This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2020.3041830, IEEE Transactions on Industrial Informatics

Sparse and Structured Function-on-Function Quality Predictive Modeling by Hierarchical Variable Selection and Multitask Learning

Kai Wang, and Fugee Tsung

***Abstract*—Modern manufacturing industries are often featured with a data-rich environment. The real-time behaviors of process variables can be completely recorded as multiple various signal signatures, and the geometric quality of finished products can be thoroughly characterized by their two-dimensional surface data. Learning the relationship between such signal predictors and surface responses, where the input and output are no longer the conventional scalar variables but are in fact both functions in the time domain and spatial domain, respectively, is critical for quality prediction in many applications nowadays. To this end, this paper proposes a novel Sparse and Structured Function-on-Function Regression (SSF^{2}R) model, where a hierarchical variable selection is developed to identify informative signals and further screen significant elements within the selected signals, and a multitask learning is devised to exploit the smoothness nature of surface response and the similarity structure among a series of sub-regression tasks. Our SSF^{2}R model is concisely formulated as a convex problem with an efficient iterative algorithm derived to obtain the global optimum. Moreover, our quality prediction can be performed dynamically during an ongoing manufacturing process when only partial observations of the signal predictors are available. The superiority of our proposed method is validated by numerical simulations and a real case study in the semiconductor industry.**

***Index Terms*—Functional data analysis, functional regression, regularization method, sparse learning, structure penalty.**

I. Introduction

RAPID advances in information technology and industrial revolution have created an unprecedented data-rich manufacturing paradigm [1]. The real-time status of all process variables can be continuously recorded by large-scale online sensing nodes, forming a multitude of time-ordered data streams knowns as profiles or signals, and the geometries of fabricated products can be thoroughly examined by modern sophisticated metrology devices, generating ample yet complex quality data such as images and surfaces. Hinging on such industrial big data, learning the relationship between product quality responses and process variable signatures is an extremely promising initiative as a gift of data availability, but also a quite challenging task due to the curse of data complexity. The learned quality predictive model assumes a fundamental role in current soft sensing applications for reducing operational costs and improving production yield [2].

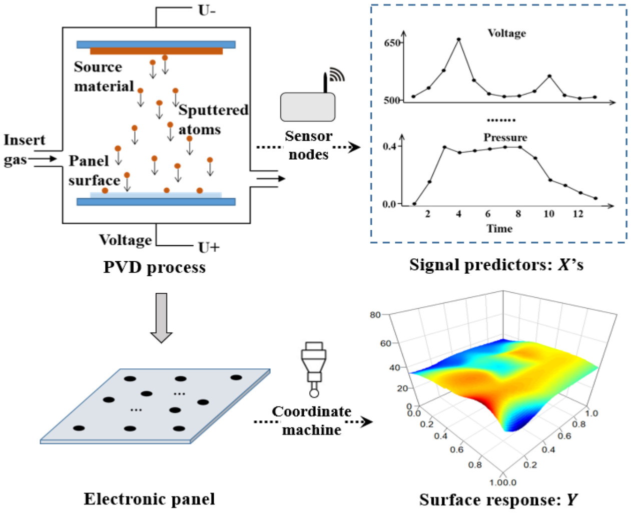


Fig. 1: PVD process with signal predictors and surface response.

In this paper, we propose a novel Functional Regression (FR) model for quality prediction where both the predictors and the response of interest are functions. Particularly, the predictors are multiple in-situ signals which are functions in the time domain, whereas the response is a product surface which is a function in the spatial domain. This work is motivated by a real manufacturing example from the semiconductor industry named the Physical Vapor Deposition (PVD) process (see Fig. 1). This process coats an electronic panel with a thin film in a high-temperature and high-pressure chamber, where the high-energy atoms ejected by a source material fly to and accumulate at the surface of the electronic panel. The signal trajectories of two process variables are demonstrated in Fig. 1. After the manufacturing, the film thickness is measured by a touch-probe coordinate machine at 17 different locations in the electronic panel. Current quality forecast approaches adopted in the semiconductor industry when the target is

Manuscript received July 21, 2020; revised October 23, 2020; accepted November 24, 2020. Date of publication xx xx, xxxx; date of current version xx xx, xxxx. This work was supported in part by the National Natural Science Foundation of China under Grant 71931006 and 71602155, in part by the Hong Kong RGC General Research Funds under Grant 16216119 and 16201718, and in part by the Fellowship of China Postdoctoral Science Foundation under Grant 2020M673430. Paper no. TII-20-3498. (*Corresponding author: Kai Wang.*)

Kai Wang is with the School of Management and State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China (e-mail: kwangai@xjtu.edu.cn).

Fugee Tsung is with the Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology, Hong Kong (e-mail: season@ust.hk).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier xxxxxxxx

1551-3203 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on April 01,2021 at 14:14:43 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2020.3041830, IEEE Transactions on Industrial Informatics

surface thickness include the popular linear regression model and some machine learning methods such as the Support Vector Regression (SVR) and Neural Network (NN) [3]–[6]. However, these existing ones all simplify the entire signal signatures and product surface into a few scalar variables, i.e., they take the mean, range or variance of predictors and response as summary statistics to reduce the data complexity, which in effect fails to utilize the available functional data completely and would cause a great information loss.

In our proposed function-on-function regression model, we follow the Functional Data Analysis (FDA) reasoning [7] where either the process signal or the product surface is regarded as a single complete functional datum of a continuity or smoothness nature, which captures maximum data information and circumvents the traditional feature extraction efforts. When predictive modeling involves numerous predictors, Variable Selection (VS) is often conducted to enhance model interpretation and generalization [8]. For the multiple functional predictors in our context, a new Two-Level Hierarchical VS (TLH-VS) is developed. The main idea is that firstly, at the function level, we identify the informative signals globally that highly impact the quality response, and secondly, at the element level, we further determine the dominant segments or parts locally within the selected signals. Such a TLH-VS accommodates to the inherent hierarchical structure of the multivariate functional predictors, and could enable a more flexible and interpretable sparisty pattern in the estimated regression coefficients.

Additionally, when the response is also conveyed by a function with a smoothness functional integrity, the quality measurements at neighbouring locations in this functional response would exhibit certain similarities. For example, in our PVD process, the film thicknesses at adjacent surface coordinates are typically of similar values. If we regard the predictive model associated with each response measurement in the coated electronic panel as a sub-regression task, we would have a series of 17 such sub-regression tasks, and among them the adjacent ones are prone to have similar coefficients as indicated by the above similarity structure. To exploit this explicitly in our predictive model, we adopt the MultiTask Learning (MTL) [9], [10] to solve these multiple interconnected sub-regression tasks jointly. Our devised mechanism is to intentionally penalize the pair-wise coefficients' differences of these sub-regression tasks and the penalties will become pronounced when two locations in the functional response get close to each other. In this way, we achieve the transfer of knowledge on model parameters across multiple related tasks, i.e., the information on the coefficient values in each task will be transferred into its adjacent tasks to make these tasks together tend to have similar coefficients.

Our proposed FR model which integrates the above two favorable properties, i.e., sparsity and structure, is formulated as a tractable convex problem. An efficient global optimization algorithm via the Alternating Direction Method of Multipliers (ADMM) [11] is derived. Furthermore, our quality predictive model is extended to a dynamical version where only partial observations of the functional predictors are available during the manufacturing progress. The superiority of our Sparse and Structured Function-on-Function Regression (SSF^{2}R) model is validated by extensive numerical simulations and a real case study on the PVD process. Compared with the recently emerging deep learning networks which are actually end-to-end black-box models [12], [13], our method with the functional linear model as one basic building block is simpler, more transparent and user-friendly. Our delivered coefficients measure the predictors' marginal effects and are easier to be understood by practitioners. Besides, here we propose variable selection to gain useful knowledge on predictor importance rather than extract high-level abstract features as in the deep learning. Finally, our model is convex and thus enjoys a global optimum while the deep learning is far more computationally demanding and could be trapped into local optimums.

To sum up, the main contributions of this paper are highlighted as below:

- A unified function-on-function regression model embedded with a sparsity and a structure regularization simultaneously is proposed to address data complexity in both the signal predictors and the surface response.
- A novel TLH-VS via a joint use of the $L_{2,1}$-norm and L_1-norm penalties is developed to identify informative signal signatures as well as significant individual elements within the selected signals.
- A tailored MTL with a weighted parameter-fusion penalty is devised to utilize the interaction structure of multiple sub-regression tasks and to boost the overall estimation and prediction performance.
- An efficient ADMM-based optimization algorithm is derived which can perform in a parallel computing fashion, scale up to big data and converge to the global optimum.
- An extension of our SSF^{2}R model to an online prediction setting is achieved by dynamically completing the unobserved parts of functional predictors in a supervised manner.

Following the Introduction, Section II surveys the related works regarding the FR and MTL. Section III elaborates the technical derivation of our SSF^{2}R model. The performance of our proposed method is investigated in Section IV by numerical simulations and a real case study. Concluding remarks are given in Section V.

II. Related Work

Linear regression is a commonly used prediction tool since it is simple to implement, easy to interpret and fast to compute [8] The FR model is an advanced variant of the linear regression model where either predictors or response or both are functions. It becomes an active area of FDA and receives most research attention recently [14]. A large body of current FR works are confined to a scalar-on-function scenario where the response is a scalar. To name a few, Zhang et al. [15] employed the cubic spline basis functions to represent the functional predictors, and then the multiple linear regression model is applied to regress the scalar response on the basis function coefficients. A second-order derivative roughness penalty is often imposed to control the degree of smoothness of the estimated coefficient functions. The functional predictors can

1551-3203 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on April 01,2021 at 14:14:43 UTC from IEEE Xplore. Restrictions apply.

also be represented in a data-driven manner by the principal components [16]. When the response is also in a functional form, the regression coefficient in this function-on-function scenario turns to be a bivariate function which is defined in the joint domain of the predictors and response, and the relevant works are comparatively less [14]. The two-dimensional (2-D) smooth splines [17] or the products of eigenfunctions of the predictors and response [18] can be taken as proper candidates when the basis functions are still used. The roughness penalty herein regulates the smoothness of the bivariate coefficient function in the marginal and diagonal directions [19].

Besides the basis function representation and the roughness level governance, the sparsity regularization has also been widely used in the FR models [14] to cope with the high-dimensional challenge, and more importantly, to make the regression model more interpretable via variable selection. When only a single functional predictor is involved, the Least Absolute Shrinkage and Selection Operator (LASSO) or L_1-norm penalty can be adopted to facilitate the basis function selection [20]. For multiple functional predictors, the basis function coefficients associated with each predictor form a group, and the group LASSO or $L_{2,1}$-norm penalty is typically used to select informative functional predictors or groups [21], [22]. The preceding VS for multivariate functional predictors, however, can only perform at the function level in our context to identify informative signal signatures as a whole. The TLH-VS as introduced in Section I which is able to further screen local elements in the selected signals, is rarely studied in the existing FR works. One exception is Paynabar et al. [23] where a two-step nonnegative garrote VS technique is proposed for a scalar response, but their step-by-step selection manner does not guarantee the globally optimal VS result.

In modern machine learning field, the MTL is an important research branch [9], [10]. It aims to learn multiple different but related tasks together and to enhance the overall prediction accuracy by sharing the knowledge learned from these tasks. Conceptually, an MTL scheme can be built via either the feature transfer that uncovers a set of common latent features across tasks [24], [25] or the parameter transfer that assigns a joint prior distribution for model parameters from all tasks [26], [27]. Technically, the regularization is one main approach to explicitly utilize the task relatedness in MTL by inducing particular structures in the estimated model parameters [28]. A variety of regularization terms have been well designed with different exploitations on the task relationship, such as the group LASSO which promotes consistent VS results among similar tasks [29] and the fused LASSO which encourages equivalent model parameters for neighbouring tasks [30]. The use of the MTL in the function-on-function regression scenario to leverage the inherent smoothness integrity of the functional response, that is, to explore the potential similarity structure among a range of sub-regression tasks as discussed in Section I, has not yet been investigated.

III. Methodology

This section first provides the preliminaries of function-on-function regression model. Then the TLH-VS via a sparsity regularization and the MTL via a structure regularization are developed to formulize the proposed SSF^{2}R model. An efficient iterative optimization algorithm is derived for model parameter estimation. The extension to dynamical prediction with partial observations is finally discussed.

A. Function-on-Function Regression Model

Suppose we have collected a training dataset which consists of N historical samples. For the ith sample, $i = 1,\dots,N$, $x_{ij}(t)$, $j = 1,\dots,P$, denotes the jth functional predictor which is a signal signature and $t \in \mathscr{T}$ indicates the time domain, and $y_i(s)$ denotes the functional response which is a product surface and $s \in \mathscr{S}$ implies the spatial domain. In the function-on-function regression model [14], we have

$$y_i(s) = \sum_{j=1}^{P} \int_{\mathscr{T}} x_{ij}(t)\theta_j(t,s)dt + \varepsilon_i(s), \quad i = 1,\dots,N, \tag{1}$$

where $\theta_j(t,s)$ is a bivariate coefficient function which measures the effect of the jth predictor on the response and such an effect could be changeable as t or s varies in their respective time or spatial domain. In addition, $\varepsilon_i(s)$ is a zero-mean random error function. Note that here $x_{ij}(t)$ and $y_i(s)$ have been centered and standardized over the N samples, so the intercept term can be omitted.

In practice, the continuous functions $x_{ij}(t)$'s and $y_i(s)$'s are always unavailable, but rather they are evaluated at a fine gird. Specifically, suppose the time domain of the signal predictors is discretized at $t = 1,\dots,T$. The spatial domain of the surface response is indexed by $s = 1,\dots,S$, where each spatial index s is attributed with an in-plane or 2-D coordinate to label its true location on the product surface. Then the discrete version of (1) is

$$y_{is} = \sum_{j=1}^{P}\sum_{t=1}^{T} x_{ijt}\theta_{jts} + \varepsilon_{is}, \quad i = 1,\dots,N, \quad s = 1,\dots,S. \tag{2}$$

Let $\mathbf{y}_s = (y_{1s},\dots,y_{Ns})^T$ be a vector including all response observations at the sth index, and then $Y = (\mathbf{y}_1,\dots,\mathbf{y}_S)$ denotes the response matrix. We also denote the design matrix and coefficient matrix by

$$X = \begin{pmatrix} \mathbf{x}_{11\cdot}^T & \cdots & \mathbf{x}_{1P\cdot}^T \\ \vdots & \ddots & \vdots \\ \mathbf{x}_{N1\cdot}^T & \cdots & \mathbf{x}_{NP\cdot}^T \end{pmatrix}_{N\times PT,} \qquad \Theta = \begin{pmatrix} \theta_{1\cdot 1} & \cdots & \theta_{1\cdot S} \\ \vdots & \ddots & \vdots \\ \theta_{P\cdot 1} & \cdots & \theta_{P\cdot S} \end{pmatrix}_{PT\times S,}$$

where $\mathbf{x}_{ij\cdot} = (x_{ij1},\dots,x_{ijT})^T$ is the vectorization of $x_{ij}(t)$, and $\theta_{j\cdot s} = (\theta_{j1s},\dots,\theta_{jTs})^T$ is the vector version of $\theta_j(t,s)$ along the index t and for a fixed index s, which links the jth predictor and the response at the sth index. The matrix form of (2) is

$$Y = X\Theta + \varepsilon, \tag{3}$$

where ε is the error matrix.

Our quality predictive model can be established by minimizing the following composite objective function:

$$\min_{\Theta} \frac{1}{2}||Y - X\Theta||_F^2 + \lambda h(\Theta), \tag{4}$$

1551-3203 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on April 01,2021 at 14:14:43 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2020.3041830, IEEE Transactions on Industrial Informatics

where $||\cdot||_F^2$ is the square of the Frobenius norm of a matrix, $h(\Theta)$ is a regularization term which is usually used to suppress overly model complexity and facilitate proper parameter structure [8], and $\lambda \geq 0$ is the tuning parameter. The specific form of $h(\Theta)$ to achieve our TLH-VS and MTL is devised in the sections below.

B. TLH-VS: Sparsity Regularization

Here we configure a sparsity regularization to realize the TLH-VS. First of all, at the function level, if a signal predictor (say the jth one) is not informative, i.e., its entire signature does not have any impact on the quality response, the regression coefficients θ_{jts}, $t = 1,\ldots,T$, associated with this signal predictor will all be zero, i.e., $\boldsymbol{\theta}_{j\cdot s} = \mathbf{0}$. To exclude these coefficients together, the group LASSO or $L_{2,1}$-norm penalty is taken as

$$h_1(\Theta) = \sum_{s=1}^{S}\sum_{j=1}^{P} ||\boldsymbol{\theta}_{j\cdot s}||_2 = ||\Theta||_{2,1}, \tag{5}$$

where $||\boldsymbol{\theta}_{j\cdot s}||_2 = \sqrt{\theta_{j1s}^2 + \ldots + \theta_{jTs}^2}$ is the L_2-norm of $\boldsymbol{\theta}_{j\cdot s}$, and $||\Theta||_{2,1}$ is defined as the sum of these L_2-norms.

Next, even though a signal predictor is informative, it is possible that only a few of its elements are significant, i.e., only some of the coefficients θ_{jts}, $t = 1,\ldots,T$, are non-zero when $\boldsymbol{\theta}_{j\cdot s} \neq \mathbf{0}$. At this element level, to zero out the negligible individuals within $\boldsymbol{\theta}_{j\cdot s}$, we further consider the regular LASSO or L_1-norm penalty:

$$h_2(\Theta) = \sum_{s=1}^{S}\sum_{j=1}^{P}\sum_{t=1}^{T} ||\theta_{jts}||_1 = ||\Theta||_1. \tag{6}$$

Our proposed sparsity regularization for the TLH-VS is a combination of $h_1(\Theta)$ and $h_2(\Theta)$ in (5)-(6):

$$h_{sp}(\Theta) = \lambda_1 h_1(\Theta) + \lambda_2 h_2(\Theta) = \lambda_1||\Theta||_{2,1} + \lambda_2||\Theta||_1, \tag{7}$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are the tuning parameters. The sparsity regularization $h_{sp}(\Theta)$ can induce an adequate sparsity pattern in both the signal signatures and the elements in the selected signals hierarchically (see Fig. 2(a) for illustration). Notably, it is also quite flexible for the surface response as it allows for heterogeneous VS results when the spatial index s is different.

C. MTL: Structure Regularization

We shall now focus on the MTL via a structure regularization. Note that the function-on-function regression model in (2) is actually comprised of a series of S sub-regression tasks, each of which predicts the surface response evaluated at one particular spatial index s. For any two spatial indexes s and s', we first calculate a similarity measure $c(s,s')$ based on the distance of their respective 2-D coordinates, where $c(s,s')$ can be numerical in $[0,1]$ or binary in $\{0,1\}$ (see Section IV for example).

Then our MTL is devised in light of the intuition that when two spatial indexes s and s' are highly adjacent in the surface response with a large $c(s,s')$, their associated sub-regression tasks would have approximate coefficients, i.e., $\boldsymbol{\theta}_{j\cdot s} \approx \boldsymbol{\theta}_{j\cdot s'}$, $j = 1,\ldots,P$. To explicitly utilize such a similarity structure among these sub-regression tasks, a weighed fused LASSO penalty is defined as:

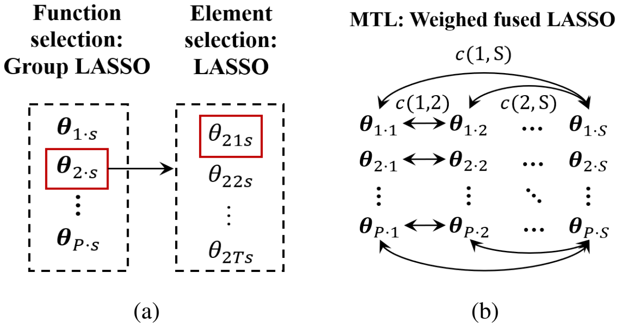

Function selection: Group LASSO
Element selection: LASSO
MTL: Weighed fused LASSO
$\boldsymbol{\theta}_{1\cdot s}$
$\boldsymbol{\theta}_{2\cdot s}$
$\boldsymbol{\theta}_{P\cdot s}$
θ_{21s}
θ_{22s}
θ_{2Ts}
$c(1,S)$
$c(1,2)$
$c(2,S)$
$\boldsymbol{\theta}_{1\cdot 1}$ $\boldsymbol{\theta}_{1\cdot 2}$ $\boldsymbol{\theta}_{1\cdot S}$
$\boldsymbol{\theta}_{2\cdot 1}$ $\boldsymbol{\theta}_{2\cdot 2}$ $\boldsymbol{\theta}_{2\cdot S}$
$\boldsymbol{\theta}_{P\cdot 1}$ $\boldsymbol{\theta}_{P\cdot 2}$ $\boldsymbol{\theta}_{P\cdot S}$
(a)
(b)

Fig. 2: (a) TLH-VS and (b) MTL.

$$h_3(\Theta) = \sum_{j=1}^{P}\sum_{t=1}^{T}\sum_{s=1}^{S-1}\sum_{s'>s} c(s,s')||\theta_{jts} - \theta_{jts'}||_1 = ||\Theta C||_1, \tag{8}$$

where $c(s,s')$ acts as a weight to impose more severe penalty when s and s' tend to be closer, and

$$C = \begin{pmatrix} c(1,2) & \cdots & c(1,S) & \cdots & 0 \\ -c(1,2) & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & c(S-1,S) \\ 0 & \cdots & -c(1,S) & \cdots & -c(S-1,S) \end{pmatrix}_{S\times\frac{S(S-1)}{2}}$$

is used to calculate the weighted pair-wise differences of coefficients in each row of Θ. See the demonstration of this penalty design in Fig. 2(b).

So far, the potential similarity structure of our bivariate coefficient function $\theta_j(t,s)$ over s has been exploited in (8), but it also possesses a smoothness nature over t when s is fixed. As such, we finally admit another penalty term [7] to our MTL which controls the coefficient function's roughness longitudinally in the time domain:

$$h_4(\theta) = \sum_{s=1}^{S}\sum_{j=1}^{P} ||\widetilde{D}\boldsymbol{\theta}_{j\cdot s}||_2^2 = ||D\Theta||_F^2, \tag{9}$$

where $\widetilde{D}$ is the numerical second-order derivative operator

$$\widetilde{D} = \begin{pmatrix} 1 & -2 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & -2 & 1 \end{pmatrix}_{(T-2)\times T,}$$

and D is a P-block diagonal matrix with $\widetilde{D}$ as its diagonals.

Our structure regularization is a summation of $h_3(\Theta)$ and $h_4(\Theta)$ in (8)-(9):

$$h_{st}(\Theta) = \lambda_3 h_3(\Theta) + \lambda_4 h_4(\Theta) = \lambda_3||\Theta C||_1 + \lambda_4||D\Theta||_F^2, \tag{10}$$

where $\lambda_3 \geq 0$ and $\lambda_4 \geq 0$ are the tuning parameters.

1551-3203 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on April 01,2021 at 14:14:43 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2020.3041830, IEEE Transactions on Industrial Informatics

D. Optimization Algorithm

Our SSF^{2}R model is finally built by combining (4), (7) and (10):

$$\min_{\Theta} \frac{1}{2}||Y-X\Theta||_F^2+\lambda\gamma\alpha||\Theta||_{2,1}+\lambda\gamma(1-\alpha)||\Theta||_1 \\ +\lambda(1-\gamma)\beta||\Theta C||_1+\lambda(1-\gamma)(1-\beta)||D\Theta||_F^2, \quad (11)$$

where the tuning parameters are reparameterized as $\lambda_1=\lambda\gamma\alpha$, $\lambda_2=\lambda\gamma(1-\alpha)$, $\lambda_3=\lambda(1-\gamma)\beta$ and $\lambda_4=\lambda(1-\gamma)(1-\beta)$ with $\lambda\geq 0$ and $\gamma,\alpha,\beta\in[0,1]$ to simplify their selections.

Since in (11) the quadratic function $||\cdot||_F^2$ and the norm functions $||\cdot||_{2,1}$ and $||\cdot||_1$ are all convex, our optimization problem as a summation of these convex functions is also convex. However, it is not differentiable since the L_1-norm function $||\cdot||_1$ is not differentiable when evaluated at the zero point. To solve this, we apply the ADMM [11] to derive an efficient algorithm. To be specific, we introduce the auxiliary variables $Z=\{Z^{(1)},Z^{(2)},Z^{(3)},Z^{(4)}\}$ and rewrite (11) as

$$\begin{aligned}\min_{\Theta,Z}\ & \frac{1}{2}||Y-X\Theta||_F^2+\lambda\gamma\alpha||Z^{(1)}||_{2,1}+\lambda\gamma(1-\alpha)||Z^{(2)}||_1 \\ & +\lambda(1-\gamma)\beta||Z^{(3)}||_1+\lambda(1-\gamma)(1-\beta)||Z^{(4)}||_F^2, \\ \text{s.t. }\ & \Theta-Z^{(1)}=\mathbf{0}, \quad \Theta-Z^{(2)}=\mathbf{0}, \\ & \Theta C-Z^{(3)}=\mathbf{0}, \quad D\Theta-Z^{(4)}=\mathbf{0}. \end{aligned} \quad (12)$$

The augmented Lagrangian $L_\rho(\Theta,Z,\eta)$ is

$$\begin{aligned}&\frac{1}{2}||Y-X\Theta||_F^2+\lambda\gamma\alpha||Z^{(1)}||_{2,1}+\lambda\gamma(1-\alpha)||Z^{(2)}||_1 \\ &+\lambda(1-\gamma)\beta||Z^{(3)}||_1+\lambda(1-\gamma)(1-\beta)||Z^{(4)}||_F^2 \\ &+\mathrm{Tr}\big(\eta^{(1)^T}(\Theta-Z^{(1)})\big)+\mathrm{Tr}\big(\eta^{(2)^T}(\Theta-Z^{(2)})\big)+ \\ &+\mathrm{Tr}\big(\eta^{(3)^T}(\Theta C-Z^{(3)})\big)+\mathrm{Tr}\big(\eta^{(4)^T}(D\Theta-Z^{(4)})\big) \\ &+\frac{\rho}{2}||\Theta-Z^{(1)}||_F^2+\frac{\rho}{2}||\Theta-Z^{(2)}||_F^2 \\ &+\frac{\rho}{2}||\Theta C-Z^{(3)}||_F^2+\frac{\rho}{2}||D\Theta-Z^{(4)}||_F^2,\end{aligned}$$

where $\eta=\{\eta^{(1)},\eta^{(2)},\eta^{(3)},\eta^{(4)}\}$ are the Lagrangian multipliers corresponding to the four constraints in (12), and $\rho>0$ is called the penalty parameter in ADMM.

Then Θ,Z,η can be estimated by proceeding the following iterative updating steps.

- Given Z and η, we have $\mathrm{vec}(\widehat{\Theta})=\mathbf{A}^{-1}\mathbf{b}$, where

$$\begin{aligned}\mathbf{A}&=\mathbf{I}_S\otimes X^TX+2\rho\mathbf{I}_{PTS}+\rho CC^T\otimes\mathbf{I}_{PT}+\rho\mathbf{I}_S\otimes D^TD \\ &=(\rho CC^T+\rho\mathbf{I}_S)\otimes\mathbf{I}_{PT}+\mathbf{I}_S\otimes(X^TX+\rho D^TD+\rho\mathbf{I}_{PT}) \\ &=(\rho CC^T+\rho\mathbf{I}_S)\oplus(X^TX+\rho D^TD+\rho\mathbf{I}_{PT}), \\ \mathbf{b}&=(\mathbf{I}_S\otimes X^T)\mathrm{vec}(Y)-\mathrm{vec}(\eta^{(1)})-\mathrm{vec}(\eta^{(2)}) \\ &-(C\otimes\mathbf{I}_{PT})\mathrm{vec}(\eta^{(3)})-(\mathbf{I}_S\otimes D^T)\mathrm{vec}(\eta^{(4)}) \\ &+\rho\mathrm{vec}(Z^{(1)})+\rho\mathrm{vec}(Z^{(2)}) \\ &+\rho(C\otimes\mathbf{I}_{PT})\mathrm{vec}(Z^{(3)})+\rho(\mathbf{I}_S\otimes D^T)\mathrm{vec}(Z^{(4)}),\end{aligned}$$

  and $\mathrm{vec}(\cdot)$, $\otimes$ and $\oplus$ are the vectorization, Kronecker product and Kronecker sum operators, respectively.

- Given Θ and η, we optimize Z as

$$\begin{aligned}\widehat{\mathbf{z}}_{j\cdot s}^{(1)}&=\Big(1-\frac{\lambda\gamma\alpha}{||\rho\theta_{j\cdot s}+\eta_{j\cdot s}^{(1)}||_2}\Big)_+\big(\theta_{j\cdot s}+\frac{\eta_{j\cdot s}^{(1)}}{\rho}\big),\ \forall s,\ \forall j, \\ \widehat{z}_{jts}^{(2)}&=\Big(1-\frac{\lambda\gamma(1-\alpha)}{||\rho\theta_{jts}+\eta_{jts}^{(2)}||_1}\Big)_+\big(\theta_{jts}+\frac{\eta_{jts}^{(2)}}{\rho}\big),\ \forall s,\ \forall j,\ \forall t, \\ \widehat{z}_{jtss'}^{(3)}&=\Big(1-\frac{\lambda(1-\gamma)\beta}{||\rho\big((C^T\otimes\mathbf{I}_{PT})\mathrm{vec}(\Theta)\big)_{jtss'}+\eta_{jtss'}^{(3)}||_1}\Big)_+ \\ &\Big(\big((C^T\otimes\mathbf{I}_{PT})\mathrm{vec}(\Theta)\big)_{jtss'}+\frac{\eta_{jts}^{(3)}}{\rho}\Big),\ \forall j,\ \forall t,\ \forall(s,s'), \\ \widehat{z}_{jts}^{(4)}&=\frac{\eta_{jts}^{(4)}+\rho\big((\mathbf{I}_S\otimes D)\mathrm{vec}(\Theta)\big)_{jts}}{\rho+2\lambda(1-\gamma)(1-\beta)},\ \forall s,\ \forall j,\ \forall t,\end{aligned}$$

  where $(a)_+$ is the soft-thresholding operator which is equal to a if $a>0$ and zero otherwise.

- Given Θ and Z, we perform

$$\widehat{\eta}^{(1)}=\widehat{\eta}^{(1)^l}+\rho(\Theta-Z^{(1)}),\ \widehat{\eta}^{(2)}=\widehat{\eta}^{(2)^l}+\rho(\Theta-Z^{(2)}), \\ \widehat{\eta}^{(3)}=\widehat{\eta}^{(3)^l}+\rho(\Theta C-Z^{(3)}),\ \widehat{\eta}^{(4)}=\widehat{\eta}^{(4)^l}+\rho(D\Theta-Z^{(4)}),$$

  where the superscript l denotes the last iteration step.

The dominant computing burden of our algorithm lies in the updating of Θ which involves the inversion of $\mathbf{A}$. Fortunately, this difficulty can be skipped based on the property of the Kronecker sum operator. Specifically, we first apply the eigendecomposition on CC^T and $X^TX+\rho D^TD$ to derive their eigenvalues and eigenvectors, and then we can obtain $\mathbf{A}^{-1}$ by properly manipulating these intermediate results. Such a trick reduces our time complexity from $O\big((SPT)^3\big)$ to $O\big(\max\{S^3,(PT)^3\}\big)$, and $\mathbf{A}^{-1}$ in fact has only to be calculated once during the entire iteration. Other terms such as $\mathbf{I}_S\otimes X^T$, $C\otimes\mathbf{I}_{PT}$ and $\mathbf{I}_S\otimes D^T$ in $\mathbf{b}$ can also be cached in advance for their repeated use. Therefore, the overall computation cost of our algorithm is modest, and the updating of Z can be performed parallelly to further enhance the computational efficiency. Based on the convergency property of the ADMM for convex optimization [11], our algorithm is guaranteed to get the global optimal solution.

E. Dynamical Prediction

The SSF^{2}R model built in the above sections takes the complete signatures of signal predictors as input and outputs a predicted surface after the manufacturing is finished, which is useful in the soft sensing applications to save measurement costs and detect quality anomalies [2], but it can also be extended to a dynamical situation where the data points of signal predictors arrive progressively during an ongoing manufacturing process. That is, at any time before the manufacturing process is completely over, we can predict the final product surface based only on the partial observations of the signal predictors. Suppose for a particular sample under study, up to current time t, $1\leq t<T$, we only have the fore parts of its functional predictors denoted by $\mathbf{x}_{jU_t}=(x_{j1},\ldots,x_{jt})^T$, $j=1,\ldots,P$, $U_t=\{1,\ldots,t\}$. To make our method applicable,

1551-3203 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on April 01,2021 at 14:14:43 UTC from IEEE Xplore. Restrictions apply.

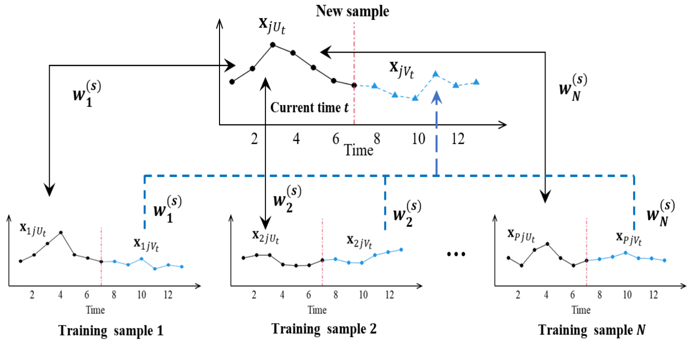


Fig. 3: SFPC for y_s prediction.

we have to complete these functional predictors, i.e., we desire to estimate $\mathbf{x}_{jV_t} = (x_{j(t+1)}, \dots, x_{jT})^T$ where $V_t = \{t+1, \dots, T\}$.

Our completion relies on the historical training samples $\mathbf{x}_{ij\cdot}$, $j = 1, \dots, P$, $i = 1, \dots, N$, each of which is also split into two parts $\mathbf{x}_{ijU_t}$ and $\mathbf{x}_{ijV_t}$. When considering the quality prediction at the sth spatial index in the surface response, we define the contribution of the ith training sample as

$$w_i^{(s)} = \exp\Big(-\kappa \sum_{j=1}^{P} \sum_{t \in U_t} (x_{ijt} - x_{jt})^2 |\widehat{\theta}_{jts}|\Big), \tag{13}$$

where κ is a scale parameter. Then

$$\widehat{\mathbf{x}}_{jV_t}^{(s)} = \sum_{i=1}^{N} w_i^{(s)} \mathbf{x}_{ijV_t} / \sum_{i=1}^{N} w_i^{(s)}, \; j = 1, \dots, P, \tag{14}$$

are weighted averages of the back parts of the training samples, and

$$\widehat{\mathbf{x}}_j^{(s)} = (\mathbf{x}_{jU_t}^T, \widehat{\mathbf{x}}_{jV_t}^{(s)^T})^T, \; j = 1, \dots, P,$$

are the estimated complete functional predictors. Finally, the quality response can be foresaw as

$$\widehat{y}_s = \sum_{j=1}^{P} \widehat{\mathbf{x}}_j^{(s)^T} \widehat{\theta}_{j\cdot s}. \tag{15}$$

Our above prediction is based on a Supervised Functional Predictor Completion (SFPC) as the regression coefficients $\widehat{\theta}_{jts}$'s which are estimated from the training dataset are utilized as weights in (13) (see Fig. 3). Such a SFPC is also adaptive since the contribution of each training sample, i.e., $w_i^{(s)}$, could be different when predicting quality at different spatial indexes. At each spatial index s, the time complexities in the above contribution calculation (13), predictor completion (14) and quality prediction (15) are $O(NPt)$, $O\big(NP(T-t)\big)$ and $O(PT)$, respectively, so the overall time complexity in predicting a whole product surface with S measurements at each time t is $O(NPTS)$.

IV. Performance Assessment

A. Numerical Simulations

Here we assess the performance of our proposed SSF^{2}R model on various aspects by extensive numerical simulations. In a synthetic dataset, we consider $N = 100$ training samples and $M = 20$ test samples. The model input includes $P = 5$ signal predictors in the time domain, each of which has $T = 15$ discrete time points, and the model output is a surface response in the spatial domain which is measured at $S = 16$ evenly distributed locations in a 4×4 regular grid. Since the signal predictors would be normalized beforehand, without loss of any generality, we generate them from the following multivariate normal distribution:

$$\mathbf{x}_{ij\cdot} \overset{i.i.d.}{\sim} \mathrm{N}(\mathbf{0}, \Sigma), \quad i = 1, \dots, N, \; j = 1, \dots, P,$$

where $\Sigma_{tt'} = 0.3^{|t-t'|}$, $t, t' = 1, \dots, T$, which accounts for the within-signal correlation. Then for each signal predictor we set the true coefficient function as below:

$$\begin{aligned}
&\theta_{1ts} = 1, \text{ for } s = 1, \dots, 16,\\
&\theta_{2ts} = \begin{cases} 0, \text{ for } s = 1,3,4,8,9,13,14,16, \\ -\sin\big((t-1)\pi/6\big) \times \mathrm{I}\{1 \le t < 13\} + 0, \text{ otherwise}, \end{cases}\\
&\theta_{3ts} = \begin{cases} 0, \text{ for } s = 1,2,4,5,9,10,12,13,14,15,16, \\ 0.5(t-1) \times \mathrm{I}\{1 \le t < 3\} + \\ \quad \big(-0.5(t-1)+2\big) \times \mathrm{I}\{3 \le t < 7\} + \\ \quad \big(0.5(t-1)-4\big) \times \mathrm{I}\{7 \le t < 9\} + 0, \text{ otherwise}, \end{cases}\\
&\theta_{4ts} = \begin{cases} 0, \text{ for } s = 1,2,3,4,5,6,7,8,10,11,12,15,16, \\ \big(1-(t-5)^2/16\big) \times \mathrm{I}\{1 \le t < 9\} + 0, \text{ otherwise}, \end{cases}\\
&\theta_{5ts} = 0, \text{ for } s = 1, \dots, 16,
\end{aligned}$$

where $\mathrm{I}\{\cdot\}$ is the indicator function which is equal to one if the involved condition is true and zero otherwise. The informative coefficient functions ($\boldsymbol{\theta}_{j\cdot s} \neq \mathbf{0}$) are depicted in the right panel of Fig. 4 as the black solid lines of constant, sinwave, linear and quadratic forms for different predictors and are associated with the dark indexes in the surface grid in Fig. 4 (left panel), while the noninformative coefficient functions ($\boldsymbol{\theta}_{j\cdot s} = \mathbf{0}$) are the gray solid lines corresponding to the gray indexes therein. The surface responses can finally be generated using (2) with $\mathrm{Var}(\varepsilon) = 0.2^2$. In the surface response, the similarity measure $c(s, s')$ is taken as one if s' is at the left, right, top or bottom of s and zero otherwise.

We apply our ADMM-based optimization algorithm derived in Section III-D to the training samples to test its computational efficiency, where Θ, Z and η are initialized as zero matrices and $\rho = 1$. As clearly shown in Fig 5., the objective function in (12) converges very quickly and only a few iterations (< 20) are required to find the global optimum. Based on the training samples, the five-fold Cross-Validation (CV) is used to determine the tuning parameters from their candidates $\lambda \in \{0.00, 0.25, \dots, 32\}$ and $\gamma, \alpha, \beta \in \{0.0, 0.1, \dots, 1.0\}$.

Now to demonstrate the superiority of our proposed SSF^{2}R model, we first show the coefficient estimation performance of our model compared with the conventional FR model which directly solves (3) and acts as a benchmark. Specifically, we take two spatial indexes $s = 6, 10$ in the surface response for example (see left panel of Fig. 4) and plot their estimated coefficient functions $\widehat{\boldsymbol{\theta}}_{j\cdot s}$, $j = 1, \dots, 4$ in the right panel of Fig. 4. It is obvious that our model (blue lines) performs much better than the benchmark FR model (red lines) as its estimates are much closer to the true ones. It can shrink both the insignificant signals and negligible elements to be exactly zero (see $j = 3, 4$) as a result of TLH-VS, and encourage

1551-3203 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on April 01,2021 at 14:14:43 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2020.3041830, IEEE Transactions on Industrial Informatics

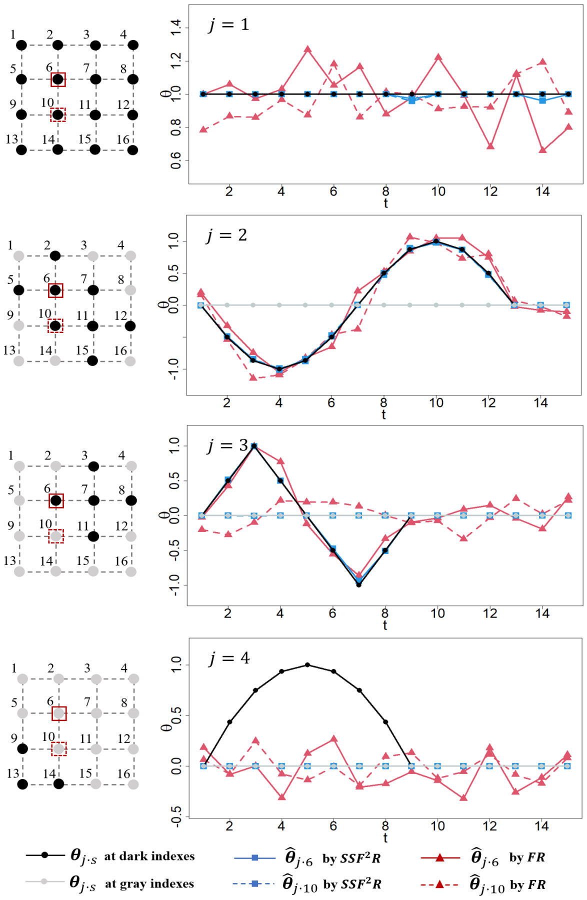


Fig. 4: True and estimated regression coefficient functions ($\theta_{5\cdot s} = \mathbf{0}$ for all s and are omitted here).

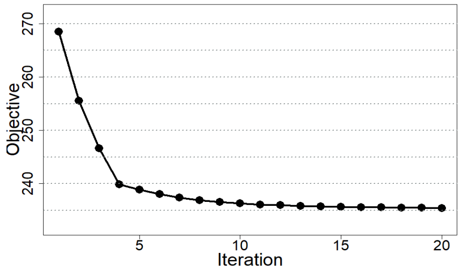


Fig. 5: Convergency analysis of our optimization algorithm.

TABLE I: Summary of competing methods and their AMSE comparison.

Notation	Method	Mean	Std
FR	Functional Regression	0.1662	0.0221
PLS	Partial Least Squares (PLS) regression [4]	0.1434	0.0220
SVR	Support Vector Regression [3]	0.1392	0.0234
NN	Neural Network [5]	0.1211	0.0188
FS	Function-level Sparse FR [21], [22]	0.0718	0.0073
ES	Element-level Sparse FR [20]	0.0725	0.0077
HS	Hierarchical or two-level Sparse FR	0.0693	0.0071
MT	MultiTask learning-based FR	0.0624	0.0070
SM	SMoothness-regularized FR	0.0917	0.0106
ST	STructure-regularized FR	0.0580	0.0054
SS	Our proposed SSF^{2}R model	0.0528	0.0046

the informative coefficient functions at neighbouring spatial indexes to be highly similar (see $j = 1, 2$) due to the MTL.

Next, we verify the prediction advantages of our SSF^{2}R model intensively over many of its counterparts. Besides the above FR model, we consider the Partial Least Squares (PLS) regression [4], SVR [3] and NN [5] as additional benchmarks which are recently used in the semiconductor engineering to predict product surface thickness. Other FR models with various regularization terms are also included. Please note that the comparison between our SSF^{2}R model and the benchmarks is used to certificate the sufficient capability of our method for quality prediction, while the comparison between our model and other regularized FR models is to validate the benefits of our joint use of the sparsity and the structure regularizations. See Table I for the summary of notations of all competing methods. The Mean Square Error (MSE) and the Average MSE (AMSE) are used to evaluate the prediction accuracy regarding a particular test sample and a whole test dataset, respectively:

$$\mathrm{MSE}_i = \frac{1}{S}\sum_{s=1}^{S}\Big(y_{is} - \sum_{j=1}^{P}\sum_{t=1}^{T} x_{ijt}\widehat{\theta}_{jts}\Big)^2, \;\; \mathrm{AMSE} = \frac{1}{M}\sum_{i=1}^{M}\mathrm{MSE}_i.$$

We repeatedly generate 200 synthetic datasets. The boxplots of the resultant 200 AMSEs obtained by different competing methods are exhibited in Fig. 6, and the mean and standard deviation (Std) of these 200 AMSEs are also listed in Table I. Our discoveries are highlighted in the following:

- First of all, our SSF^{2}R model outperforms the benchmark FR, PLS, SVR and NN models as it is equipped with two sophisticatedly tailored regularization terms (sparsity and structure) to address the inherent data complexity in both the functional predictors and the functional response.
- The HS model with a two-level penalty (7) behaves better than the FS with only a function-level penalty (5) and the ES with only an element-level penalty (6), which indicates that our TLH-VS is more powerful when multiple functional predictors are studied.
- The MT model with a structure penalty (8) achieves rather small prediction errors, which justifies the capability of the MTL in predicting a functional response. The addition of a smoothness regularization in the ST model which has a hybrid penalty (10) further improves the prediction.
- Lastly, in conjunction with the benefits of the TLH-VS and the MTL ingredients discussed above, our proposed SSF^{2}R model in (11) enjoys the best generalization performance with the smallest prediction errors.

We also analyze the robustness of our SSF^{2}R model. Note that the informative or nonzero true coefficient functions $\boldsymbol{\theta}_{j\cdot s}$'s are identical for each j in the above simulations. To relax this, we take $j = 2$ as an example, and make the modification:

$$\theta_{2ts} = -(1 + r_s)\sin\big((t-1)\pi/6\big) \times \mathrm{I}\{1 \le t < 13\} + 0,$$
$$\text{for } s = 2, 5, 6, 7, 10, 11, 12, 15,$$

1551-3203 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on April 01,2021 at 14:14:43 UTC from IEEE Xplore. Restrictions apply.

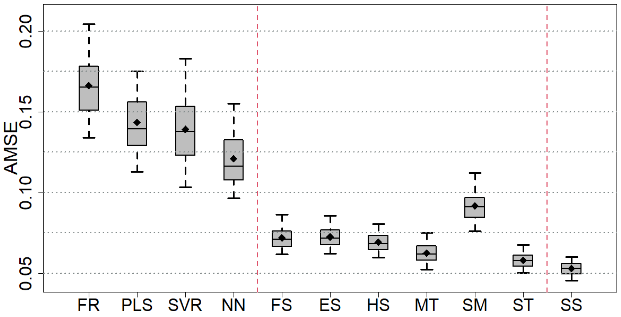


Fig. 6: Boxplots of AMSEs of competing methods.

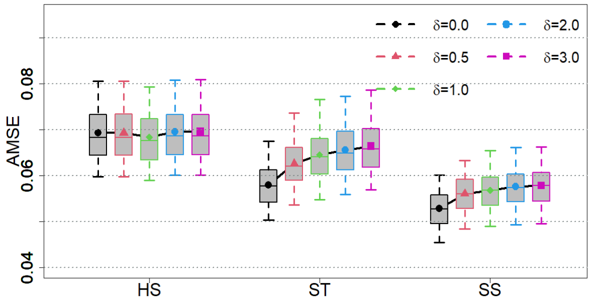


Fig. 7: Boxplots of AMSEs in robustness analysis.

where $r_s \overset{i.i.d.}{\sim} \text{Uniform}(0,\delta)$. When $\delta > 0$, the generated $\theta_{2\cdot s}$'s are only approximate but not the exactly same over s. We consider another four simulation cases with $\delta = 0.5, 1, 2, 3$ and their AMSEs are plotted in Fig. 7. It can be observed that the HS is rarely affected, whereas the ST degenerates since a larger δ would lead to more heterogeneous coefficient functions. Our SS is still the best one in all cases due to its use of both the sparsity and the structure regularizations, and the AMSE curve becomes even flat when δ gets larger. Therefore, our method is robust and can adjust to model coefficient structures of moderate and low degrees of similarity.

Finally, the dynamical prediction by our method is investigated in Fig. 8, where one training dataset with 100 training samples is concerned and the MSEs of 200 test samples are plotted. When $t = 0$ with no predictor observations at all, we just take the mean of the training surfaces as the predicted surface. When $t = 15$, we have the complete signal predictors to perform prediction. When $1 \leq t \leq 14$, we first complete the partial signals using our proposed SFPC, and then proceed the prediction based on the estimated coefficients of the SSF^{2}R model. We also demonstrated the MSEs by an UnSpervised Functional Predictor Completion (USFPC) where the estimated coefficient Θ is not utilized to calculate the training sample contributions in (13). Our dynamical prediction is effective as the MSEs when $1 \leq t \leq 14$ are smaller than those obtained by a blind guess at $t = 0$. Additionally, the proposed SFPC is better than the USFPC, which verifies the advantages of our training sample discrimination via the use of estimation information in the completion of signal predictors.

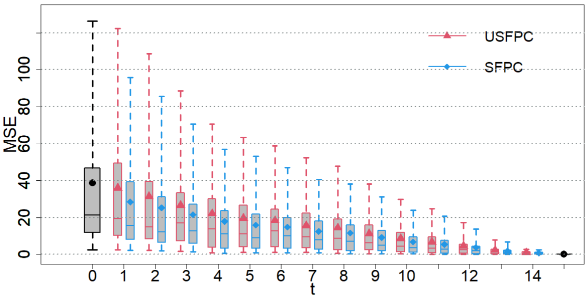


Fig. 8: Boxplots of MSEs in dynamical prediction.

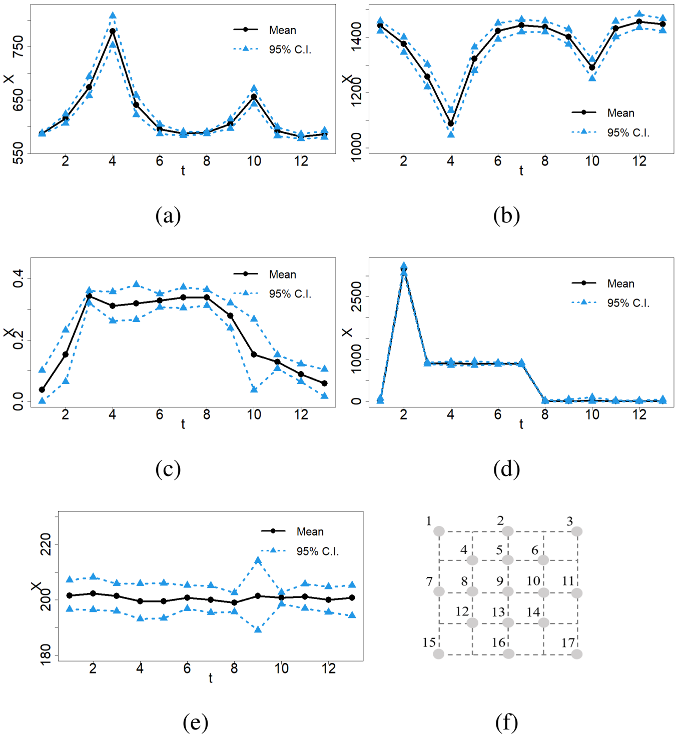
B. Real Case Study

We now revisit the PVD process example as introduced in Section I. This real dataset includes 117 samples, from which 100 are randomly drawn for training and the left ones for test. We have five process variables, which are often selected as the chamber-related predictors for the film thickness prediction [3], [5]. Each variable is collected and stored every second by the deployed sensor node as the PVD process proceeds, and finally forms a signal signature of 15 time points. The means and 95% point-wise Confidence Intervals (C.I.) of these signal signatures are plotted in Fig. 9(a)-(e). The 17 measurement locations in the electronic panel are scaled into a $[0,1] \times [0,1]$ square for simplicity as shown in Fig. 9(f), and the film thicknesses are measured at these locations by a touch-probe coordinate machine. Based on Fig. 9(f), we calculate the similarity measure as $c(s,s') = \exp(-0.2 \times \text{dist}(s,s'))$ where $\text{dist}(s,s')$ is the Euclidean distance between two spatial indexes s and s'. We further eliminate $c(s,s')$'s that are less than 0.05.

Our proposed SSF^{2}R model in (11) takes the above $P = 5$ process variables, each including $T = 15$ time points, as input and the surface of an electronic panel with $S = 17$ thickness measurements as output. The tuning parameters are decided as $\lambda = 5.0$, $\gamma = 0.6$, $\alpha = 0.8$ and $\beta = 0.8$ by performing the five-fold CV among the training samples, which implies that the optimal SSF^{2}R model leans on both the sparsity and the structure regularization almost equally ($\gamma \approx 0.5$), but it puts more weights on the group LASSO penalty ($\alpha > 0.5$) and the fused LASSO penalty ($\beta > 0.5$) than the LASSO penalty and the smoothness penalty, respectively.

The AMSEs of all considered methods for the test samples are listed in Table II, and the pair-wise Student's t test between our SS and its counterparts validates the significantly superior prediction performance of our model. At each spatial index s in the surface response, we also calculate the Square Errors (SEs) of our prediction at this spatial index for the 17 test samples, and present the SE boxplots along these spatial indexes in Fig. 10. We find that there is no discernible difference between the predictions in the center and boundary spatial indexes, so our model works well in the boundary regions of the electronic panel. The dynamical prediction is also conducted in Fig. 11, where our method is again effective with lower

1551-3203 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on April 01,2021 at 14:14:43 UTC from IEEE Xplore. Restrictions apply.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 9: (a) Voltage, (b) Current, (c) Pressure, (d) Air flow rate, (e) Temperature and (f) Measurement grid in the surface response.

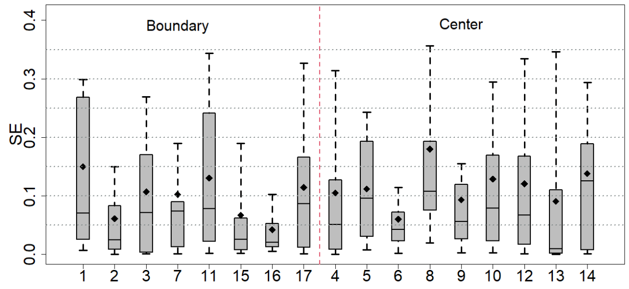


Fig. 10: Boxplots of SEs over spatial indexes in surface response.

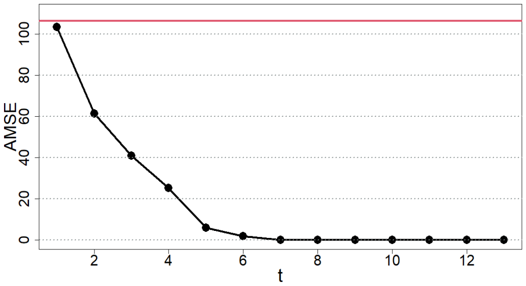
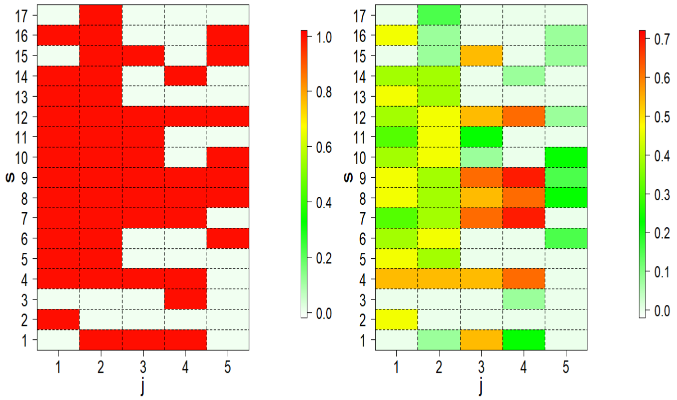


Fig. 11: AMSEs in dynamical prediction.



(a)

(b)

Fig. 12: (a) VS in the function level and (b) VS in the element level.

TABLE II: AMSEs of competing methods.

	FR	PLS	SVR	NN	FS	ES
AMSE	0.2230	0.2208	0.1895	0.1446	0.1233	0.1124
p-value	0.0000	0.0000	0.0005	0.0014	0.0044	0.0254
	HS	MT	SM	ST	SS	
AMSE	0.1113	0.1191	0.1541	0.1140	0.1061	
p-value	0.0107	0.0236	0.0000	0.0054	–	

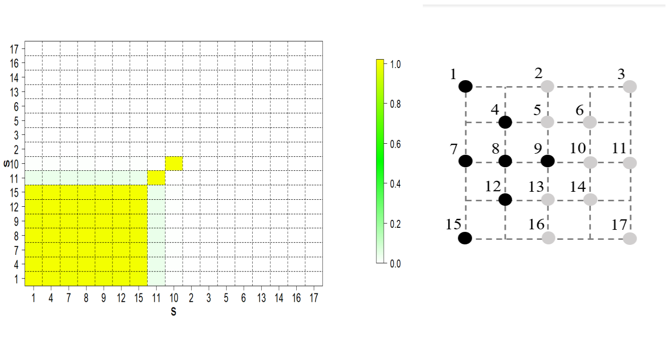
errors than the blind guess (red line), and we suggest that at least a half of the signal signatures ($t > T/2$) should be accumulated so as to obtain a reliable prediction result. Our algorithm is programmed by Python 3.7 software in a personal computer with 1.60-GHz i5-10210U CPUs. For the model parameter estimation from the training samples, it takes about 1.27 hours to search the optimal tuning parameters among their candidates by the five-fold CV. For the dynamical prediction on a test sample, it only consumes 0.0176 seconds to obtain the predicted surface at each time point during the PVD process.

The prediction superiority of our SSF^{2}R model to other regularized ones in Table II evidences the adequacy of the hierarchical sparsity and the similarity structure in the regression coefficient functions. Fig. 12(a) demonstrates whether the estimated coefficient functions $\widehat{\theta}_{j\cdot s}$'s over j and s are selected or not (one or zero), whereas Fig. 12(b) visualizes the proportions (between zero and one) of the significant elements in these coefficient functions $\widehat{\theta}_{j\cdot s}$'s. It can be seen that for each spatial index s, our TLH-VS indeed identifies some important signal predictors at the function level, and further selects only a few significant elements in the informative signals at the element level. For different spatial indexes s's, the VS results could be different. The distinct VS patterns at the boundary indexes, such as 1, 2, 3, 15, 16, 17, are possibly due to the fact that these boundary indexes are adjacent to different modules in the PVD chamber. For example, the spatial indexes 1 and 3 are near the air-in and air-out pipes, and the spatial index 2 is near the power module.

We can also explore the similarity among the obtained informative coefficient functions $\widehat{\theta}_{j\cdot s}$'s associated with each signal predictor. For example, we consider the pressure variable in Fig. 9(c) where $j = 3$. We calculate the Degree of Similarity

1551-3203 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on April 01,2021 at 14:14:43 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2020.3041830, IEEE Transactions on Industrial Informatics



(a) (b)

Fig. 13: (a) DoS heatmap and (b) Spatial indexes in a cluster with high DoSs.

(DoS) between any two coefficient functions as

$$\mathrm{DoS}(\widehat{\theta}_{3\cdot s}, \widehat{\theta}_{3\cdot s'}) = 1 - ||\widehat{\theta}_{3\cdot s} - \widehat{\theta}_{3\cdot s'}||_2 / \max\{||\widehat{\theta}_{3\cdot s}||_2, ||\widehat{\theta}_{3\cdot s'}||_2\}$$

which is zero if $\widehat{\theta}_{3\cdot s} = \mathbf{0}$ or $\widehat{\theta}_{3\cdot s'} = \mathbf{0}$. The DoS heatmap is exhibited in Fig. 13(a), where the spatial index s is reordered and $\widehat{\theta}_{3\cdot s}$'s in the left-bottom cluster have pretty high DoSs (> 0.99). They correspond to the left region in the surface response (see Fig. 13(b)), which indicates the pressure variable mainly influences that particular region of the electronic panel.

V. Conclusion

This paper has proposed a novel function-on-function regression model to predict a surface response based on multiple signal predictors. To conclude, our SSF^{2}R model has the following main advantages: (a) it is highly interpretable as it can produce coefficient functions with a hierarchical sparsity pattern and a similarity structure, (b) it enjoys favorable efficiency due to an ADMM-based iterative algorithm with a modest complexity and a parallel fashion to seek the global optimum, (c) it is extendable to the in-situ manufacturing process to perform reliable prediction progressively and (d) it is effective with superior inference and generalization performance for both synthetic and real datasets.

Though originally motivated by a PVD process example, our method is readily applicable to many other applications in genetics, climatology, economics, and chemical industry whenever the predictor and response are both functions. For long signal predictors with too many time points, the piecewise constant basis function in [7] can be used as a preprocessing step to reduce the computational cost of our model. Finally, as the emerging deep learning networks are capable of approximating almost any flexible relationships between predictors and responses, in our future work, we will take them as a basic framework, and design more tailored layers and special architectures to exploit the similarity structure and control the model complexity when both the model input and output are functions as in our context.

References

[1] E. Sisinni, A. Saifullah, S. Han, U. Jennehag, and M. Gidlund, "Industrial internet of things: Challenges, opportunities, and directions," *IEEE Trans. Ind. Informat.*, vol. 14, no. 11, pp. 4724–4734, 2018.

[2] D. Wang, J. Liu, and R. Srinivasan, "Data-driven soft sensor approach for quality prediction in a refining process," *IEEE Trans. Ind. Informat.*, vol. 6, no. 1, pp. 11–17, 2010.

[3] H. Purwins, B. Barak, A. Nagi, R. Engel, U. Hockele, A. Kyek, S. Cherla, B. Lenz, G. Pfeifer, and K. Weinzierl, "Regression methods for virtual metrology of layer thickness in chemical vapor deposition," *IEEE/ASME Trans. Mechatron.*, vol. 19, no. 1, pp. 1–8, 2014.

[4] H.-J. Roh, S. Ryu, Y. Jang, N.-K. Kim, Y. Jin, S. Park, and G.-H. Kim, "Development of the virtual metrology for the nitride thickness in multi-layer plasma-enhanced chemical vapor deposition using plasma-information variables," *IEEE Trans. Semicond. Manuf.*, vol. 31, no. 2, pp. 232–241, 2018.

[5] D. H. Kim, J. E. Choi, T. M. Ha, and S. J. Hong, "Modeling with thin film thickness using machine learning," *J. Semicond. Disp. Technol.*, vol. 18, no. 2, pp. 48–52, 2019.

[6] K. B. Lee and C. O. Kim, "Recurrent feature-incorporated convolutional neural network for virtual metrology of the chemical mechanical planarization process," *J. Intell. Manuf.*, vol. 31, no. 1, pp. 73–86, 2020.

[7] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*. Springer Science & Business Media, 2005.

[8] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*. Springer Science & Business Media, 2001.

[9] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.

[10] F. Tsung, K. Zhang, L. Cheng, and Z. Song, "Statistical transfer learning: A review and some extensions to statistical process control," *Qual. Eng.*, vol. 30, no. 1, pp. 115–128, 2018.

[11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed Optimization and Statistical Learning Via the Alternating Direction Method of Multipliers*. Now Publishers Inc., 2011.

[12] R. Liu, B. Yang, and A. G. Hauptmann, "Simultaneous bearing fault recognition and remaining useful life prediction using joint-loss convolutional neural network," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 87–96, 2020.

[13] J. Long, S. Zhang, and C. Li, "Evolving deep echo state networks for intelligent fault diagnosis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4928–4937, 2020.

[14] J. S. Morris, "Functional regression," *Annu. Rev. Stat. Appl.*, vol. 2, pp. 321–359, 2015.

[15] D. Zhang, X. Lin, and M. R. Sowers, "Two-stage functional mixed models for evaluating the effect of longitudinal covariate profiles on a scalar outcome." *Biom.*, vol. 63, no. 2, pp. 351–362, 2007.

[16] J. Goldsmith, J. Bobb, C. M. Crainiceanu, B. S. Caffo, and D. Reich, "Penalized functional regression." *J. Comput. Graph. Stat.*, vol. 20, no. 4, pp. 830–851, 2011.

[17] A. E. Ivanescu, A.-M. Staicu, F. Scheipl, and S. Greven, "Penalized function-on-function regression," *Comput. Stat.*, vol. 30, no. 2, pp. 539–568, 2015.

[18] H.-G. Muller and F. Yao, "Functional additive models," *J. Am. Stat. Assoc.*, vol. 103, no. 484, pp. 1534–1544, 2008.

[19] X. Sun, P. Du, X. Wang, and P. Ma, "Optimal penalized function-on-function regression under a reproducing kernel hilbert space framework." *J. Am. Stat. Assoc.*, vol. 113, no. 524, pp. 1601–1611, 2018.

[20] Y. Zhao, H. Chen, and R. T. Ogden, "Wavelet-based weighted lasso and screening approaches in functional linear regression," *J. Comput. Graph. Stat.*, vol. 24, no. 3, pp. 655–675, 2015.

[21] X. Qi and R. Luo, "Function-on-function regression with thousands of predictive curves," *J. Multivariate Anal.*, vol. 163, pp. 51–66, 2018.

[22] M. R. Gahrooei, K. Paynabar, M. Pacella, and J. Shi, "Process modeling and prediction with large number of high-dimensional variables using functional regression," *IEEE Trans. Autom. Sci. Eng.*, vol. 17, no. 2, pp. 684–696, 2020.

[23] K. Paynabar, J. Jin, and M. P. Reed, "Informative sensor and feature selection via hierarchical nonnegative garrote," *Technometrics*, vol. 57, no. 4, pp. 514–523, 2015.

[24] C. Hong, J. Yu, J. Zhang, X. Jin, and K.-H. Lee, "Multimodal face-pose estimation with multitask manifold deep learning," *IEEE Trans. Ind. Informat.*, vol. 15, no. 7, pp. 3952–3961, 2019.

[25] C. Sun, M. Ma, Z. Zhao, S. Tian, R. Yan, and X. Chen, "Deep transfer learning based on sparse autoencoder for remaining useful life prediction of tool in manufacturing," *IEEE Trans. Ind. Informat.*, vol. 15, no. 4, pp. 2416–2425, 2019.

[26] S.-T. Tseng, N.-J. Hsu, and Y.-C. Lin, "Joint modeling of laboratory and field data with application to warranty prediction for highly reliable products," *IIE Trans.*, vol. 48, no. 8, pp. 710–719, 2016.

1551-3203 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on April 01,2021 at 14:14:43 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2020.3041830, IEEE Transactions on Industrial Informatics

[27] C. Shao, J. Ren, H. Wang, J. J. Jin, and S. J. Hu, "Improving machined surface shape prediction by integrating multi-task learning with cutting force variation modeling," *J. Manuf. Sci. Eng.*, vol. 139, no. 1, p. 11014, 2017.

[28] J. Zhou, J. Chen, and J. Ye, "Malsar: Multi-task learning via structural regularization," *Arizona State University*, vol. 21, 2011.

[29] S. H. G. Oliveira, A. R. Gonçalves, and F. J. V. Zuben, "Group lasso with asymmetric structure estimation for multi-task learning." in *Proc. of 28th Int. Conf. Artif. Intell.*, 2019, pp. 3202–3208.

[30] J. C. Beer, H. J. Aizenstein, S. J. Anderson, and R. T. Krafty, "Incorporating prior information with fused sparse group lasso: Application to prediction of clinical measures from neuroimages." *Biom.*, vol. 75, no. 4, pp. 1299–1309, 2019.

Kai Wang received the Ph.D. degree in Industrial Engineering and Logistics Management from Hong Kong University of Science and Technology, Hong Kong, in 2018, and the B.S. degree in Industrial Engineering from Xi'an Jiaotong University, Shaanxi, China, in 2014.

He is currently an Assistant Professor with the Department of Industrial Engineering, School of Management in Xi'an Jiaotong University, Xi'an China. His research focuses on statistical process control, industrial big data analytics, statistical machine learning and transfer learning.

Fugee Tsung received the Ph.D. and M.Sc. degrees from the University of Michigan, Ann Arbor, MI, USA., and the B.Sc. degree from National Taiwan University, Taipei, Taiwan.

He is currently a Chair Professor with the Department of Industrial Engineering and Decision Analytics and the Director of the Quality and Data Analytics Laboratory in Hong Kong University of Science and Technology, Hong Kong. His research interests include quality analytics in advanced manufacturing and service processes, industrial big data and statistical process control, monitoring, and diagnosis.

Based on his pioneer contribution to Quality Analytics research and education, Prof. Tsung has been elected Academician of the International Academy for Quality (IAQ), Fellow of the Institute of Industrial Engineers (IIE), Fellow of the American Statistical Association (ASA), Fellow of the American Society for Quality (ASQ) and Elected Member of the International Statistical Institute (ISI).

1551-3203 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Hong Kong University of Science and Technology. Downloaded on April 01,2021 at 14:14:43 UTC from IEEE Xplore. Restrictions apply.