detailed discussions and investigations on developing tailored tools for our proposed framework go beyond the focus of this work, as our contribution lies in proposing a novel transfer learning framework that incorporates both individual-to-individual and process-to-process correlations.

4 Numerical Study

In this section, we demonstrate the effectiveness of the proposed transfer learning framework by comparing it with various benchmark methods under different signal settings. More specifically, we introduce five different benchmarks, and each of them represents one type of simplification of the proposed method. We also implement the comparisons under three different signal settings, where the scarce data collected at random locations, at segment locations, and under large-scale processes are considered.

4.1 General Setting and Benchmarks

The proposed method aims to address the individualization and data-scarce issues associated with applying SK to predict non-i.i.d. functional data. To compare with the proposed method, we set up five different benchmarks, and each one represents a type of simplification of the proposed strategy. We elaborate on each of the benchmarks as follows:

1) Relax penalization. This benchmark method uses the same framework, i.e., Fig. 3, as our proposed method, but it does not apply the penalization term, i.e., $\mathbb{P}_\gamma(\boldsymbol{\theta_0})$, in Eq. 13. Thus, its likelihood function is $L(\boldsymbol{\theta})$ instead of $L_{\mathbb{P}}(\boldsymbol{\theta})$. This benchmark is proposed to test the effectiveness of mitigating negative transfer, and it is denoted as "Unpenalized".
2) Down-sampling to resolve heterogeneous replications. An alternative way for SK to resolve the heterogeneous replications in each process is to down-sample the replications in each process so that only one replication is retained. In this case, a regular MGP structure can be used to model the one replication in each process. To retain information for each process, we select the replication index that has the largest number of data points in each process, i.e., $\dot{m}_i = \underset{m\in\{1,\cdots,m_i\}}{\operatorname{argmax}} |\boldsymbol{x}_i^{(m)}|$. It is clear this benchmark method drops a significant amount of data, thus it might not be robust to the non-i.i.d. noise. This benchmark method is a variant of the method in [20], where we impose the non-i.i.d. structure. We denote this model as "Minimal Transfer".
3) Relax non-i.i.d. assumption. This benchmark method assumes i.i.d. noise, i.e., $\epsilon_i^{(m)}(x_{i,j}) \sim N(0, \sigma^2)$, for all processes. In this case, the prediction of the individualized replication (I_t) might fail because the deviation from the underlying mean trend becomes purely stochastic under i.i.d. noise. As a result, this benchmark will focus on predicting the mean trend, instead of the individual replication. This model is denoted as "i.i.d. MGP".
4) Relax non-i.i.d. assumption with enforced correlation. This is based on the third benchmark method, but the between-process correlation model in [10] is used, where a fully correlated covariance matrix can be constructed. A feature of the model in [10] is that it assumes all processes are always mutually correlated (enforced

correlation). We denote this benchmark as "i.i.d. Enforced".

5) No transfer case. To test the effectiveness of transfer learning, this benchmark uses only data in the target process. However, to facilitate the prediction of individual replication, we still split the replications in the target process to I_r and I_t processes. Thus, this is a modified single SK method, and we denote this benchmark method as "Individualized SK".

To evaluate the performance of the prediction result, the root mean square deviation (RMSE) is used to quantify the deviation between prediction and observed data in the I_t replication. The RMSE for each method will be calculated and compared, and each experiment will repeat 100 times to show the box plot of RMSE.

4.2 Setting I (randomly collected data)

In this section, we set the total number of processes as $I = 5$, i.e., 4 source processes:

$$\begin{aligned} y_1(x_1) &= 2\cos(5x_1) + \epsilon_1^{(m)}(x_1), \\ y_2(x_2) &= 2\sin(5x_2) + \epsilon_2^{(m)}(x_2), \\ y_3(x_3) &= \frac{1}{2}\exp(x_3) + \epsilon_3^{(m)}(x_3), \\ y_4(x_4) &= \frac{1}{5}(x_4+3)(x_4+4) + \epsilon_4^{(m)}(x_4), \end{aligned} \tag{15}$$

and 1 target process:

$$y_5(x_5) = \cos(5x_5) + \sin(5x_5) + \epsilon_5^{(m)}(x_5), \tag{16}$$

where the $\mathbf{\Sigma}_{i,i}$ for $\epsilon_i^{(m)}(\cdot)$ is with the equalcorrelation structure, and $\rho_i = 0.9$ and $\sigma_i = 1, \forall i \in S \cup I$.

The sample space x_i for each process is defined on a regularly spaced domain $\boldsymbol{\mathcal{X}}$, which is on $[0, \pi]$. To ensure the source process has more data points and replications than the target process, we set $m_i = 5$ for $i = 1, \cdots, 4$ and $m_5 = 3$. For each replication in the source, we first randomly select an integer number between 25 and 30 to determine the number of data points in the replication, i.e., $|\boldsymbol{x}_i^{(m)}|$. The space $\boldsymbol{\mathcal{X}}$ is randomly occupied by these $|\boldsymbol{x}_i^{(m)}|$ points. The target process generates data in a similar way, and the only difference is that we only allow 10 to 15 data in each replication. One of the replications in the target process will be randomly selected as the I_tth process, and the remaining two replications construct the two replications in the I_rth process.

We provide the visualization of sample means of 5 processes in Fig. 4 (a), i.e., $\bar{\boldsymbol{y}}_1, \cdots, \bar{\boldsymbol{y}}_4$, and $\bar{\boldsymbol{y}}_{I_r}$, where the 4 sources are with sample mean of 5 replications and the I_rth target process is with sample mean of 2 replications. The individual replication for prediction, i.e., I_tth process, is shown as a green solid line in Fig. 4 (b) with green dots as data samples. The prediction results using different methods are shown in Fig. 4 (c)-(h). There are some interesting observations about the prediction performance among different methods in Fig. 4:

- The proposed method achieves the best performance in predicting the individual replication in the target process. This is demonstrated as the accurate mean prediction (red solid curve) and the reasonable confidence interval (grey area).
- The Unpenalized method achieves a similar performance to the proposed method in terms of mean prediction, but it provides a much larger confidence interval, indicating the prediction quality is not as good as the proposed method. This is because sources 3 and 4 in Fig. 4 (a) have large deviations from the target, and the inclusion of these two sources

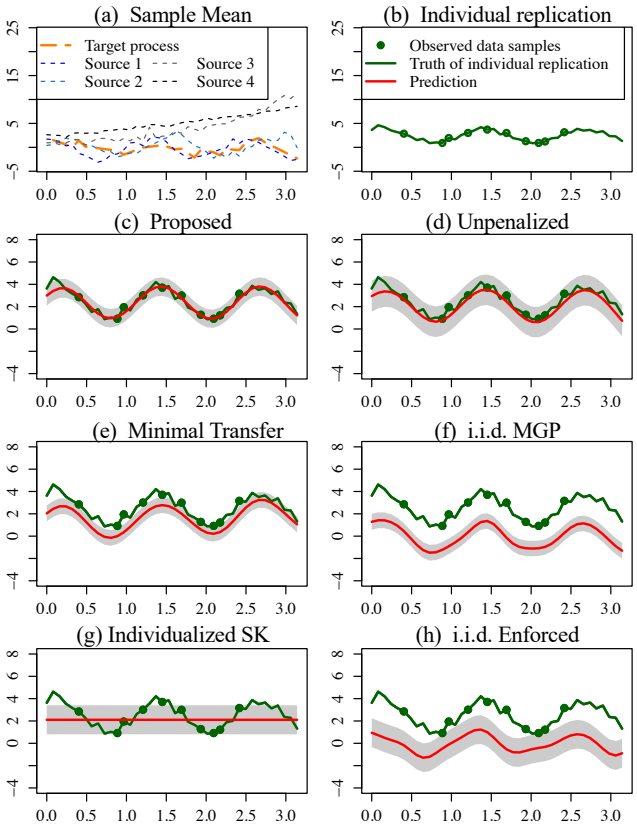


Fig. 4: Prediction results in one specific trial of Setting I

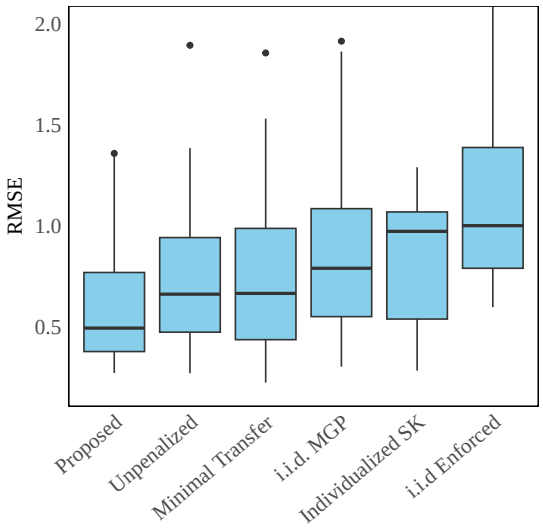


Fig. 5: Setting I RMSE with 100 trials

generates inferior prediction performance.

- The Minimal Transfer achieves inferior performance in mean and confidence interval, where the prediction of mean deviates from the observed data and most of the data are at the boundary or outside of the confidence interval. This is because the Minimal Transfer drops a significant amount of data during parameter estimation, and the remaining data cannot capture the correct trend of the individual replication. Such a strategy also indicates the prediction quality of Minimal Transfer highly depends on the remaining limited number of training data, which can cause robustness issues in the prediction results.
- The i.i.d. MGP provides a large deviation in the prediction result. This is expected since the i.i.d. MGP ignores the deviation resulting from the non-i.i.d. noise and only focuses on the prediction of the signal mean. Unfortunately, the individual replication in the I_tth process has a large auto-correlated deviation from the mean. As a result, the prediction performance of the i.i.d. MGP is unsatisfactory.
- The i.i.d. Enforced provides even worse performance compared to the i.i.d. MGP. This is because the negative sources affect prediction accuracy and increase uncertainty. In fact, the difference between the proposed method and the unpenalized method, as well as the difference between the i.i.d. MGP and the i.i.d. Enforced, share a similar pattern, where the impaired prediction accuracy and increased uncertainty are due to the neglect of negative transfer.
- The Individualized SK provides a straight line for mean prediction and a large confidence interval that covers most of the data. This is because the limited data in the target process cannot support the accurate identification of the mean trend, which results in a rough estimation of the mean with a large confidence interval.

It is worth noting that the results in Fig. 4 is just one specific trial of the signals in Setting I. The randomness in data generation will impact the training and prediction results. We repeat the procedures in Fig. 4 100 times and report the box plot in Fig. 5. It is clear that the proposed method still achieves the best performance among 100 trials. An interesting observation from Fig. 5 is that both Minimal Transfer and i.i.d. MGP/Enforced have larger variances than other methods. We want to point out that the root cause of uncertainties in these two methods is different. For the Minimal Transfer, the larger uncertainty is mainly from dropping data, i.e., the uncertainty is due to lacking sufficient data. On the other hand, the uncertainty of i.i.d. MGP/Enforced is rooted in the model structure, in which only the prediction of the mean is considered. This is visualized in Fig. 4 (a), where the RMSE of prediction highly depends on the observed data: If the observed data is generated close to the mean, then the RMSE can be small, otherwise the RMSE is large. We also provide the results of $\hat{\boldsymbol{\theta}}_0$ in Table 1, where the proposed method correctly penalizes sources 3 and 4 (with estimated values close to 0). The Unpenalized benchmark still heavily relies on the information from source 4. This validates the effectiveness of the penalization term in mitigating the negative transfer, which is a critical advantage in transfer learning. Note the i.i.d. Enforced does not allow penalization, and it is not provided in Table 1.

TABLE 1: Median of estimated scaling parameter, Setting I

	Proposed	Unpenalized	Minimal Transfer	i.i.d MGP
α_{1,I_r}	0.76	0.81	0.90	1.05
α_{2,I_r}	1.41	-1.19	0.77	0.73
α_{3,I_r}	0.01	-0.08	0.15	0.28
α_{4,I_r}	0.03	-2.40	0.02	0.07

4.3 Setting II (segmentally collected data)

In this section, we apply the methods to a different signal setup, where the data in the target process are only available at the beginning section of $\boldsymbol{\mathcal{X}}$. More specifically, we have the 4 source processes as:

$$\begin{aligned} y_1(x_1) &= 0.3(x_1-3)^3 + \epsilon_1^{(m)}(x_1), \\ y_2(x_2) &= 3(x_2)^3 + 2\sin(2x_2) + \epsilon_2^{(m)}(x_2), \\ y_3(x_3) &= (x_3-2)^2 + \epsilon_3^{(m)}(x_3), \\ y_4(x_4) &= (x_4-1)(x_4-2)(x_4-4) + \epsilon_4^{(m)}(x_4), \end{aligned} \tag{17}$$

and the target process as:

$$y_5(x_5) = 0.2(x_5-3)^2 + 0.15(x_5)^2 + \sin(2x_5) + \epsilon_5^{(m)}(x_5), \tag{18}$$

where the non-i.i.d. noise setting and the sampling procedures are the same as those in the Setting I. The only difference is that the 10 to 15 data points in the target process will only appear in the beginning 40% section of $\boldsymbol{\mathcal{X}}$. In this case, there will be no data in the remaining 60% section of $\boldsymbol{\mathcal{X}}$, and the prediction for the remaining section becomes an extrapolation task. Such a signal setting is commonly used in system condition prognosis [34, 35]. We visualize the generated sample mean in Fig. 6, where the target process mean in Fig. 6 (a) covers only a limited section in $\boldsymbol{\mathcal{X}}$. The extrapolation results are shown in Fig. 6 (c) to (h).

The result in Fig. 6 shows the proposed method again achieves the best result. However, there are some different observations comparing with the results in Fig. 4. First, the difference between the proposed and the Unpenalized is very minor, which is due to the much closer source and target data in Fig. 6. It is also worth noting that the Minimal Transfer even provides a wrong trend in extrapolation. This is because of the influence of source 4, and it indicates that Minimal Transfer is sensitive to the trend in sources due to the limited data included in the training stage. Finally, the i.i.d. MGP/Enforced provides a reasonable prediction. This is because the individual replication (observed data) in this specific trial is very close to the mean trend (as opposed to the case in Fig. 4). To provide a more comprehensive view of the performance in Setting II, we repeat the procedures 100 times, and the box plot is in Fig. 7. A notable observation is that the variance of the Minimal Transfer becomes larger compared with that in Fig. 5. This is expected since the Minimal Transfer suffers from uncertainties from both the data generation and extrapolation in Setting II, which also explains its high RMSE. We also provide the results of $\hat{\boldsymbol{\theta}}_0$

in Table 2, where the proposed method identifies source 3 as the negative source. This is a reasonable result because the source 3 is the most deviated signal at the beginning of $\boldsymbol{\mathcal{X}}$, which can also be visualized in Fig. 6 (a). It is also worth noting in Table 2 that although Minimal Transfer also identifies the correct negative source, its performance is the worst among all the methods. This observation shows that successful transfer learning depends not only on the selection of the source but also on the correct way of leveraging data. The Minimal Transfer deals with heterogeneous replications in an inappropriate way (dropping too much information), which results in inferior performance.

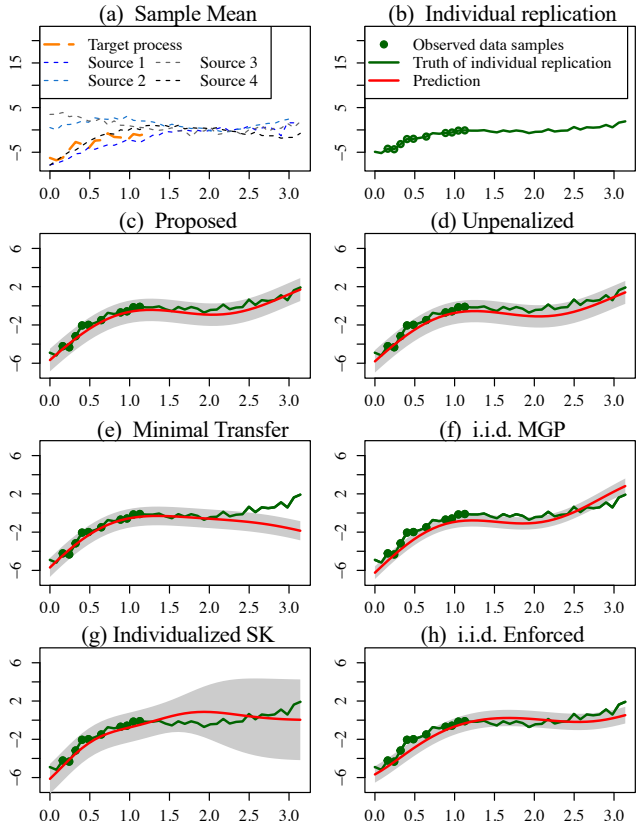


Fig. 6: Prediction results in one specific trial of Setting II

TABLE 2: Median of estimated scaling parameter, Setting II

	Proposed	Unpenalized	Minimal Transfer	i.i.d MGP
α_{1,I_r}	5.90	8.30	3.88	3.95
α_{2,I_r}	2.12	2.24	1.70	1.45
α_{3,I_r}	0.03	-0.97	$1.35e^{-3}$	0.24
α_{4,I_r}	0.75	0.34	1.70	2.01

4.4 Setting III (large scale processes)

In this section, we test the performance of the proposed method and five benchmark methods with large scale source processes. We make slight modifications on the Setting II

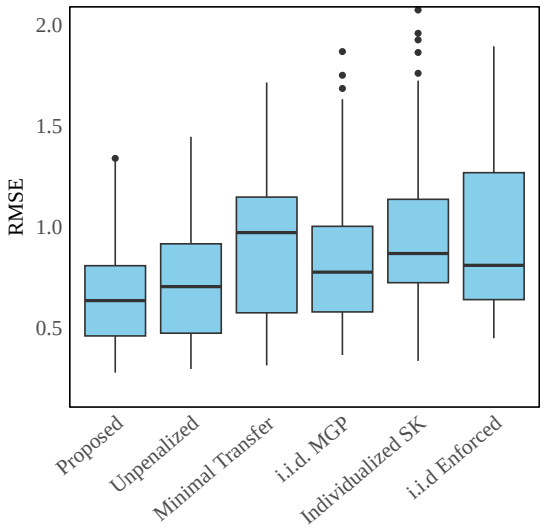


Fig. 7: Setting II RMSE with 100 trials.

and construct the setting of sources as follows:

$$\begin{aligned}
y_q(x_q) &= 0.3(x_q - 2.5 - e_1^q)^3 + \epsilon_q^{(m)}(x_q),\\
y_{\lambda+q}(x_{\lambda+q}) &= 3(x_{\lambda+q})^3 + 2\sin(2x_{\lambda+q} + e_2^q) + \epsilon_{\lambda+q}^{(m)}(x_{\lambda+q}),\\
y_{2\lambda+q}(x_{2\lambda+q}) &= (x_{2\lambda+q} - 1.5 - e_3^q)^2 + \epsilon_{2\lambda+q}^{(m)}(x_{2\lambda+q}),\\
y_{3\lambda+q}(x_{3\lambda+q}) &= (x_{3\lambda+q} - 1)(x_{3\lambda+q} - 2)\cdot(x_{3\lambda+q} - 3.5 - e_4^q)\\
&\quad + \epsilon_{3\lambda+q}^{(m)}(x_{3\lambda+q})
\end{aligned} \tag{19}$$

and the target process is:

$$\begin{aligned}
y_{4\lambda+1}(x_{4\lambda+1}) =& 0.2(x_{4\lambda+1} - 2.5 - e_1^1)^2 + 0.15(x_{4\lambda+1})^2\\
&+ \sin(2x_{4\lambda+1} + e_2^1) + \epsilon_{4\lambda+1}^{(m)}(x_{4\lambda+1})
\end{aligned} \tag{20}$$

where the source processes are with four different groups, the λ is the number of source processes in each group, $q \in \{1, \cdots, \lambda\}$ is the index of source process in each group, and $e_1^q, e_2^q, \cdots, e_4^q$ are all independently sampled from $U(0,1)$. The sampling procedures for x_i are the same as those in Setting II, but we allow more replications in source processes, i.e., $6 \le m_i \le 10$. The target process still has $m_I = 3$ replications. We select $\lambda = 4$ and $\lambda = 25$, i.e., the total number of processes $I = 17$ and $I = 101$, to demonstrate the performance. Due to the large number of processes, we directly show the box plot of these two scenarios in Fig. 8.

The performance among different methods in Fig. 8 is consistent with observations in Setting I and Setting II, where the proposed method achieves the best prediction performance (RMSE median and variance). Besides, there are some critical observations under the large-scale case. First, the variances in all methods increase significantly from $I = 17$ to $I = 101$ except for the Individualized SK. This is because the Individualized SK is the only method that does not involve transfer learning (not affected by the number of sources). Second, the RMSE median of the Minimal Transfer and the i.i.d. MGP is similar to that in Individualized SK. This is a strong indication that as the number of processes

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2025.3607773

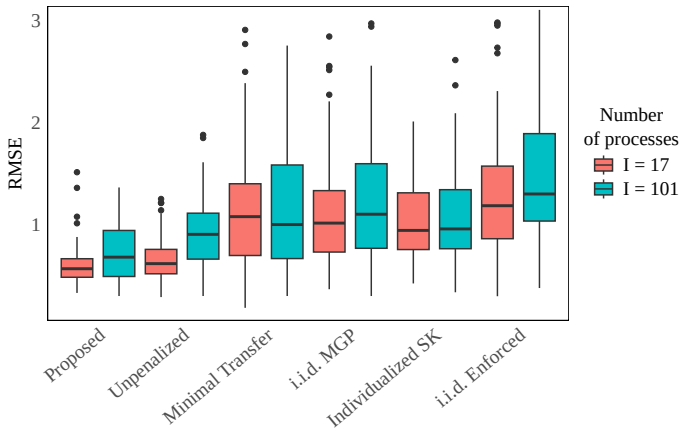


Fig. 8: Setting III RMSE under a large number of processes.

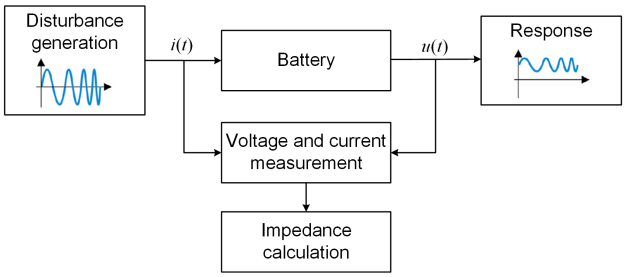


Fig. 9: Measurement diagram of battery impedance.

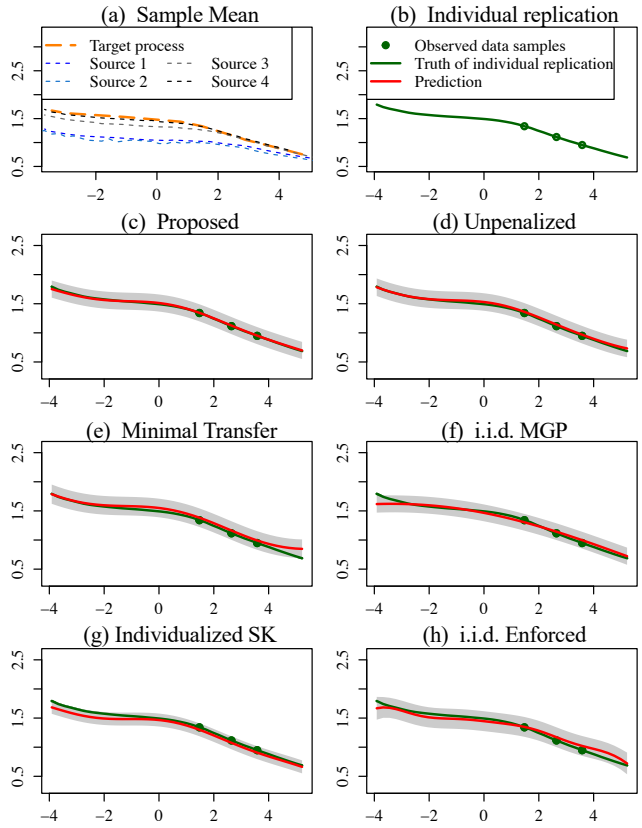


Fig. 10: Prediction results in one trial of EIS battery data.

increases, the correct way of modeling the data becomes vital to the performance. Otherwise, the transfer learning results will be quite similar to a non-transfer case. Finally, the i.i.d. Enforced achieves the worse performance, which is due to the inappropriate correlation structure and the violation of the non-i.i.d. assumption. These observations again validate the effectiveness and contribution of the proposed transfer learning framework in dealing with issues of individualization and scarce data.

5 Case Study

In this section, we validate the performance of the proposed method using data in two case studies: the electrochemical impedance spectroscopy (EIS) test and the reduced graphene oxide (RGO) field-effect transistors (FET) based sensors.

The EIS test is a widely used technique to measure the internal state of electrochemical systems, such as lithium-ion batteries [36]. The measurement diagram is shown in Fig. 9, where a disturbance current is fed to the battery to get an impedance response. It is desired to obtain the impedance response under different frequencies of the disturbance signal so that a comprehensive understanding of the battery can be obtained. However, the EIS test can only input one frequency disturbance signal at a time, and the measurement process is time-consuming. In practice, the EIS test is usually conducted in batches, where batteries under the same condition, e.g., state of charge (SoC), will be tested under different frequencies to obtain the responses. For example, Fig. 1 demonstrates the response data of three batteries with the same SoC (samples belonging to the same curve are from one battery). As we stated the response functions of these three batteries are highly non-linear and auto-correlated with non-i.i.d. noise, and it is difficult to predict the underlying truth (the solid line) of any of these batteries with few data points. To boost the understanding of each of the individual battery's responses, it is intuitive to leverage testing results from batteries in other SoCs. In our experiment setting, there are five different SoC groups, and each group contains 6 batteries. We randomly select one SoC group as our target process and treat the remaining as sources. In each source process, we randomly select 5 to 10 frequencies and their responses in each battery. In the target process, we randomly select 3 frequencies and their responses for each battery. The mean responses of 6 batteries in each group are shown in Fig. 10 (a), the individual replication to be predicted is in Fig. 10 (b), and the prediction performance using different methods is in Fig. 10 (c) to (h).

The prediction performance in Fig. 10 validates that the proposed method can achieve decent predictions even with only 3 available data points in an individual replication. The Unpenalized method also achieves a reasonable trend, but the prediction quality is inferior to the proposed method (when x is around 0 and 4). All other three benchmark methods provide poor predictions or even wrong trends. The same pattern is more clear in the box plot of 80 trials in Fig. 11, where the proposed method achieves the best performance. The Minimal Transfer has the largest variance among all methods. This is because the Minimal Transfer drops most of the replications in sources, and the remaining replication only has 5 to 10 data points. The

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2025.3607773

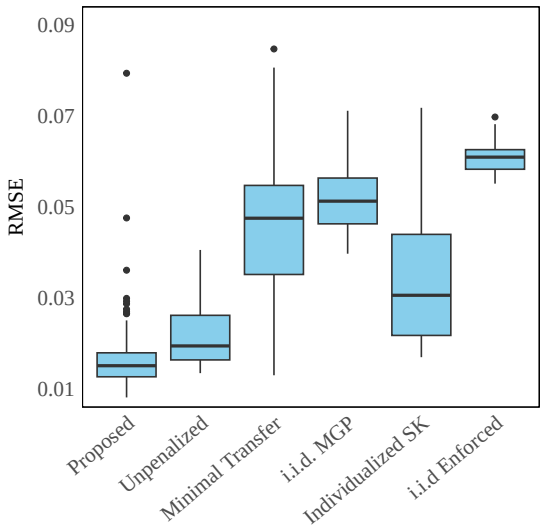


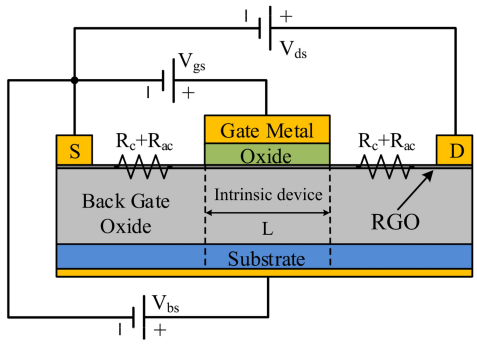
Fig. 11: EIS battery impedance data RMSE with 80 trials.

limited data points and significant information loss together lead to the large uncertainty in Minimal Transfer. The i.i.d. MGP/Enforced, on the other hand, perform relatively consistently, but they have larger RMSE values, showing worse prediction quality. This is mainly because of the strong non-i.i.d. feature and significant heterogeneity among replications in the case study data (as shown in Fig. 1). As a result, the prediction that focuses on the mean trend cannot adapt to the functional shape of each individual replication. This also demonstrates the importance of incorporating non-i.i.d. features into the prediction. The Individualized SK also suffers from a large variance, which is due to the few data points in the target process (only three in each replication).

The second case study is to predict the functional response from reduced graphene oxide field-effect transistors based sensors, which have wide applications in bioengineering and environment protection [37, 38]. The basic structure of a RGO FET is shown in Fig. 12 (a), where the V_{gs} is the gate voltage and the V_{ds} is the drain-source voltage. When an object to be monitored, e.g., protein molecule or chemical ion, touches the RGO, the reaction between the RGO and the object will change the resistance between the drain and the source so that the change in I_{ds} can report the detection of the object. Due to different reaction mechanisms between the RGO and the objects to be monitored , different V_{bs} and V_{ds} values will be used for sensing different objects/materials [39]. However, every RGO FET sensor is disposable, which means it is desirable to use as few sensors as possible to predict the V_{gs} vs. I_{ds} relationship for a new detection task, i.e., the combination of V_{bs} and V_{ds}.

Such task is feasible using transfer learning because there are many V_{gs} vs. I_{ds} relationships available in previous tasks. This is shown in Fig. 12 (b), where the V_{gs} vs. I_{ds} curves under a specific V_{ds} construct a "process". In each process, different V_{bs} provide different curves. In practice, a "process" (V_{ds}) represents an intensity of the materials to be sensed, e.g., lead ion or mercury icon, and the V_{bs} represents the variations in sensing environment. It is clear the functional curves have strong within-and between-process

(a) The basic structure of a RGO FET



(b) V_{gs} vs. I_{ds} curves in RGO FET

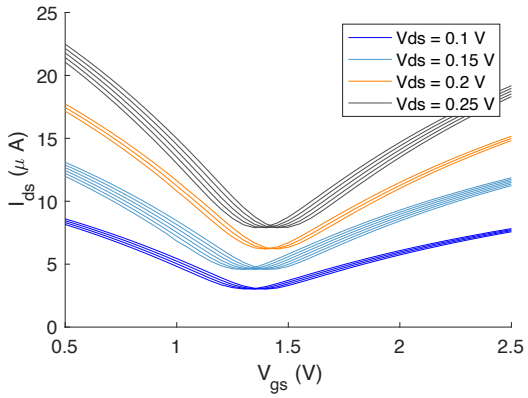


Fig. 12: RGO FET setup and signals.

correlations in Fig. 12 (b). To validate the effectiveness of our proposed method, we randomly select one curve in one process as the replication of interest in the target, i.e., I_t. Figure 13 presents the prediction performance in one specific replication, where the proposed again achieves the best prediction result. It is worth noting that the Unpenalized achieves similar prediction results with a bit larger confidence interval. This is indeed an expected case in this case study because the sources and the target are quite similar (as shown in Fig. 12 (b)), which explains the minor difference between the proposed and the Unpenalzied. This observation is further confirmed in the box plot of 80 trials in Fig. 14, where the Unpenalized is similar to the proposed except for a larger variance. Other benchmark methods all perform worse than the proposed, and the consistent conclusion can be made from this case study is that our proposed framework can achieve superior performance in transfer learning for individualized prediction.

6 Conclusion

In this paper, we propose a novel transfer learning framework to deal with the individualization and data-scarce issues in traditional SK. The proposed framework features a within-process model for facilitating individual prediction and a between-process model for mitigating negative transfer. The within-and between-process models are integrated through a tailored convolution process, which quantifies the within-and between-process interactions with a specially designed covariance matrix and corresponding kernel parameters. We analyze the parameter estimation of

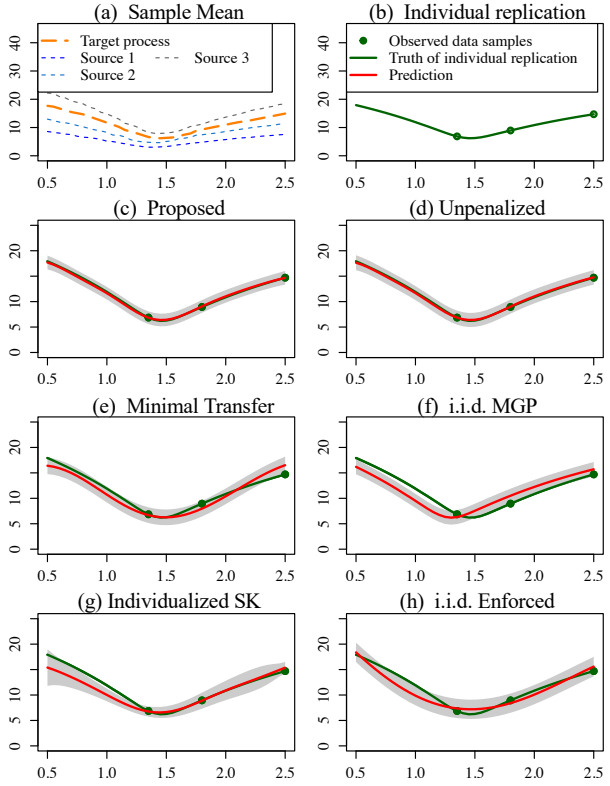


Fig. 13: Prediction results in one trial of RGO FET data.

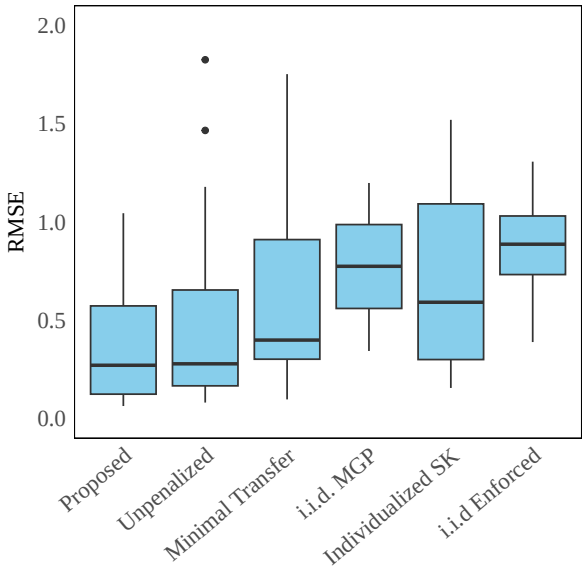


Fig. 14: RGO FET data RMSE with 80 trials.

the proposed framework and provide theoretical guarantees of the transfer learning performance. Various numerical studies are designed to test and compare the performance of the proposed method with alternative solutions. A real case study also validates the effectiveness of the proposed method in dealing with individual prediction of non-i.i.d. data. The superior performance in both numerical and case studies provides evidence that the proposed method is an effective solution to the individualization and data-scarce issues.

There are some opening topics based on our study in this paper. First, the non-i.i.d. noise can take more flexible forms. In this paper, we apply the widely used equicorrelation matrix to represent the non-i.i.d. noise, and other parametric/non-parametric formulations of the non-i.i.d. noise can be considered. However, a more complicated structure of non-i.i.d. noise, e.g., using another CP to construct such structures, may require an additional layer of non-convex objectives in the optimization procedure, which raises concerns about numerical and practical efficiency. Another interesting topic is to apply computationally efficient algorithms to boost the computational speed of the proposed method. For example, variational inference of MGP [40] and online MGP [41] can be considered. However, directly applying these methods might not provide satisfactory prediction results since none of the existing variational methods can simultaneously model the process mean and individualized replication in a transfer learning context. As a result, novel modeling frameworks and inference techniques are necessary to further reduce the computational load of the proposed method. We will explore these topics in our future work.

REFERENCES

[1] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006, vol. 2.
[2] A. Bhosekar and M. Ierapetritou, "Advances in surrogate based modeling, feasibility analysis, and optimization: A review," *Computers & Chemical Engineering*, vol. 108, pp. 250–267, 2018.
[3] J. Q. Shi and T. Choi, *Gaussian process regression analysis for functional data*. CRC press, 2011.
[4] J. Wang, "An intuitive tutorial to gaussian processes regression," *arXiv preprint arXiv:2009.10862*, 2020.
[5] Z. Wang, Q. Zhai, and P. Chen, "Degradation modeling considering unit-to-unit heterogeneity-a general model and comparative study," *Reliability Engineering & System Safety*, vol. 216, p. 107 897, 2021.
[6] C. J. Taylor *et al.*, "A brief introduction to chemical reaction optimization," *Chemical Reviews*, vol. 123, no. 6, pp. 3089–3126, 2023.
[7] B. Ankenman, B. L. Nelson, and J. Staum, "Stochastic kriging for simulation metamodeling," in *2008 Winter simulation conference*, IEEE, 2008, pp. 362–370.
[8] J. Staum, "Better simulation metamodeling: The why, what, and how of stochastic kriging," in *Proceedings of the 2009 Winter Simulation Conference (WSC)*, IEEE, 2009, pp. 119–133.
[9] M. Binois, R. B. Gramacy, and M. Ludkovski, "Practical heteroscedastic gaussian process modeling for large simulation experiments," *Journal of Computational and Graphical Statistics*, vol. 27, no. 4, pp. 808–821, 2018.
[10] P. Wei, T. V. Vo, X. Qu, Y. S. Ong, and Z. Ma, "Transfer kernel learning for multi-source transfer gaussian process regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3862–3876, 2022.
[11] P. Wei, Y. Ke, Y. S. Ong, and Z. Ma, "Adaptive transfer kernel learning for transfer gaussian process regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
[12] B. Cao, S. J. Pan, Y. Zhang, D.-Y. Yeung, and Q. Yang, "Adaptive transfer learning," in *proceedings of the AAAI Conference on Artificial Intelligence*, vol. 24, 2010, pp. 407–412.
[13] P. Goovaerts, *Geostatistics for natural resources evaluation*. Oxford University Press, USA, 1997.
[14] M. Goulard and M. Voltz, "Linear coregionalization model: Tools for estimation and choice of cross-variogram matrix," *Mathematical Geology*, vol. 24, pp. 269–286, 1992.

[15] 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.

[16] M. G. Genton and W. Kleiber, "Cross-covariance functions for multivariate geostatistics," 2015.

[17] B. Rakitsch, C. Lippert, K. Borgwardt, and O. Stegle, "It is all in the noise: Efficient multi-task gaussian process inference with structured residuals," *Advances in neural information processing systems*, vol. 26, 2013.

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

[19] R. Kontar, G. Raskutti, and S. Zhou, "Minimizing negative transfer of knowledge in multivariate gaussian processes: A scalable and regularized approach," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3508–3522, 2020.

[20] X. Wang, C. Wang, X. Song, L. Kirby, and J. Wu, "Regularized multi-output gaussian convolution process with domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 6142–6156, 2022.

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

[22] A. Fallahdizcheh and C. Wang, "Variational inference-based transfer learning for profile monitoring with incomplete data," *IISE Transactions*, pp. 1–16, 2024.

[23] R. Kontar, S. Zhou, C. Sankavaram, X. Du, and Y. Zhang, "Nonparametric modeling and prognosis of condition monitoring signals using multivariate gaussian convolution processes," *Technometrics*, vol. 60, no. 4, pp. 484–496, 2018.

[24] B. Matérn, *Spatial variation*. Springer Science & Business Media, 2013, vol. 36.

[25] B. O'Neill, "The double-constant matrix, centering matrix and equicorrelation matrix: Theory and applications," *arXiv preprint arXiv:2109.05814*, 2021.

[26] R. H. Shumway, D. S. Stoffer, and D. S. Stoffer, *Time series analysis and its applications*. Springer, 2000, vol. 3.

[27] P. J. Huber, "Robust smoothing," *Robustness in statistics*, pp. 33–47, 1979.

[28] D. Duvenaud, "The kernel cookbook: Advice on covariance functions," *URL https://www. cs. toronto. edu/duvenaud/cookbook*, 2014.

[29] M. G. Genton, "Classes of kernels for machine learning: A statistics perspective," *Journal of machine learning research*, vol. 2, no. Dec, pp. 299–312, 2001.

[30] M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur, "Gaussian processes and kernel methods: A review on connections and equivalences," *arXiv preprint arXiv:1807.02582*, 2018.

[31] K. Ritter, *Average-case analysis of numerical problems*. Springer Science & Business Media, 2000.

[32] X. Yue and R. A. Kontar, "Joint models for event prediction from time series and survival data," *Technometrics*, vol. 63, no. 4, pp. 477–486, 2021.

[33] N. Lawrence, "Gaussian process latent variable models for visualisation of high dimensional data," *Advances in neural information processing systems*, vol. 16, 2003.

[34] W. Q. Wang, M. F. Golnaraghi, and F. Ismail, "Prognosis of machine health condition using neuro-fuzzy systems," *Mechanical Systems and Signal Processing*, vol. 18, no. 4, pp. 813–831, 2004.

[35] B. Wang, Y. Lei, N. Li, and N. Li, "A hybrid prognostics approach for estimating remaining useful life of rolling element bearings," *IEEE Transactions on Reliability*, vol. 69, no. 1, pp. 401–412, 2018.

[36] X. Wang *et al.*, "A review of modeling, acquisition, and application of lithium-ion battery impedance for onboard battery management," *ETransportation*, vol. 7, p. 100 093, 2021.

[37] D. Wu *et al.*, "Microvesicle detection by a reduced graphene oxide field-effect transistor biosensor based on a membrane biotinylation strategy," *Analyst*, vol. 144, no. 20, pp. 6055–6063, 2019.

[38] X. Chen *et al.*, "Real-time and selective detection of nitrates in water using graphene-based field-effect transistor sensors," *Environmental Science: Nano*, vol. 5, no. 8, pp. 1990–1999, 2018.

[39] G. Zhou, J. Chang, S. Cui, H. Pu, Z. Wen, and J. Chen, "Real-time, selective detection of pb2+ in water using a reduced graphene oxide/gold nanoparticle field-effect transistor device," *ACS applied materials & interfaces*, vol. 6, no. 21, pp. 19 235–19 241, 2014.

[40] J. Zhao and S. Sun, "Variational dependent multi-output gaussian process dynamical systems," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 4134–4169, 2016.

[41] Z. Hu and C. Wang, "Nonlinear online multioutput gaussian process for multistream data informatics," *IEEE transactions on industrial informatics*, vol. 18, no. 6, pp. 3885–3893, 2021.

Jinwei Yao received B.S. degree in Mathematics from the Stony Brook University in 2018, and the M.S. degree in Financial Engineering from the University of Southern California in 2020. He is currently a PhD student at the Department of Industrial and Systems Engineering at the University of Iowa. His research interest includes transfer learning, multi-system

Jianguo Wu received the B.S. degree in Mechanical Engineering from Tsinghua University, China in 2009, the M.S. degree in Mechanical Engineering from Purdue University in 2011, and M.S. degree in Statistics in 2014 and Ph.D. degree in Industrial and Systems Engineering in 2015, both from University of Wisconsin-Madison. Currently, he is an Associate Professor in the Dept. of Industrial Engineering and Management at Peking University, Beijing, China. He was an Assistant Professor at the Dept. of IMSE at UTEP, TX, USA from 2015 to 2017. His research interests are mainly in quality control and reliability engineering of intelligent manufacturing and complex systems through engineering informed machine learning and advanced data analytics.

Yongxiang Li received his Ph.D. degree in data science from City University of Hong Kong in 2019. Currently, he is an Associate Professor in the Department of Industrial Engineering and Management at Shanghai Jiao Tong University, Shanghai, China. His research focuses on both the theoretical and applied aspects of data science integrated with domain knowledge for quality and reliability engineering using methodologies from statistics, machine learning, and signal processing. He has been working on the research directions such as computer experiments, quality control, anomaly detection, and fault diagnostics.

Chao Wang is an Assistant Professor in the Department of Industrial and Systems Engineering at the University of Iowa. He received his B.S. from the Hefei University of Technology in 2012, and M.S. from the University of Science and Technology of China in 2015, both in Mechanical Engineering, and his M.S. in Statistics and Ph.D. in Industrial and Systems Engineering from the University of Wisconsin-Madison in 2018 and 2019, respectively. His research interests include statistical modeling, analysis, monitoring and control for complex systems.