Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: https://www.tandfonline.com/loi/utch20

Real-Time Monitoring of High-Dimensional Functional Data Streams via Spatio-Temporal Smooth Sparse Decomposition

Hao Yan, Kamran Paynabar & Jianjun Shi

To cite this article: Hao Yan, Kamran Paynabar & Jianjun Shi (2018) Real-Time Monitoring of High-Dimensional Functional Data Streams via Spatio-Temporal Smooth Sparse Decomposition, Technometrics, 60:2, 181-197, DOI: 10.1080/00401706.2017.1346522

To link to this article: https://doi.org/10.1080/00401706.2017.1346522

TECHNOMETRICS
2018, VOL. 60, NO. 2, 181–197
https://doi.org/10.1080/00401706.2017.1346522

Taylor & Francis
Taylor & Francis Group

Real-Time Monitoring of High-Dimensional Functional Data Streams via Spatio-Temporal Smooth Sparse Decomposition

Hao Yan, Kamran Paynabar, and Jianjun Shi

Georgia Institute of Technology, Atlanta, GA

ABSTRACT

High-dimensional data monitoring and diagnosis has recently attracted increasing attention among researchers as well as practitioners. However, existing process monitoring methods fail to fully use the information of high-dimensional data streams due to their complex characteristics including the large dimensionality, spatio-temporal correlation structure, and nonstationarity. In this article, we propose a novel process monitoring methodology for high-dimensional data streams including profiles and images that can effectively address foregoing challenges. We introduce spatio-temporal smooth sparse decomposition (ST-SSD), which serves as a dimension reduction and denoising technique by decomposing the original tensor into the functional mean, sparse anomalies, and random noises. ST-SSD is followed by a sequential likelihood ratio test on extracted anomalies for process monitoring. To enable real-time implementation of the proposed methodology, recursive estimation procedures for ST-SSD are developed. ST-SSD also provides useful diagnostics information about the location of change in the functional mean. The proposed methodology is validated through various simulations and real case studies. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received February 2016
Revised May 2017

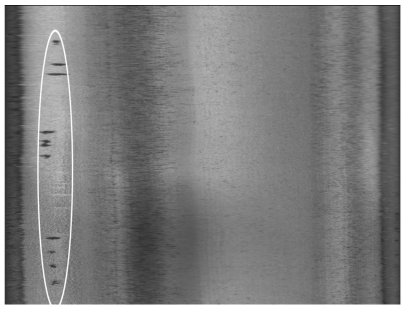
KEYWORDS
Anomaly detection; Functional data; Real-time monitoring; Recursive estimation; Smooth sparse decomposition; Spatio-temporal data

1. Introduction

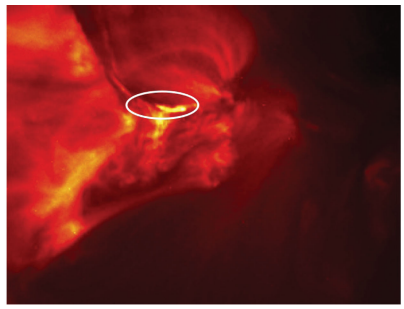
Nowadays most manufacturing processes are instrumented with sensing systems comprised of hundreds of sensors to monitor process performance and product quality. The low implementation cost, high acquisition rate, and high variety of such sensing systems lead to rich data streams that provide distinctive opportunities for performance improvement. Real-time process monitoring and control, accurate fault diagnosis, and online product inspection are among the benefits that can be gained from effective modeling and analysis of streaming data. However, the complex characteristics of these data streams pose significant analytical challenges yet to be addressed. Common characteristics of these data streams include (1) *High variety*: Various types of sensors generate a high variety of data streams, including profiles or waveform signals (e.g., an exerted force profile during a forging operation, Lei, Zhang, and Jin 2010), images (e.g., an image of a bar surface after rolling, Jin, Zhou, and Chang 2004), and videos (e.g., a video of an industrial flame in steel tube manufacturing, Yan, Paynabar, and Shi 2015b); (2) *High dimensionality*: A typical image used for surface inspection is on the order of 1M pixels (Jin, Zhou, and Chang 2004); (3) *High velocity*: In recent years, the speed of data collection has significantly increased so that it can keep up with almost any production rate. For example, a commercially available ultrasonic sensor can easily record data at the rate of 1 KHz, and a high-speed industrial camera is capable of scanning a product surface with the rate of 80 million pixels per second or faster (Jin, Zhou, and Chang 2004); (4) *Spatial and temporal structure*: Another layer of complexity arises because of the spatio-temporal structure of streaming data. Data points in a profile or pixels within an image are spatially correlated (e.g., neighbor pixels often exhibit high correlations) and corresponding data points or pixels across sequential samples are often temporally correlated with nonstationary behavior.

Examples of such high-dimensional (HD) data are shown in Figure 1. In Figure 1(a), a sample of a bar surface used for monitoring of a rolling process is shown (Jin, Zhou, and Chang 2004). In the second example, shown in Figure 1(b), a sequence of solar images captured by a satellite is used to monitor solar activities and detect solar flares. The streams of solar and rolling images can be seen in two video clips in the online appendix. Figure 1(c) shows 20 normal and 20 faulty multi-channel tonnage profiles used for monitoring a forging process (Lei, Zhang, and Jin 2010). As can be seen from the figures and clips, an in-control HD data stream can typically be represented by a functional mean with a smooth spatial structure that gradually changes over time. However, this gradual change manifests inherent dynamics of the process and should not be considered as an out-of-control situation. Anomalies, on the other hand, are in the form of abrupt changes with a spatio-temporal structure different from the functional mean. The smooth temporal change of the functional mean may significantly increase the false alarm rate of a monitoring procedure if not appropriately modeled. This makes monitoring of HD data streams even more challenging. Most

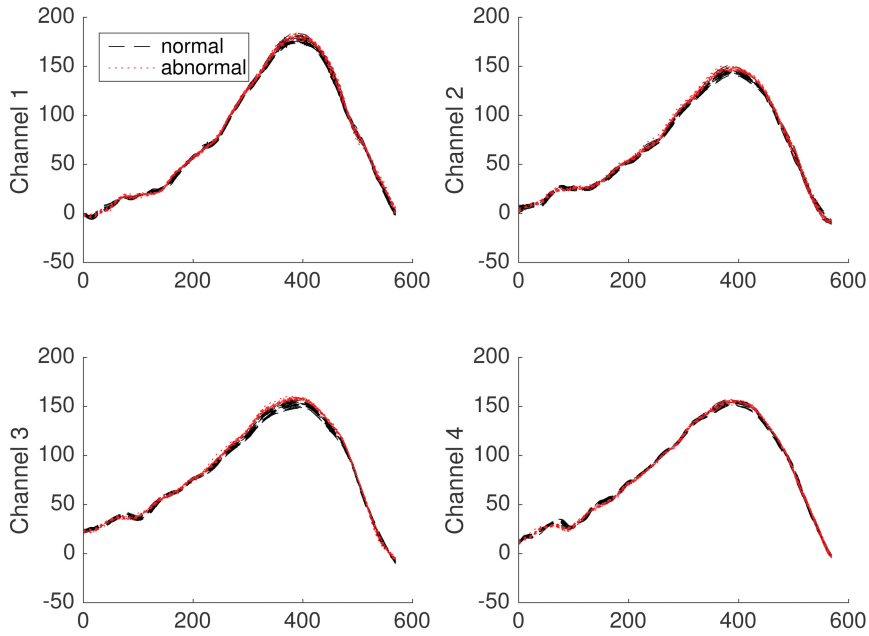
CONTACT Kamran Paynabar kamran.paynabar@isye.gatech.edu Georgia Institute of Technology, Atlanta, GA, 30332.
Color versions of one or more of the figures in this article are available online at www.tandfonline.com/r/TECH.
Supplementary materials for this article are available online. Please go to http://www.tandfonline.com/r/TECH.



(a) Surface image of a steel bar



(b) Temperature map of sun surface



(c) Tonnage signals in a forging process

Figure 1. Example of HD streaming data with anomalies.

of existing HD monitoring methods fail to model the temporal trend of the functional mean, and only focus on change detection by assuming that the in-control functional mean is constant over time.

To address the aforementioned challenges, this article develops a new scalable spatio-temporal methodology for real-time monitoring and diagnosis of HD high-velocity streaming functional data with time-varying means. This methodology is also capable of identifying the location of the change, which is important for diagnosis. Our proposed methodology is inspired by the recent development of smooth-sparse decomposition (SSD) for anomaly detection in images (Yan, Paynabar, and Shi 2015a). SSD can separate anomalies from the image background by using the spatial structure of an image. The key idea is to extend the SSD methodology so that it can incorporate temporal information of an HD data stream in addition to the spatial information of a single sample. However, this extension is nontrivial because adding the time dimension significantly increases the dimensionality of the problem, given the high rate data acquisition. In this article, we begin with extending the SSD method to spatio-temporal SSD so it can include temporal information and model smooth temporal trend of a data stream. Assuming that the functional mean of the data stream is spatially and temporally smooth and process changes/anomalies are nonsmooth and sparse in a certain basis representation, our proposed spatio-temporal SSD decomposes an HD data stream into a smooth spatio-temporal functional mean, sparse anomalous features, and random noises. This model serves as a dimension reduction technique, which reduces the HD data stream to a small set of features. We then develop recursive estimation procedures that significantly reduce the computational complexity and enable the real-time implementation of the method. Finally, we combine the proposed model with a likelihood-ratio test (LRT) to monitor the process based on the detected anomalies/features.

The remainder of the article is organized as follows. Section 2 reviews the relevant literature. Section 3 elaborates the proposed spatio-temporal SSD for HD functional data streams. In Section 4, reproducing kernel and roughness penalization are proposed for temporal modeling and developing recursive estimation procedures for real-time analysis. In Section 5, monitoring and diagnosis methods are proposed by combining

LRT with spatio-temporal SSD. To evaluate and compare the proposed methodology with existing methods, simulated data based on thermodynamic principles of heat transfer are used in Section 6. In Section 7, we illustrate how our proposed method can be used in real world using three case studies including monitoring of a rolling process, detection of solar flares, and monitoring of a forging process. We conclude the article and discuss future research directions in Section 8.

2. Literature Review

There is a considerable body of literature on monitoring and diagnosis of HD streaming data. Current research in this area can be classified into three groups: monitoring methods for HD multivariate data streams, profile monitoring techniques, and monitoring methods based on dimension reduction. In the first group, HD data are treated as multiple univariate data streams. For example, a profile stream with a length of 200 generates 200 individual data streams. Under the assumption that data streams are independent, Mei (2010) proposed a monitoring scheme based on the sum of the local CUSUM statistics for individual streams. Liu, Mei, and Shi (2014) extended this method and developed an adaptive sensing scheme assuming that only partial observations are available. Zou et al. (2015) developed a powerful goodness-of-fit test for monitoring independent HD data streams. However, these methods assume that the data streams are independent and therefore, ignore their temporal and spatial structures. To monitor univariate data streams with a temporal trend, Qiu and Xiang (2014) and Xiang, Qiu, and Pu (2013) combined nonparametric regression with longitudinal modeling techniques. However, they did not consider the spatial structure of the functional mean and anomalies. The literature on nonlinear profile monitoring is rich, which includes various parametric and nonparametric methods. For example, for monitoring smooth profiles, there are various nonparametric methods based on local kernel regression (Zou, Tsung, and Wang 2008; Zou, Qiu, and Hawkins 2009; Qiu, Zou, and Wang 2010) and splines (Chang and Yadama 2010). Paynabar and Jin (2011) used wavelets to model and monitor nonsmooth profiles. These methods, however, are not applicable to profiles with time-varying means. Moreover, most of these methods are specifically designed for profile motioning, and their generalization to image and video streams is nontrivial. Among the dimension reduction approaches, principal component analysis (PCA) is the most popular method for HD data monitoring because of its simplicity, scalability, and data compression capability. For example, Liu (1995) used PCA to reduce the dimensionality of streaming data and constructed T^2 and Q charts to monitor extracted features and residuals, respectively. Paynabar, Qiu, and Zou (2015) proposed a monitoring approach for multichannel signals by combining multivariate functional PCA and change-point models. Yan, Paynabar, and Shi (2015b) developed a tensor-based principal component analysis that can model both the spatial and spectral structures of an image sequence. Bakshi (1998) proposed a multi-resolution PCA for profile monitoring by integrating PCA with wavelets. The main drawback of PCA-based methods is that they cannot be directly used for nonstationary data streams with a time-varying mean. To address the drawbacks of existing methods, we propose a new spatio-temporal smooth sparse decomposition for monitoring and diagnosis of HD data streams.

3. Spatio-Temporal Smooth Sparse Decomposition

In this section, we develop the spatio-temporal model by extending SSD so it can model the temporal trend in addition to the spatial structure of functional data streams. We also propose efficient algorithms for fast implementation of spatio-temporal SSD (ST-SSD) for a given data sample. For simplicity, we begin with profile data (i.e., 1D functional data). Suppose a sequence of profiles $y_t; t = 1, \ldots, n$ is available where y_t is a profile of size $p \times 1$ recorded at time t. We combine all profiles into a matrix $Y = (y_1, y_2, \ldots, y_n)$ of size $p \times n$ and define $y = \text{vec}(Y)$ as the vectorized matrix (i.e., y is a $pn \times 1$ vector). Following Yan, Paynabar, and Shi (2015a), we aim to decompose y into three components: A functional mean μ, anomalies a, and noises e as $y = \mu + a + e$, where $a = \text{vec}(a_1, \ldots, a_n)$ and $e = \text{vec}(e_1, \ldots, e_n)$ with a_t and e_t as anomaly features and noise in y_t. We assume that the dynamic functional mean μ has a smooth spatio-temporal structure and a is sparse or can be sparsely represented by a certain basis. To model both spatial and temporal structures and at the same time reduce data dimensions, we define B_s and B_t as smooth spatial and temporal bases for the mean, and B_{as} and B_{at} as spatial and temporal bases for anomalies, respectively. The spatio-temporal bases for the mean and anomalies are obtained by the tensor product of these bases, that is, $B = B_t \otimes B_s$ and $B_a = B_{at} \otimes B_{as}$. Consequently, the functional mean and anomalies are modeled as $\mu = (B_t \otimes B_s)\theta$ and $a = (B_{at} \otimes B_{as})\theta_a$ resulting in $y = (B_t \otimes B_s)\theta + (B_{at} \otimes B_{as})\theta_a + e$, where $\theta = \text{vec}(\theta_1, \theta_2, \ldots, \theta_n)$ and $\theta_a = \text{vec}(\theta_{a,1}, \ldots, \theta_{a,n})$, and θ_t and $\theta_{a,t}$ are the spatio-temporal coefficients of the functional mean and anomalies at time t, correspondingly. We assume that noise components are normally independently distributed, that is, $e \sim \text{NID}(0, \sigma^2)$. To estimate θ and θ_a, we propose a penalized regression model, called spatio-temporal smooth sparse decomposition (ST-SSD), as follows:

$$\underset{\theta,\theta_a}{\text{argmin}} \|e\|^2 + \theta^T R\theta + \gamma\|\theta_a\|_1 \text{ s.t. } y = (B_t \otimes B_s)\theta + (B_{at} \otimes B_{as})\theta_a + e, \quad (1)$$

where $\|\cdot\|$ and $\|\cdot\|_1$ are L_2 and L_1 norm operators, and γ is a tuning parameter to be determined by the user. The Matrix R is the regularization matrix that controls the smoothness of the mean function, and the L_1 penalty term, $\gamma|\theta_a|_1$, encourages the sparsity of the anomalous regions. In this article, inspired by Xiao, Li, and Ruppert (2013), we define the regularization matrix R as $R = R_t \otimes B_s^T B_s + B_t^T B_t \otimes R_s + R_t \otimes R_s$, where R_s and R_t are the regularization matrices that control the smoothness in the spatial and temporal directions. For tensors with smooth structure, it has shown in Xiao, Li, and Ruppert (2013) and Yan, Paynabar, and Shi (2015a) that the penalty term defined with this tensor structure is able to achieve high precision with small computational time and asymptotically achieve the optimal rate of convergence under some mild conditions. The spatial regularization matrix R_s can be defined as $R_s = \lambda_s D_s^T D_s$ (Ruppert 2012), where D_s is the first-order difference matrix since the smoothness of the function is directly related to the difference between the neighbor coefficients. That is, $D_s = [d_{pq}] = 1_{q=p} -$

$1_{q=p+1}$, with 1_A as an indicator function, that is, it is 1 when A is true, and 0 otherwise. λ_s is the tuning parameter controlling the spatial smoothness of the functional mean. The choice of R_t depends on the temporal model and will be discussed in Section 4. It was shown by Yan, Paynabar, and Shi (2015a) that if θ_a is given, $\mu = B\theta$ can be solved by $\mu = H(y - B_a\theta_a)$, where $H = B(B^TB + R)^{-1}B^T$ is the projection matrix. They also showed that (1) is equivalent to a weighted lasso formulation, that is, $\min_{\theta_a}(y - B_a\theta_a)^T(I - H)(y - B_a\theta_a) + \gamma\|\theta_a\|_1$, thus can be efficiently solved by the APG algorithm. The reason for defining the regularization matrix in the foregoing form is that under this definition of R, the projection matrix of ST-SSD, denoted by H, can be further decomposed by the tensor product of two spatial and temporal projection matrices, that is, $H = H_t \otimes H_s$, where $H_s = B_s(B_s^TB_s + R_s)^{-1}B_s^T$ and $H_t = B_t(B_t^TB_t + R_t)^{-1}B_t^T$, as shown in Appendix A. This will help significantly reduce the computational complexity of the optimization algorithm for solving Equation (1). Equation (1) is a convex optimization problem that can be solved via a general convex solver such as the interior point method. However, the interior point method is slow and cannot be used in HD settings. Therefore, similar to Yan, Paynabar, and Shi (2015a), the accelerated proximal gradient (APG) algorithm is used to solve (1) iteratively, as given in Algorithm 1.

Algorithm 1. Optimization algorithm for solving SSD

initialize

$L = 2\|B_{as}\|_2^2$, $x^{(0)} = 0$, $\theta_a^{(0)} = 0$, $t_0 = 1$

end

Compute

$$H_s = B_s\left(B_s^TB_s + R_s\right)^{-1}B_s^T$$
$$H_t = B_t\left(B_t^TB_t + R_t\right)^{-1}B_t^T \tag{2}$$

for $k = 1, 2, \cdots$ **do**

Update

$$a^{(k-1)} = (B_{at} \otimes B_{as})x^{(k-1)}$$
$$\mu^{(k-1)} = (H_t \otimes H_s)(y - a^{(k-1)}) \tag{3}$$

$$\theta_a^{(k)} = S_{\frac{\gamma}{L}}\left(x^{(k-1)} + \frac{2}{L}\left(B_{at}^T \otimes B_{as}^T\right)(y - a^{(k-1)} - \mu^{(k-1)})\right) \tag{4}$$

$$t_k = \frac{1 + \sqrt{1 + 4t_{k-1}^2}}{2}$$

$$x^{(k)} = \theta_a^{(k)} + \frac{t_{k-1} - 1}{t_k}\left(\theta_a^{(k)} - \theta_a^{(k-1)}\right)$$

If $|\theta_a^{(k)} - \theta_a^{(k-1)}| < \epsilon$ **then**

Stop

end

end

In Algorithm 1, $S_\gamma(x) = \text{sgn}(x)(|x| - \gamma)_+$ is a soft-thresholding operator, in which $\text{sgn}(x)$ is the sign function and $x_+ = \max(x, 0)$. The θ is not explicitly update since it is updated with μ as $\mu^{(k)} = B\theta^{(k)}$. Note that the convergence of Algorithm 1 is guaranteed and can be proved similarly as shown in Yan, Paynabar, and Shi (2015a).

To generalize the ST-SSD model to l-dimensional data (e.g., $l = 2$ for images or multichannel signals), we represent a single sample by tensor $\mathcal{Y}$ of size $p_1 \times \cdots \times p_l$. For computational efficiency, the spatial basis B_s and B_{as} are defined as the tensor product of multiple 1D bases, that is, $B_s = \otimes_{i=1}^l B_{si}$ and $B_{as} = \otimes_{i=1}^l B_{asi}$, where $\otimes_{i=1}^l B_{si} := B_{s1} \otimes \cdots \otimes B_{sl}$. It is shown in Appendix A that if we set $R_s = \otimes_{i=1}^l (B_{si}^TB_{si} + R_{si}) - B_s^TB_s$, the projection matrix becomes decomposable, that is, $H_s = \otimes_{i=1}^l H_{si}$ with $H_{si} = B_{si}(B_{si}^TB_{si} + R_{si})^{-1}B_{si}^T$. For example, if a B-spline basis is used, R_{si} can be defined as $R_{si} = \lambda_{si}D_i^TD_i$. Furthermore, to increase the computational efficiency of the optimization algorithm, we use the well-known relationship between the Kronecker and tensor products to compute $y = (\otimes_{i=1}^l B_{si})x$ by $\mathcal{Y} = \mathcal{X} \times_{i=1}^l B_{si} := \mathcal{X} \times_1 B_{s1} \times_2 B_{s2} \cdots \times_l B_{sl}$, in which $\mathcal{X} \times_n B_{sn}$ is the n-mode tensor product defined by $(\mathcal{X} \times_n B_{sn})(i_1, \ldots, i_l) = \sum_{j_n} \mathcal{X}(i_1, \ldots, j_n, \ldots, i_l)B_{sn}(i_n, j_n)$. A summary of the optimization algorithm for solving the generalized ST-SSD problem is given in Algorithm 2. In this algorithm, since the matrix inversion can be performed in each dimension separately, that is, $B_{si}^TB_{si} + R_{si}; i = 1, \ldots, l$, the total complexity of the matrix inversion is reduced from $O(n^3\prod_i k_i^3)$ to $O(n^3 + \sum_i k_i^3)$, assuming B_{si} is of size $p_i \times k_i$.

Algorithm 2. Optimization algorithm for solving SSD based on APG

initialize

$$\Theta_a^{(0)} = 0, \quad \mathcal{X}^{(0)} = 0, \quad t_0 = 1$$
$$L = 2\prod_i \|B_{ai}\|_2^2$$
$$H_{si} = B_{si}\left(B_{si}^TB_{si} + R_{si}\right)^{-1}B_{si}^T, \quad i = 1, \ldots, k$$
$$H_t = B_t\left(B_t^TB_t + R_t\right)^{-1}B_t^T$$

end

for $k = 1, 2, \cdots$ **do**

Update $\mathcal{A}^{(k-1)} = \mathcal{X}^{(k-1)} \times_{i=1}^k B_{si} \times_t B_{st}$

$$\mathcal{M}^{(k)} = (\mathcal{Y} - \mathcal{A}^{(k-1)}) \times_{i=1}^k H_{si} \times_t H_t$$
$$\Theta_a^{(k)} = S_{\frac{\gamma}{L}}\left(\mathcal{X}^{(k-1)} + \frac{2}{L}(\mathcal{Y} - \mathcal{A}^{(k-1)} - \mathcal{M}^{(k-1)}) \times_{i=1}^k B_{si}^T \times_t B_{st}^T\right)$$
$$t_k = \frac{1 + \sqrt{1 + 4t_{k-1}^2}}{2}$$
$$\mathcal{X}^{(k)} = \Theta_a^{(k)} + \frac{t_{k-1} - 1}{t_k}\left(\Theta_a^{(k)} - \Theta_a^{(k-1)}\right)$$

if $|\Theta_a^{(k-1)} - \Theta_a^{(k)}| < \epsilon$ **then**

Stop

end

end

Selection of an appropriate basis for the functional mean and anomaly is important to model the spatio-temporal structure

of a data stream. Therefore, due to its computational efficiency and flexibility, the B-Spline basis is commonly used for modeling nonlinear smooth functions. In this article, for the spatial basis, we assume that the functional mean is smooth and can be modeled with B-spline basis. Selecting a basis for anomalous regions depends on the type of anomalies. For example, if anomalies are randomly scattered over the mean, it is recommended to use an identity basis, that is, $B_{as} = I$. If anomalies form clustered regions, a spline basis can be a better choice. More details about the spatial basis selection of the functional mean and anomalies are given in Yan, Paynabar, and Shi (2015a). We also assume that anomalies appear abruptly, and hence they do not have a specific temporal structure. Therefore, we use the identity matrix as the temporal basis for anomalies, that is, $B_{at} = I$. In the following section, we will discuss the choice of temporal basis for the functional mean and the recursive estimation of ST-SSD.

4. ST-SSD For Streaming Data And Recursive Estimation

The proposed ST-SSD can effectively model both the temporal and spatial structure of HD data streams. However, the estimation method given in Algorithm 2 is only efficient for a given data stream with a fixed number of observations, n. In the context of statistical process control (SPC), process monitoring includes two stages known as Phase I and Phase II. Since the functional mean is unknown in the beginning, we use n in-control (IC) observations in Phase I to learn the distribution of the monitoring statistic and the control limit. The baseline control chart estimated in Phase I can then be used for real-time and online monitoring in Phase II. Therefore, the proposed method can be used to conduct Phase I analysis offline on n observations collected offline. However, for online (phase II) analysis of HD data where streaming samples are being recorded in short sampling intervals, Algorithm 2 with the complexity of $O(n^3 + \sum_i k_i^3)$ loses its efficiency over time as n grows linearly by time. Specifically, when a new sample is recorded at time t, the length of y increases by the dimensions of the recorded data. Consequently, after some time, the dimensions of Problem (1) become so large that it cannot be solved by any optimization algorithms. To address this issue, the key idea is to develop a recursive estimation procedure that only requires the previous estimations and current data to solve the optimization problem. This recursive algorithm significantly reduces the computation time and required memory, which enables real-time implementation of the method. For this purpose, we use special temporal bases for the functional mean, B_t, and penalization term, R_t. In the following subsections, we propose two temporal models based on reproducing kernels and roughness minimization and present a recursive estimator for each model.

4.1 Reproducing Kernels

Reproducing kernel Hilbert space (RKHS) is a functional space widely used for modeling smooth functional forms using kernels (Berlinet and Thomas-Agnan 2011). From the representer theorem (Schölkopf, Herbrich, and Smola 2001), it is known that any function in an RKHS can be written as a linear combination of kernel functions evaluated at time t. Hence, the gram matrix K_t, defined as $(K_t)_{ij} = \kappa(i, j)$ $(i, j = 1, \ldots, t)$, can be used as the temporal basis (i.e., $B_t = K_t$) in (1), where $\kappa(i, j)$ is the kernel function. In this article, we use the Gaussian kernel to model the smooth temporal structure defined as $\kappa(i, j) = \exp(-\frac{(i-j)^2}{2c^2})$ (Babaud et al. 1986), where c is the bandwidth of the Gaussian kernel). To control the smoothness of the temporal trend, we use Hilbert norm penalization (Schölkopf, Herbrich, and Smola 2001), which is equivalent to defining $R_t = \lambda_t K_t$ in Equation (1). λ_t is the tuning parameter controlling the temporal smoothness of the functional mean. Consequently, the projection matrix H_t can be computed by

$$H_t = K_t\left(K_t^2 + \lambda_t K_t\right)^{-1} K_t = K_t K_{\lambda_t,t}, \tag{5}$$

where $K_{\lambda_t,t} = (K_t + \lambda_t I_t)^{-1}$. However, since computing (5) requires inversion of $K_t + \lambda_t I_t$, which is a $t \times t$ matrix, the total complexity is $O(t^3)$ at time t. Eventually, computing (5) is not feasible due to the increasing number of observations and the limited computational resources. To reduce the computational complexity, we propose to solve the estimation recursively with only recent w observations since earlier observations typically have little impact on the current estimation. We define $K_t = \kappa(i, j) \quad i, j = t - w + 1, \ldots, t$ and $\tilde{K}_t = \kappa(i, j) \quad i, j = t - w, \ldots, t$ as windowed kernel functions, and define $K_{\lambda_t,t} = (K_t + \lambda_t I)^{-1}$ and $\tilde{K}_{\lambda_t,t} = (\tilde{K}_t + \lambda_t I)^{-1}$, accordingly. Proposition 1 shows that H_t and $K_{\lambda_t,t}$ can be computed recursively.

Proposition 1. The following update rules hold for H_t and $K_{\lambda_t,t}$

$$\tilde{H}_t = \begin{bmatrix} \tilde{H}_{t-1} - k_{t-1} r_{t-1}^T g_{t-1}(I_{t-1} - \tilde{H}_{t-1}) & (I_{t-1} - \tilde{H}_{t-1})k_{t-1} g_{t-1} \\ r_{t-1}^T(I_{t-1} + k_{t-1} r_{t-1} g_{t-1} - g_{t-1}) & (1 - r_{t-1}^T k_{t-1}) g_{t-1} \end{bmatrix}$$

$$\tilde{K}_{\lambda_t,t} = \begin{bmatrix} K_{\lambda,t-1} + r_{t-1} r_{t-1}^T g_{t-1} & -r_{t-1} g_{t-1} \\ -r_{t-1}^T g_{t-1} & g_{t-1} \end{bmatrix}, \tag{6}$$

where $r_t = \tilde{K}_{\lambda_t,t} k_t$, $H_t = \tilde{H}_t(2:t, 2:t)$, $K_{\lambda_t,t} = \tilde{K}_{\lambda_t,t}(2:t, 2:t)$ $k_t = [\kappa(t - w + 1, t), \ldots, \kappa(t - 1, t)]^T$, $g_{t-1} = (1 + \lambda_t - r_{t-1}^T k_{t-1})^{-1}$.

$\tilde{K}_{\lambda_t,t}(2:t, 2:t)$ denotes the reduced matrix $\tilde{K}_{\lambda_t,t}$ after removing the first row and column of the matrix. The proof of Proposition 1 is given in Appendix B. With this recursive updating rule, it is not hard to show that the total complexity of Algorithm 1 will reduce to $O(w^2)$ at each sampling time t, which is more efficient compared to the nonrecursive case with $O(t^3)$. The fact that the complexity does not grow with the rate of $O(t^3)$ enables the real-time implementation of ST-SSD for online monitoring of HD streaming data. Finally, the optimization (estimation) algorithm can be updated by replacing the computation of the projection matrix H_t in Algorithm 1 with the updating procedure in (6).

4.2 Roughness Minimization

In this section, we propose an alternative approach for temporal modeling that can achieve even faster computational speed than reproducing kernels. In cases where the functional mean is less volatile over time, we suggest a simple temporal basis and roughness matrix, namely, $B_t = I_t$ and $R_t = D_t^T D_t$ in (1), in which D_t is the first-order difference matrix of size $(t - 1) \times t$

defined as $D_t = [d_{pq}] = 1_{q=p} - 1_{q=p+1}$. By choosing R_t to be $D_t^T D_t$, the temporal penalization term, $\theta^T R_t \theta = \theta^T D_t^T D_t \theta = \sum_{i=2}^{t} \|\theta_i - \theta_{i-1}\|^2$, becomes roughness penalization that penalizes the first-order difference of θ_t for a smoother estimation over time. Therefore, the temporal projection matrix is given by

$$H_t = \left(I_t + \lambda_t D_t^T D_t\right)^{-1}. \tag{7}$$

The next step is to design a recursive estimator for the roughness minimization model. As mentioned earlier, for a system with a gradual temporal trend, it is often true that recent observations have more impact and therefore are more important for parameter estimation and updating. Therefore, an approximate, yet accurate, approach is to estimate only the most recent coefficient θ_t without changing the previous estimations of $\theta_1, \ldots, \theta_{t-1}$. This is equivalent to solve (1) for only θ_t and $\theta_{a,t}$. In this way, the ST-SSD model in (1) can be reduced to the following model, which only requires the estimation of θ_t and $\theta_{a,t}$ at time t:

$$\begin{aligned} &\underset{\theta_t, \theta_{a,t}}{\operatorname{argmin}} \|e\|^2 + \theta^T R \theta + \gamma \|\theta_a\|_1, \\ &\text{subject to } y_t = B_s \theta_t + B_{as} \theta_{a,t} + e_t, \end{aligned} \tag{8}$$

where $R = I_t \otimes R_s + \lambda_t D_t^T D_t \otimes B_s^T B_s + \lambda_t D_t^T D_t \otimes R_s$. As shown in Proposition 2, given $\theta_{a,t}$ and previous estimates, θ_t has a closed-form solution.

Proposition 2. Suppose the previous estimation $\hat{\theta}_1, \ldots, \hat{\theta}_{t-1}$, $\hat{\theta}_{a,1}, \ldots, \hat{\theta}_{a,t-1}$, and $\hat{\theta}_{a,t}$ are known, then the solution of θ_t (or equivalently $\mu_t = B_s \theta_t$) to (8) is given by

$$\hat{\mu}_t = B_s \hat{\theta}_t = (1 - \tilde{\lambda}_t)\hat{\mu}_{t-1} + \tilde{\lambda}_t H_s (y_t - \hat{a}_t), \tag{9}$$

where $\tilde{\lambda}_t = \frac{1}{1+\lambda_t}$ and $\hat{a}_t = B_{at} \hat{\theta}_{a,t}$.

The proof is shown in Appendix C. Note that in (9), the temporal structure of μ_t is modeled by the weighted average of the previous estimation $\hat{\mu}_{t-1}$ and the current estimation of $H_s(y_t - \hat{a}_t)$, which is a recursive equation similar to the monitoring statistic of the EWMA control chart. Therefore, for a stationary process, $\hat{\mu}_t$ can help average the noise over time, which leads to a stationary distribution with a much smaller variance than the original data. However, different from the EWMA control chart, we use (9) to estimate the true dynamic trend $\hat{\mu}_t$ in dynamic processes. The spatial structure of μ_t is captured by applying the projection matrix H_s. However, $\hat{\theta}_{a,t}$ (or equivalently $\hat{a}_t = B_{at} \hat{\theta}_{a,t}$) is unknown and should also be estimated. To efficiently solve for $\theta_{a,t}$, we first show that the loss function is equivalent to a weighted lasso formulation, which can be solved via an accelerated proximal gradient algorithm.

Proposition 3. Suppose the previous estimation $\hat{\theta}_1, \ldots, \hat{\theta}_{t-1}, \hat{\theta}_{a,1}, \ldots, \hat{\theta}_{a,t-1}$ are known, then Problem (8) is equivalent to the following weighted lasso formulation:

$$\begin{aligned} \min_{\theta_{a,t}} F(\theta_{a,t}) = \min_{\theta_{a,t}} &(y_t - B_{as}\theta_{a,t})^T (I - \tilde{\lambda}_t H_s)(y_t - B_{as}\theta_{a,t}) \\ &- 2(1 - \tilde{\lambda}_t)(y_t - B_{as}\theta_{a,t})^T y_{t-1} + \gamma \|\theta_{a,t}\|_1, \end{aligned} \tag{10}$$

where $\tilde{\lambda}_t = \frac{1}{1+\lambda_t}$.

The proof is given in Appendix D. To efficiently solve this weighted lasso formulation, we propose to use the proximal gradient method, which is a class of optimization algorithms focusing on minimization of the summation of a group of convex functions, some of which are nondifferentiable. The function $F(\theta_{a,t})$ in (10) is comprised of a differentiable convex function $f(\theta_{a,t}) = (y_t - B_{as}\theta_{a,t})^T (I - \tilde{\lambda}_t H_s)(y_t - B_{as}\theta_{a,t}) - 2(1 - \tilde{\lambda}_t)(y_t - B_{as}\theta_{a,t})^T y_{t-1}$ and a nondifferentiable L_1 penalty $g(\theta_a) = \gamma \|\theta_{a,t}\|_1$. It can be proved that the proximal gradient algorithm converges to a global optimum given R_s is a positive semidefinite matrix. This is true because $f(\theta_{a,t})$ is convex and Lipchitz continuous (see Appendix E for the proof of convexity and Appendix F for the proof of Lipchitz continuity.) According to the following proposition, the proximal gradient method leads to a closed-form solution for $\theta_{a,t}$ in each iteration of the optimization algorithm.

Proposition 4. The proximal gradient problem for (10), given by $\theta_{a,t}^{(k)} = \operatorname{argmin}_{\theta_{a,t}} \{ f(\theta_{a,t}^{(k-1)}) + \langle \theta_{a,t} - \theta_{a,t}^{(k-1)}, \nabla f(\theta_a^{(k-1)}) \rangle + \frac{L}{2} \|\theta_{a,t} - \theta_{a,t}^{(k-1)}\|^2 + \gamma \|\theta_{a,t}\|_1 \}$, has a closed-form solution in each iteration k, in the form of a soft-thresholding function as follows:

$$\theta_{a,t}^{(k)} = S_{\frac{\gamma}{L}} \left(\theta_{a,t}^{(k-1)} + \frac{2}{L} B_{as}^T \left(y_t - B_{as} \theta_{a,t}^{(k-1)} - \mu_t^{(k)} \right) \right), \tag{11}$$

where $L = 2\|B_{as}\|_2^2$.

The proof is given in Appendix G. Finally, by combining the estimator from both (9) and (11), Problem (8) can be solved iteratively and recursively with the accelerated proximal gradient algorithm as shown in Algorithm 3. The accelerated proximal gradient algorithm is an *accelerated* version of the proximal gradient (PG) algorithm, which is able to achieve a better convergence rate than the PG algorithm.

Algorithm 3. Recursive algorithm for roughness minimization

initialize

$$\theta_a^{(0)} = 0, \quad L = 2\|B_{as}\|_2^2, \quad t_0 = 1, \quad x_t^{(0)} = 0$$
$$H_s = B_s \left(B_s^T B_s + R_s \right)^{-1} B_s^T$$

for each time t

While $|\theta_{a,t}^{(k-1)} - \theta_{a,t}^{(k)}| > \epsilon$ **do**

Update

$$\begin{aligned} a_t^{(k-1)} &= B_{as} x_t^{(k-1)} \\ \mu_t^{(k-1)} &= (1 - \tilde{\lambda}_t)\hat{\mu}_{t-1} + \tilde{\lambda}_t H_s \left(y_t - a_t^{(k-1)} \right) \\ \theta_{a,t}^{(k)} &= S_{\frac{\gamma}{L}} \left(\theta_{a,t}^{(k-1)} + \frac{2}{L} B_{as}^T \left(y_t - a_t^{(k-1)} - \mu_t^{(k-1)} \right) \right) \\ t_k &= \frac{1 + \sqrt{1 + 4t_{k-1}^2}}{2} \\ x_t^{(k)} &= \theta_{a,t}^{(k)} + \frac{t_{k-1} - 1}{t_k} \left(\theta_{a,t}^{(k)} - \theta_{a,t}^{(k-1)} \right) \end{aligned}$$

end

4.3 ST-SSD for Stationary Processes

In a stationary process where the functional mean of the data stream is constant when the process is in-control, ST-SSD is simplified by removing the temporal basis of the mean, that is, $\mu = B_s\theta$. Hence, Equation (8) becomes $\operatorname{argmin}_{\theta_t,\theta_{a,t}} \|e\|^2 + \theta^T R\theta + \gamma\|\theta_a\|_1$, subject to $y_t = B_s\theta + B_{as}\theta_{a,t} + e_t$, which can be solved by Algorithm 3 with a slight modification in estimating μ. As the functional mean is constant, the temporal projection matrix reduces to a sample average function. Consequently, the functional mean in Algorithm 3 is estimated by $\hat{\mu}^{(k)} = H_s(\frac{1}{n}\sum_{i=1}^{n}(y_i - a_i^{(k-1)}))$. It is noteworthy that the ST-SSD model for stationary processes is a special case of the roughness minimization model with $\lambda_t \to \infty$ and the kernel model with $c \to 0$. More detailed discussions are given in Appendix H.

5. Online Process Monitoring and Diagnostics

In this section, we propose a monitoring procedure that combines the ST-SSD model with a sequential likelihood ratio test. We also discuss how ST-SSD can be used for diagnosis after a change is detected.

5.1 Construct Monitoring Statistics

We propose an online monitoring method using the estimated sparse anomalous features from ST-SSD. If the sparse vector of anomalies detected by ST-SSD, that is, $\hat{a}$ is statistically significant, it can be implied that a process change has occurred. In this article, we focus on two types of temporal changes: the first type, studied in the simulation study, is based on the change-point model where the anomaly appears after a time point τ. In the second type, discussed in the case study, the anomaly happens only in short-time windows. It should be noted that in both cases, the anomaly is nonsmooth in the temporal domain due to the sudden jump. We denote the detected anomaly at time t as $\hat{a}_t$. Therefore, at each time t, we test whether the expected residuals after removing the functional mean, denoted by $\mu_{r,t}$, is zero or has a mean shift in the direction of $\hat{a}_t$. That is,

$$H_0 : \mu_{r,t} = 0 \quad \text{versus} \quad H_1 : \mu_{r,t} = \delta\hat{a}_t;\ \delta > 0.$$

To test these hypotheses, a likelihood ratio test is applied to the residuals at each sampling time t, that is, $r_t = y_t - \mu_t$. This leads to the test statistic $T_\gamma(t) = \frac{(\hat{a}_t^T r_t)^2}{\hat{a}_t^T \hat{a}_t}$ (Hawkins 1993), in which it is assumed that the residuals r_t are independent after removing the functional mean and their distribution before and after the change remains the same. However, the test statistics $T_\gamma(t)$ relies on the selection of γ since it directly controls the sparsity of $\hat{a}_t$. To construct a more stable hypothesis test, inspired by Zou and Qiu (2009), we develop a monitoring statistic by combining multiple tuning parameters. Zou and Qiu (2009) proposed to use different values of the tuning parameter γ obtained from the breakpoints of the piecewise linear solution path of LASSO. This is a very time-consuming process. For example, for an image stream with the size of 350×350, the LARS algorithm finds the entire solution path in about 60 hr, which makes it impractical for real-time monitoring purposes. Consequently, we use a smaller set of possible tuning parameters denoted by Γ_{n_γ}. It is known that when γ is large enough, that is, $\gamma \geq \gamma_{\max}$, every element of coefficient θ_a will become 0. Therefore, we define the set of tuning parameter γ as $\Gamma_{n_\gamma} = \{\frac{\gamma_{\max} i}{n_\gamma} | i = 0, 1, \ldots, n_\gamma\}$ by dividing $(0, \gamma_{\max}]$ equally into n_γ intervals. The choice of $\gamma_{\max}$ is discussed in the next subsection. Thus, the combined test statistic can be defined as

$$\tilde{T}(t) = \max_{\gamma \in \Gamma_{n_\gamma}} \frac{T_\gamma(t) - E(T_\gamma(t))}{\sqrt{\operatorname{var}(\tilde{T}_\gamma(t))}}, \tag{12}$$

where $E(T_\gamma(t))$ and $\operatorname{var}(\tilde{T}_\gamma(t))$, respectively, are the mean and variance of $T_\gamma(t)$ under H_0, that are estimated using a set of in-control data. An out-of-control sample is detected when its corresponding monitoring statistic $\tilde{T}(t)$ is greater than a control limit h.

5.2 Control Limit Determination

The value of the control limit is computed based on a predetermined in-control average run length (or equivalently, Type I error rate) and the set of the tuning parameter values, Γ_{n_γ}. Zou and Qiu (2009) suggested to determine Γ_{n_γ} by using the least angle regression (LARS) algorithm (Efron et al. 2004) that provides the entire solution path. The breakpoints in such a solution path define the set Γ_q. However, the complexity of the LARS algorithm with p covariates is $O(np + p^3)$, which is infeasible for HD data. Alternatively, to define Γ_q, we use equidistant values of γ within a certain range. The procedure for computing the control limit h in Phase I analysis using an in-control sample of HD is summarized as follows: First, an ST-SSD algorithm such as Algorithm 1 (for 1D profile) or Algorithm 2 (for image or high-dimensional tensor) is applied to an in-control sample $Y = (y_1, y_2, \ldots, y_n)$ to estimate μ and a. The parameters λ_s and λ_t are tuned via the GCV criterion as proposed in Yan, Paynabar, and Shi (2015b) and the kernel bandwidth c are selected by using the cross-validation criterion. Next, the set of tuning parameter is defined by $\Gamma_{n_\gamma} = \{\frac{\gamma_{\max} i}{n_\gamma} | i = 0, 1, \ldots, n_\gamma\}$, $\gamma_{\max}$ is determined such that $\theta_a = 0$ for all the IC samples. Larger values of n_γ increase the detectability of the monitoring procedure. However, if too large, the monitoring procedure becomes computationally inefficient. In this article, based on numerical experiments, we found that for $n_\gamma \geq 20$ the detection power in detecting small shifts are similar. Therefore, we use $n_\gamma = 20$ in this article. After that, similar to Zou and Qiu (2009), assuming the dynamic mean can be estimated accurately (this is validated in the simulation study), we generate iid Gaussian random draws to simulate the residuals r_t. We then apply the ST-SSD on the simulated data, compute the monitoring statistics $\tilde{T}(t)$, and estimate its empirical distribution. Finally, the control limit is determined as a certain quantile of the empirical distribution of the monitoring statistics based on a predetermined IC average run length.

5.3 Diagnosis of Detected Changes

After the proposed control chart triggers an out-of-control signal, the next step is to diagnose the detected change. Diagnosis for functional data is defined by determining portions of data that have a different structure from the functional mean. In many cases, especially in the HD setting, estimating the location

of anomalies responsible for the out-of-control signal is important. This information would help process engineers identify and eliminate the potential root causes. Suppose that the control chart triggers a signal at time τ, we then apply the LRT test procedure described in the previous section to determine which γ provides the largest test statistics in (12), denoted by $j^* = \arg\max_{j=1,\cdots,q_t} \frac{T_{\gamma_j}(\tau)-E(T_{\gamma_j}(\tau))}{\sqrt{\text{var}(\tilde{T}_{\gamma_j}(\tau))}}$. Vector $\hat{a}_{\gamma_{j^*},\tau} = B_{as}\hat{\theta}_{a\tau}$ is the estimated anomalies for the optimal γ_{j^*} at time τ. Since a localized basis (e.g., band matrix) is used for B_{as}, the sparsity of $\hat{\theta}_{a\tau}$ leads to the sparsity of $\hat{a}_{\gamma_{j^*},\tau}$. Therefore, the nonzero elements of $\hat{a}_{\gamma_{j^*},\tau}$ can be used to identify the location of anomalies. If a nonlocalized basis is chosen, one may use thresholding to determine the anomalous region by $1(\hat{a}_{\gamma_{j^*},\tau} > \omega)$, where ω can be chosen by Ostu's method (Otsu 1975).

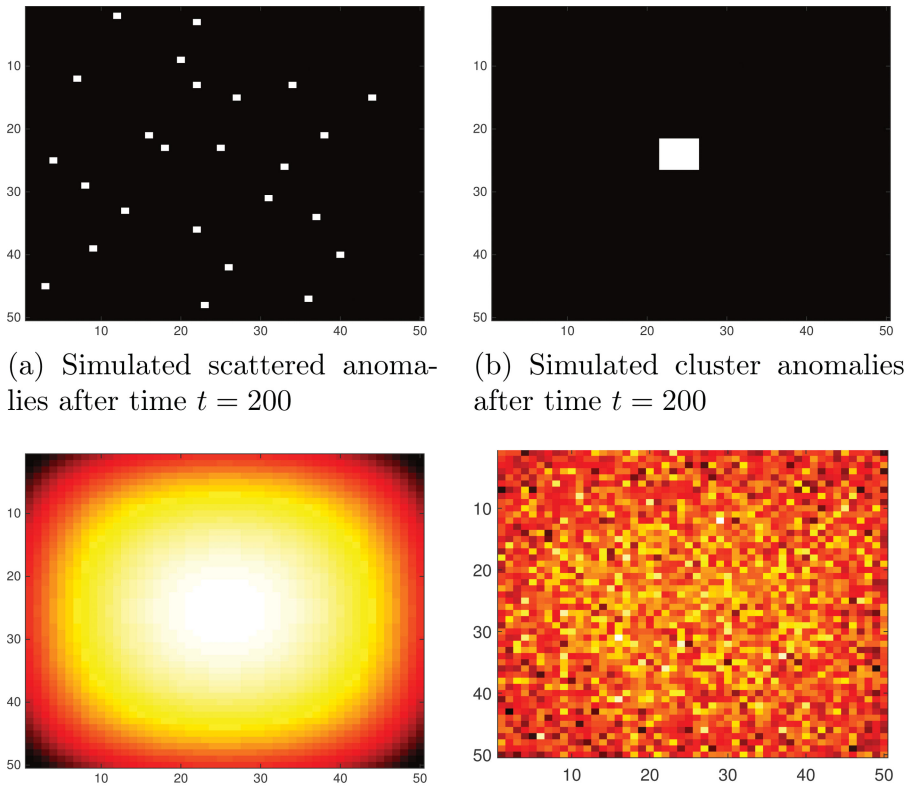
6. Simulation Study

In this section, the performance of the proposed methodology is evaluated by using simulated streams of images with a dynamic functional mean (background). To simulate the functional mean with smooth spatial and temporal structures, we mimic a heat transfer process, in which a 2D temperature map, $M(x, y, t)$, is generated according to the following heat transfer equation (Patankar 1980):

$$\frac{\partial M}{\partial t} - \alpha\left(\frac{\partial^2 M}{\partial x^2} + \frac{\partial^2 M}{\partial y^2}\right) = f,$$

where x, y, $0 \le x, y \le 1$ denote pixel locations on an image, α is the thermal diffusivity constant describing how fast a material can conduct thermal energy, and f describes the internal heat generation of the entire surface. In this simulation study, we set $\alpha = 1$. The initial and boundary conditions are set as $M|_{t=0} = 0$ and $M|_{x=0} = M|_{x=1} = M|_{y=0} = M|_{y=1} = 1$, respectively. At each time t, the functional mean $M(x, y, t)$ is recorded at points $x = \frac{i}{m+1}, y = \frac{j}{m+1}$; $i, j = 1, \ldots, m$, which results in an $m \times m$ matrix denoted by $M(t)$. In this study, we consider two types of anomalies, namely, clustered and scattered anomalies. Both types of anomalies are generated based on $S_0 = \delta I(s \in S_A)1(t > t_1)$, in which S_A is the set of anomalous pixels, δ characterizes the intensity difference between anomalies and the functional mean, $1(\cdot)$ is an indicator function, and t_1 is the time of the change. For the scattered case, S_A is a set of 25 pixels randomly selected throughout the image. For the clustered case, S_A is a randomly generated 5×5 square. Finally, the matrix of random noises, that is, $E_i \sim \text{NID}(0, \sigma^2)$ with $\sigma = 0.1$, is added to the generated image streams. A sample of simulated scattered and square anomalies, the simulated functional mean, and an example of simulated noisy image are shown in Figure 2(a)–2(d), respectively.

To model the spatial structure of each image $M(t)$, we use cubic B-spline basis with 10 knots in both x and y directions. For scattered anomalies, since the size of anomalies is very small and their locations are randomly chosen, an identity matrix can be used as the spatial basis. For the clustered anomalies, however, since anomalies form small continuous regions, a cubic B-spline



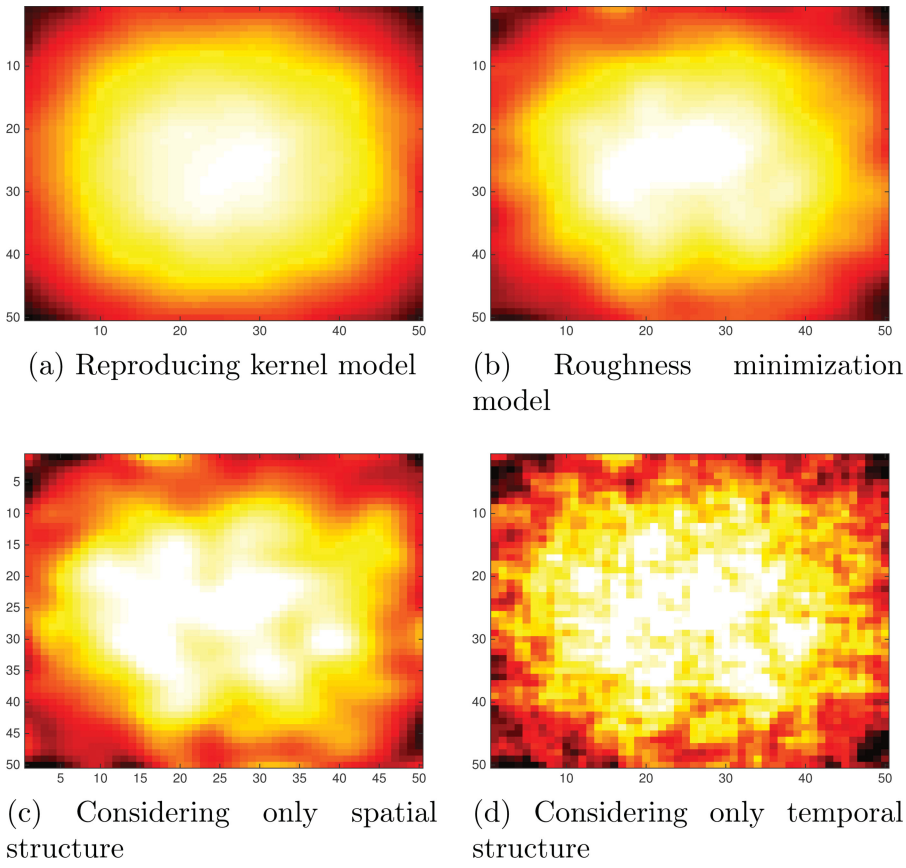
(a) Simulated scattered anomalies after time $t = 200$

(b) Simulated cluster anomalies after time $t = 200$

(c) Simulated functional mean at time $t_1 = 201$

(d) Simulated image coupled with noise and anomalies at time $t_1 = 201$

Figure 2. Simulated images with both functional mean and anomalies at time $t = 201$.



(a) Reproducing kernel model

(b) Roughness minimization model

(c) Considering only spatial structure

(d) Considering only temporal structure

Figure 3. Functional mean estimation results.

basis with 30 knots is used in both x and y directions. We also include the results of using an identity basis for the clustered case to study the sensitivity of the proposed method to the choice of bases. We then apply both versions of the ST-SSD model (i.e., kernel and roughness minimization) to the simulated streaming images.

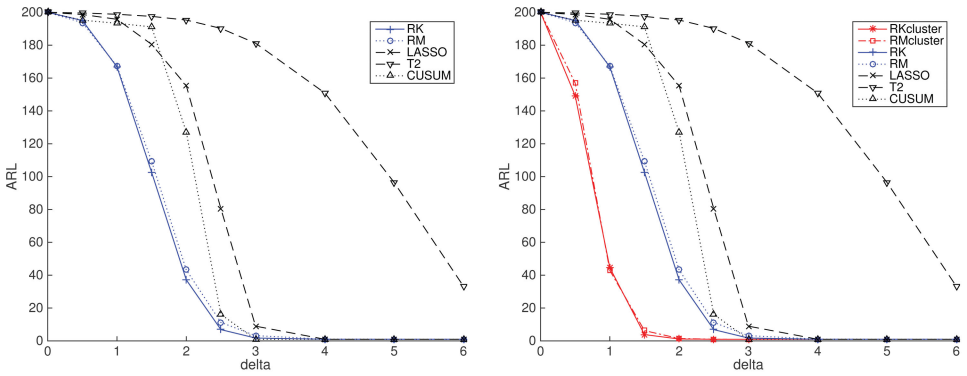
We first begin with evaluating the effectiveness of ST-SSD in estimating the functional mean. The estimated functional mean from a sample of data streams using both reproducing kernel (RK) and roughness minimization (RM) models are shown in Figure 3(a) and 3(b). The mean square errors (MSE) of the estimated mean are 2.320×10^{-5} and 8.400×10^{-5} for RK and RM, respectively, which indicates a slight advantage of the kernel basis due to its flexibility. Also, to show the importance of considering both spatial and temporal structures of data, in Figure 3(c) and 3(d), we plot the estimated functional mean when only either spatial or temporal structure is modeled. To estimate the functional mean with only spatial structure, we apply SSD on each single image with the same spatial spline basis used in the ST-SSD. To estimate the functional mean considering only the temporal structure, we apply the proposed RM method with the identity matrix as the spatial basis. The MSEs of the estimated mean for spatial and temporal models are, receptively, 2.32×10^{-4} and 3.92×10^{-4}, both larger than that of RK and RM. By comparing Figure 3 with Figure 2(c), it is clear that the estimated functional mean by our proposed ST-SSD is much closer to the true functional mean as it takes both spatial and temporal structures into account.

Next, we compare the performance of our method with a few benchmark methods in the literature. Specifically, we compare the proposed reproducing kernel (designated as RK for the identity spatial basis and as RKcluster for the cubic B-spline basis) and roughness minimization (designated as RM for the identity spatial basis and RMcluster for the cubic B-spline basis) methods with the Hotelling T^2 control chart (designated as "T2"), Lasso-based control chart proposed by Zou, Jiang, and Tsung (2011) (designated as LASSO) and local CUSUM control chart (Mei 2010, designated as CUSUM). It should be noted that none of benchmark methods can remove the temporal trend. Therefore, to have a fair comparison, we use a simple moving average filter with the window size of 5 to remove the temporal trend before applying the benchmark methods. We fix the in-control ARL_0 for all methods to be 200 and compare the out-of-control ARL_1 under different anomaly intensity levels δ. As per reviewer's suggestion, we also conduct the simulation in the case of static functional mean and the result is included in the online appendix.

The average time of computing the monitoring statistics for a sample is given in Table 1, and the out-of-control ARL curves of clustered and scattered anomalies obtained from 1000 simulation replications are shown in Figure 4(a) and 4(b), respectively. In both cases of scattered and clustered anomalies, it is clear

Table 1. Computation time of ST-SSD and other benchmark methods.

	RK	RM	LASSO	CUSUM	T2
Time	0.13 sec	0.015 sec	5.2e-3 sec	2.0e-4 sec	1.6e-4 sec



(a) ARL for scattered defect in different δ

(b) ARL Run Length for clustered defect in different δ

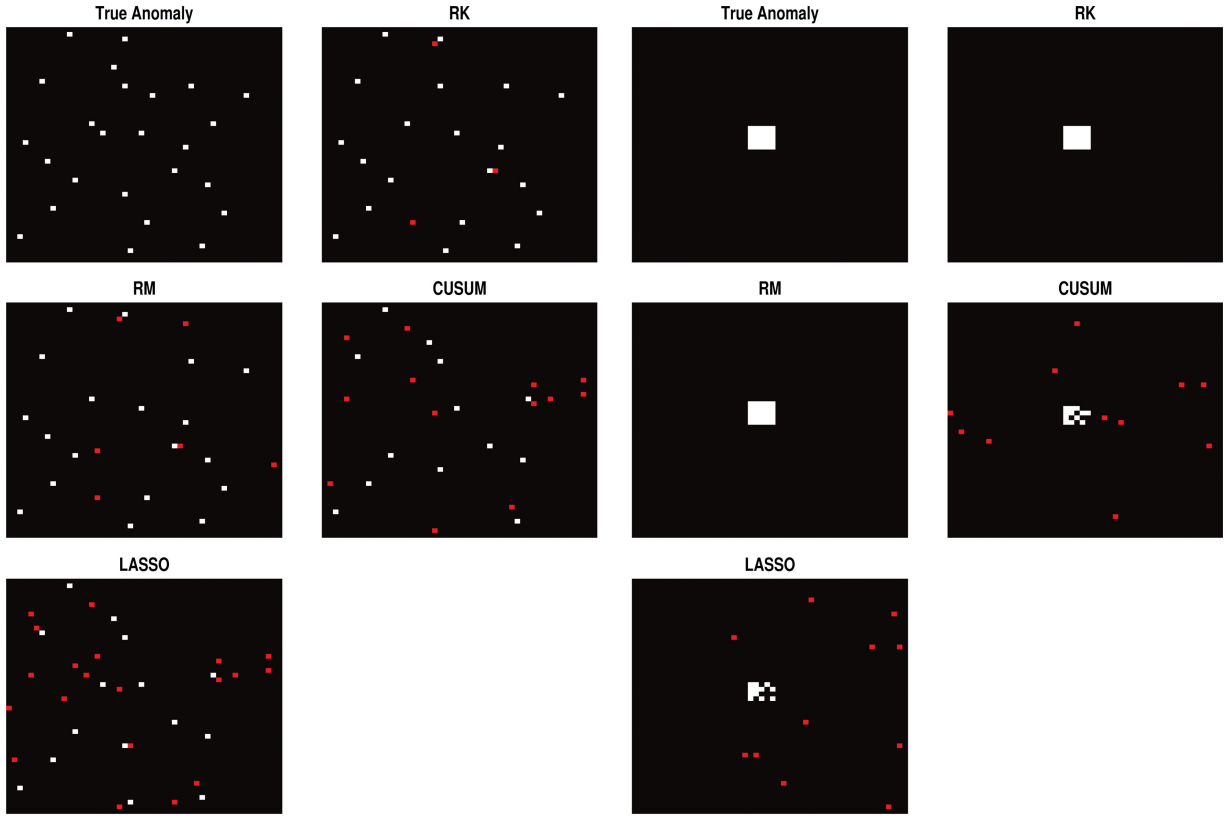
Figure 4. Detection power comparison based on ARL.

that RK and RM models have better detection performance than other benchmark methods. The RK method performs slightly better than RM due to its accuracy in modeling the temporal trend. However, RM is slower in terms of the computation time because of its higher modeling complexity. The reason for the poor performance of the local CUSUM and lasso-based control charts is that they lack the ability to model both the spatial structure and the temporal trend at the same time. Hotelling T^2 control chart performs the worst because it is based on a multivariate hypothesis test, whose power deteriorates as the data dimensions increase, hence, not scalable to HD data streams. Moreover, in the case of clustered anomalies, the proposed RK and RM models with spline basis detect the changes significantly quicker than those with identity basis. For example, in the clustered anomaly case, for a small shift with $\delta = 1$, the ARL for both RKcluster and RMcluster is around 40, while the ARL for other methods without considering the spatial structure are at least about four times larger (≥ 170). This indicates the importance of accurate modeling of the spatial structure in addition to the temporal tend. The ARL of the benchmark methods for such a shift is close to the in-control ARL of 200, indicating that these methods are not capable of detecting small changes. In conclusion, even if the computational time of RK and RM is much larger than LASSO, T2, and CUSUM, it is still small enough to be used for online monitoring. Furthermore, the performance of RK and RM is much better especially in the clustered anomaly case. A video of one simulation run along with the ST-SSD results and the corresponding control chart is given in the online appendix.

Finally, we evaluate and compare the performance of the diagnosis method with benchmark methods. For this purpose, we compute the following four criteria after a shift is detected: (i) precision, defined as the proportion of detected anomalies that are true anomalies; (ii) recall, defined as the proportion of the anomalies that are correctly identified; (iii) F measure, a single criterion that combines the precision and recall by calculating their harmonic mean; and (iv) the corresponding ARL. The average values of these criteria over 1000 simulation replications for $\delta = 2$ and $\delta = 3$ are given in Table 2. An example of detected anomalies for both scattered and clustered cases with $\delta = 3$ are also shown in Figure 5, in which incorrectly classified points are shown in red. It is clear from this figure and Table 2 that the proposed RK and RM models have a much better diagnostics performance than other benchmark methods. This difference is more pronounced in the clustered case where the benchmark methods fail to model the spatial structure of anomalies. For example, for the scattered case, the F measure of both RK and LASSO is around 0.25. However, in the clustered case, this measure is 0.84 for kernel, while lasso's measure remains the same. Moreover, the diagnostics measures of ST-SSD methods (i.e., RK and RM) in the clustered case is much better than the corresponding measures in the scattered case. This is because the spatial structure of defects in the clustered case is well captured by the B-spline basis.

Table 2. Monitoring and diagnostics result when $\delta = 2$ and $\delta = 3$ (precision, recall, and F, the larger the better; ARL, the smaller the better)

	Scattered anomalies $\delta = 2$				Scattered anomalies $\delta = 3$			
Methods	Precision	Recall	F	ARL	Precision	Recall	F	ARL
RK	0.2357	0.2764	0.2544	37.17	0.6106	0.5500	0.5738	1.73
RM	0.2535	0.2532	0.2533	43.11	0.5851	0.5560	0.5656	1.83
LASSO	0.2553	0.2204	0.2366	155.39	0.5719	0.4892	0.5257	8.96
CUSUM	0.1092	0.1136	0.1114	124.88	0.5187	0.5394	0.5289	1.86
T2	—	—	—	195.32	—	—	—	181.23
Methods	Clustered anomalies $\delta = 2$				Clustered anomalies $\delta = 3$			
RKcluster	0.8515	0.8596	0.8415	1.11	0.9424	0.9464	0.9444	1.00
RMcluster	0.8490	0.7934	0.8202	1.46	0.9163	0.9474	0.9316	1.00
LASSO	0.2498	0.2160	0.2297	153.88	0.5880	0.4952	0.5333	8.57
CUSUM	0.1100	0.1144	0.1121	121.85	0.5195	0.5402	0.5296	1.95
T2	—	—	—	195.32	—	—	—	181.23



(a) Scattered defects with $\delta = 3$ (b) Clustered defects with $\delta = 3$

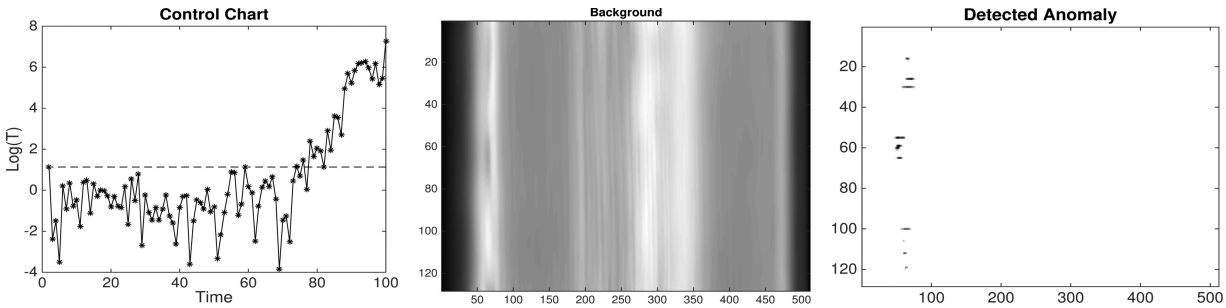
Figure 5. Detected anomalies by using different methods (incorrectly identified pixels are shown in red).

7. Case Study

In this section, the proposed monitoring method is applied to three real datasets collected from a steel rolling process, a solar data observatory, and a stamping process. In the first two cases, we analyze images with a dynamic functional mean and in the third case we study multi-channel profiles with a static functional mean.

7.1 Online Seam Detection in Steel Rolling Process

Rolling is a high-speed deformation process that uses a set of rollers to reduce the cross-section of a long steel bar by applying compressive forces for achieving certain uniform diameters (Kalpakjian, Schmid, and Kok 2008). Surface defects such as seam defects can result in stress concentration on the bulk material that may cause failures when a steel bar is used. Therefore, early detection of anomalies is vital to prevent product damage and to reduce manufacturing costs. Traditionally, due to the high speed of the rolling process (e.g., 225 mile per hour), seam detection has been limited to off-line manual inspection. In recent years, with the development of advanced sensing and imaging technologies, vision sensors have been successfully adopted in rolling processes, collecting high-resolution images of the product surface with a high data acquisition rate. In this case study, a stream of surface images of a rolling bar is used to validate our methodology. We collect a sample of 100 images with the size of 128×512 pixels. The first 50 images are in-control samples with no defects. As an example, one frame of the image stream is shown in Figure 1(a). An image of a rolling bar is generally



(a) Log of testing statistics (b) Functional mean estimation (c) Anomalies estimation

Figure 6. Detection results for rolling example at time $t = 97$.

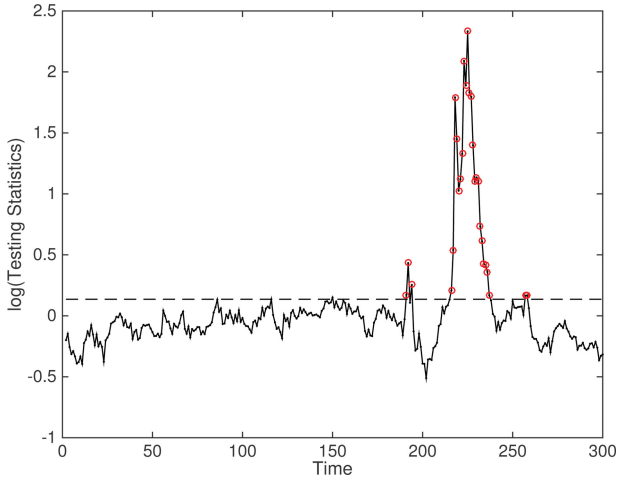


Figure 7. Log of testing statistics in solar flare monitoring.

smooth in the rolling direction (vertical direction). Moreover, seam defects that have a high contrast against the functional mean (image background) are typically sparse (Jia et al. 2004), which justifies the use of ST-SSD model for analyzing this data stream. We apply the proposed RM method to monitor rolling process and detect potential defects on the surface. To model the functional mean in y direction, a B-Spline basis with five knots is used for B_y and $B_x = I_x$. We also use an identity matrix basis for anomalies in both the x and y directions, that is, $B_{ax} = I_{ax}$ and $B_{ay} = I_{ay}$.

Since the dynamic behavior of the functional mean is not intricate, the roughness minimization model described in Section 4.2, is used. The testing statistic in (12) is calculated and plotted in Figure 6(a). The control limit for this case and other examples presented in this section is determined using the in-control data and according to the procedure presented in Section 5.2. Seam defects often occur toward the end of the rolling bar. It is clear from the image stream (see the online appendix), the first defect appears at time $t = 76$, which is the first out-of-control point in the control chart. The computational time is 0.35 sec per sample, which is sufficiently fast for online monitoring. To illustrate the effectiveness of the diagnosis procedure, the estimated functional mean and detected defects in one out-of-control image recorded at $t = 97$ is shown in Figure 6(b) and 6(c), respectively. The original image is also shown in Figure 1(a). As we can see from Figure 6, the estimated functional mean (background) is smooth in the y direction and the detected defects are sparse and demonstrate

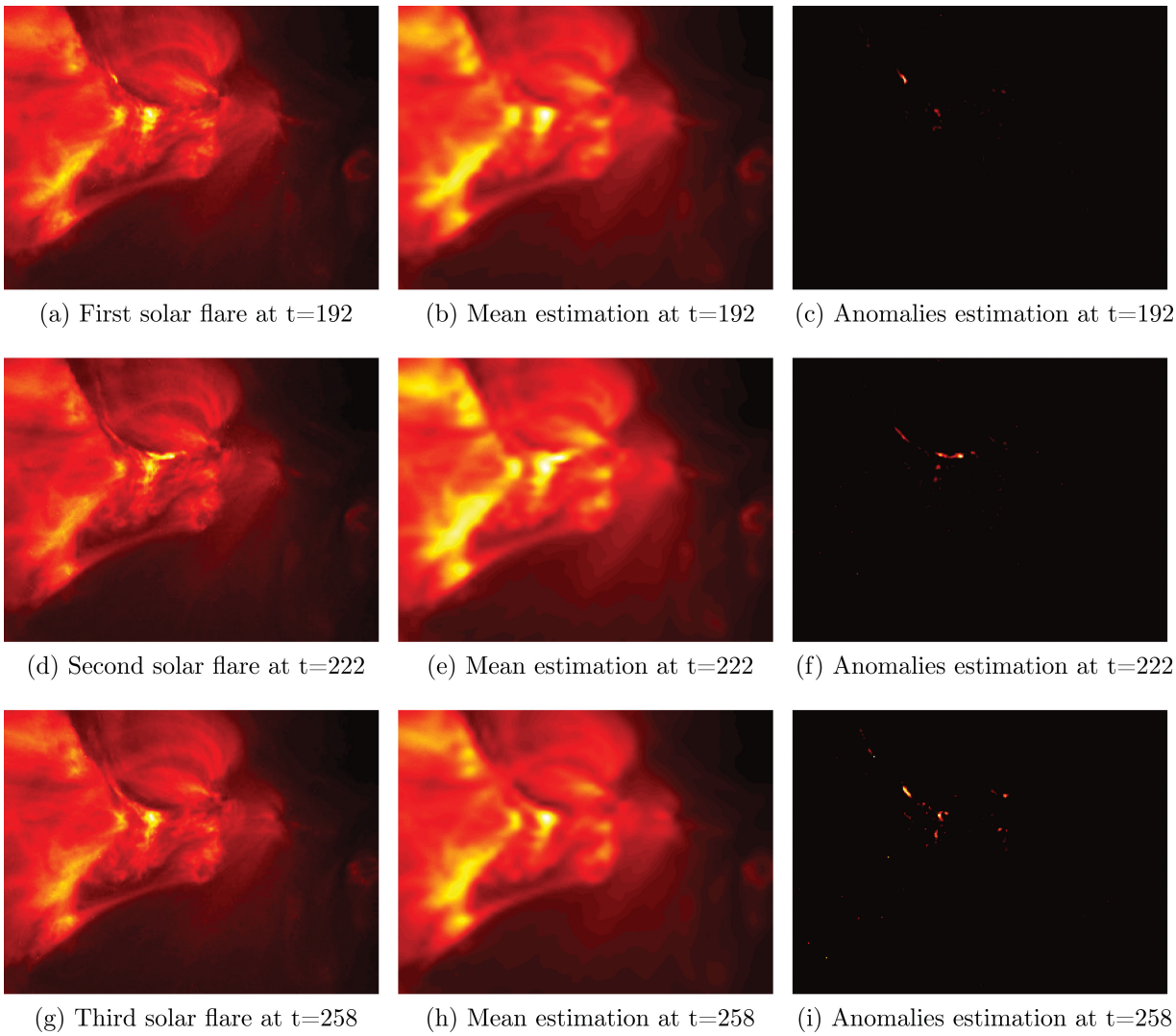


Figure 8. Detection results in three solar frames at time $t = 192, 222, 258$.

certain repeated patterns suggesting that the roller may be damaged.

7.2 Online Monitoring of Solar Activity

In the second example, a stream of solar images are used for monitoring of solar activities and detection of solar flares. A solar flare emits a large number of energetic charged particles, which may potentially cause the failure of large-scale power-grids. Thus, quick detection of solar flares is important for preventive and corrective actions. The solar temperature slowly changes over time and solar bursts are sparse in both the time and space, which makes process monitoring challenging. Existing detection methods that simply remove the functional mean (background) by subtracting the sample mean are incapable of detecting small transient flares in the dynamic system (Xie, Huang, and Willett 2013).

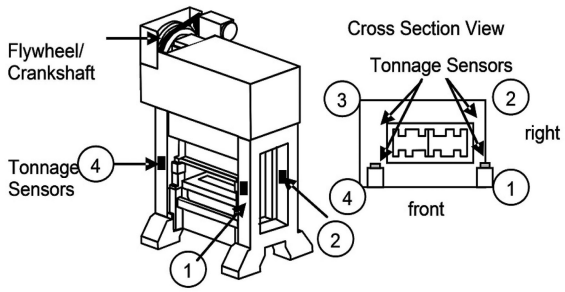
This dataset is publicly available online at *http://nislab.ee.duke.edu/MOUSSE/index.html*. In this dataset, a sequence of images of size 232×292 pixels was captured by satellite. A sample of 300 frames is used in this case study and the first 100 frames are considered as the in-control sample. To detect the solar flare in real-time, the proposed RM monitoring method is applied with the following specification: To model the smooth functional mean (background), B-Spline basis with 50 knots are used as B_x and B_y; to model the sparse anomalies (solar flares), we select the identity matrix for the anomalies in both the x and y directions, that is, $B_{ax} = I_{ax}$ and $B_{ay} = I_{ay}$. The logarithm of the test statistic obtained from (12) is plotted in Figure 7. As can be seen form the control charts, three solar flares are detected. The first two solar flares occurred at intervals [191, 194] and [216, 237], which is compatible with the results reported in Xie, Huang, and Willett (2013) and Liu, Mei, and Shi (2014). Additionally, we are able to detect a third small flare at the interval [257, 258], which was not detected by the existing two-step approaches (i.e., Xie, Huang, and Willett 2013; Liu, Mei, and Shi 2014). Computation time is about 0.12 sec per frame, which enables online monitoring. Note that although image frames in both case studies have similar number of pixels, the computation time for the analysis of solar images is smaller than that of rolling images. The reason is that the computational complexity for the proposed algorithm is $O(n_x^3 + n_y^3)$, which is in order of 1.2×10^8 and 4×10^7 for rolling and solar images, receptively. This makes the computation time for solar images approximately three times lower.

Furthermore, to find the location of the solar flares in out of control images, the estimated functional mean (the background) and anomalies (solar flares) corresponding to time $t = 192, 222, 258$ are shown in Figure 8. As can be seen form the figure, the proposed method not only is able to detect the changes, but also can identify the location of solar flares in different time frames.

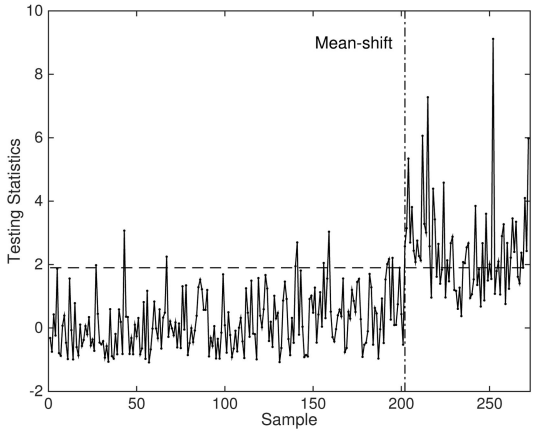
7.3 Tonnage Signal Monitoring

We also use the proposed methodology to monitor multi-channel tonnage profiles collected in a multi-operation forging process. In this process, four strain gauge sensors, each mounted on one column of the forging machine, measure the exerted tonnage force of the press uprights as shown in Figure 9(a). This results in a four-channel tonnage profile in each cycle of operation. The dataset used in this case study contains 202 in-control profiles collected under normal production condition and 69 out-of-control profiles in which there is a missing part in the piercing operation die. As pointed out by Lei, Zhang, and Jin (2010), Paynabar, Jin, and Pacella (2013), and Paynabar, Qiu, and Zou (2015), a missing part only affects certain segments of the tonnage profile, which implies that the change is sparse. Hence, in this case study, we only focus on the peak area of the tonnage profile, which is mostly affected by a missing part. The length of the peak profiles for each channel is 569. Examples of peak profiles for both normal and faulty conditions are shown in Figure 1(c).

Since the signal mean is static, the proposed static model is applied. However, to model the spatial structure of the profile mean and anomalies, cubic B-spline bases with 10 and 90 knots are used, respectively. We use the sequence of in-control profiles to estimate the control limit. Out of 202 samples collected under the normal operations, 9 samples are specified as out-of-control. After removing these outlier samples and recalculating the control limit, the proposed monitoring method is applied to the sequence of faulty profiles and the resulting control chart is shown in Figure 9(b). As shown in the figure, there is a clear change in the mean of the monitoring statistic, indicating that the monitoring method can detect the profile changes caused by missing parts. Overall 44 out of 69 faulty samples are beyond the control limit, which is roughly equivalent to the out-of-control

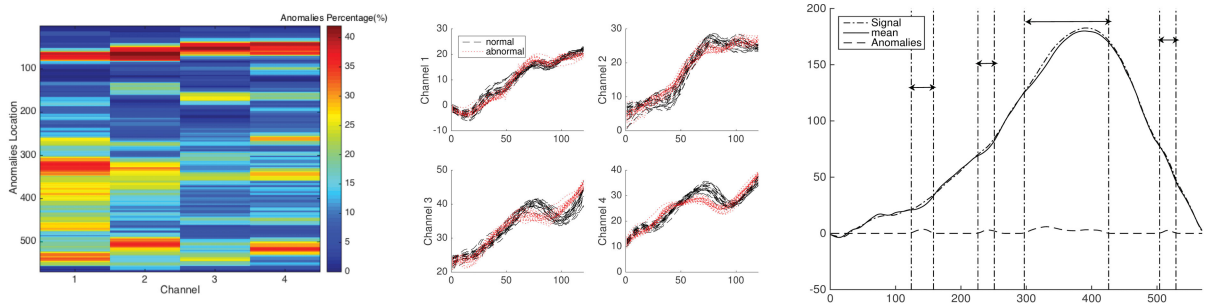


(a) A Stamping press and tonnage sensor location



(b) Tonnage signal monitoring result

Figure 9. Monitoring forging process using multi-channel tonnage signal.



(a) Diagnostics of the tonnage signal (b) 4-channel tonnage signals between $0-120$ (c) Tonnage signal anomalies location

Figure 10. Tonnage signal diagnostics.

ARL of 1.5. The computational time on average is 0.25 sec per sample.

Moreover, we use all out-of-control samples to perform diagnosis analysis. The percentage of identified anomalies by our diagnosis method across different channels and segments are shown in a colormap in Figure 10(a). Warmer colors imply that more out-of-control samples contain anomalies in the corresponding channel segment. As can be seen in Figure 10, anomalies mostly occur in the segment [44, 100], segment [319, 346], and segment [497, 535]. For example, to visualize differences between normal and faulty signals, we plotted 20 normal and 20 abnormal signals for all channels in the segment [0, 120] in Figure 10(b). As can be seen from the figure, there is a clear local difference between normal and faulty signals in the segment [44, 100] although it is not clear in Figure 1 where the whole signal is shown. The proposed diagnosis algorithm is capable of locating such small clustered changes as shown in Figure 10(a), which Channel 1 has most clustered changes. This is because Sensor 1 is mounted on the front side of the forging machine where the die with missing parts is located. Figure 10(c) shows one example of faulty profile recorded by Sensor 1 along with the profile mean and the identified anomalous segment. As can be seen from Figure 10(c), the main difference between the signal and the profile mean is picked up by the diagnosis procedure. These findings are consistent with those in Lei, Zhang, and Jin (2010); Paynabar, Qiu, and Zou (2015).

8. Conclusion

Online monitoring of high-dimensional streaming data with complex spatio-temporal structure is very important in various manufacturing and service applications. In this article, we proposed a novel methodology for real-time monitoring of HD data streams. In our methodology, we first developed ST-SSD that effectively decomposes a data stream into a smooth functional mean and sparse anomalies by considering the difference in the spatio-temporal structures of the functional mean and anomalies. Similar to SSD, we formulated ST-SSD in the form of high-dimensional regression augmented with penalty terms to encourage both the smoothness of the spatio-temporal functional mean and the sparsity of anomalies. To effectively solve this large-scale convex optimization problem, we used APG methods and developed efficient iterative algorithms that have closed-form solutions in each iteration. This method can be applied to identify anomalies and the functional mean for a fixed number of samples, which can only be applied in offline phase-I monitoring. To handle challenges of the increasing number of observations in online monitoring, reproducing kernel and roughness minimization models were developed as two temporal modeling methods that provide a recursive estimation scheme for ST-SSD. This enables real-time implementation of ST-SSD. Then, a sequential likelihood-ratio-test-based control chart was proposed for monitoring. In the simulation study, we showed that the proposed methods outperform existing process monitoring approaches that fail to effectively model both the spatial structure and temporal trend. Finally, the proposed method was applied to three real case studies including steel rolling, solar activity, and tonnage signal monitoring. The results from all case studies demonstrated the capability of the proposed methods in identifying not only the time of process changes, but also the location of detected anomalies.

There are several potential research directions to be investigated. One possible nontrivial extension is to generalize SSD for other types of spatial and temporal structures such as non-smooth and/or periodic functional mean. To model different types of spatial and temporal structures, one may adjust the basis, for example, by using Fourier or wavelet basis.

Appendix A: Decomposition of The Projection Matrix

Since $B_s^T B_s + R_s = \otimes_{i=l}^{1}(B_{si}^T B_{si} + R_{si})$, from the property of Kronecker product, we know that $(B_s^T B_s + R_s)^{-1} = \otimes_{i=l}^{1}(B_{si}^T B_{si} + R_{si})^{-1}$.

Finally, we have $H_s = B_s(B_s^T B_s + R_s)^{-1} B_s^T = \otimes_{i=l}^{1} B_{si}(B_{si}^T B_{si} + R_{si})^{-1} B_{si}^T = \otimes_{i=l}^{1} H_{si}$.

Appendix B: Prove of The Recursive Estimation of H_t

We can apply the standard block matrix inversion formula as follow, $M = \begin{bmatrix} A & b \\ b^T & d \end{bmatrix} \in \mathbb{R}^{n\times n}$, $A \in \mathbb{R}^{(n-1)\times(n-1)}$, $b \in \mathbb{R}^{(n-1)\times 1}$, g is a scalar, then $M^{-1} = \begin{bmatrix} A^{-1}(I + bb^T A^{-1} g) & -A^{-1} bg \\ -b^T A^{-1} g & g \end{bmatrix}$, with $g = (d - b^T A^{-1} b)^{-1}$.

Therefore, $K_{\lambda_t,t} = (K_t + \lambda_t I)^{-1} = \begin{bmatrix} K_{t-1} + \lambda_t I_{t-1} & k_{t-1} \\ k_{t-1}^T & 1 + \lambda_t \end{bmatrix}^{-1}$. Following this , it is straightforward to show that $K_{\lambda_t,t} = \begin{bmatrix} K_{\lambda_t,t-1}(I + k_{t-1} k_{t-1}^T K_{\lambda_t,t-1} g_{t-1}) & -K_{\lambda_t,t-1} k_{t-1} g_{t-1} \\ -k_{t-1}^T K_{\lambda_t,t-1} g_{t-1} & g_{t-1} \end{bmatrix}$, where $g_{t-1} = (1 +$

$\lambda_t - r_{t-1}^T k_{t-1})^{-1}$. Therefore, the $\tilde{H}_t$ can be computed recursively by

$$\tilde{H}_t = K_t K_{\lambda_t,t} = \begin{bmatrix} K_{t-1} & k_{t-1}^T \\ k_{t-1} & 1 \end{bmatrix} \times \begin{bmatrix} K_{\lambda_t,t-1}(I + k_{t-1}k_{t-1}^T K_{\lambda_t,t-1} g_{t-1}) & -K_{\lambda_t,t-1}k_{t-1}g_{t-1} \\ -k_{t-1}^T K_{\lambda_t,t-1} g_{t-1} & g_{t-1} \end{bmatrix} = \begin{bmatrix} H_{t-1} - k_{t-1}r_{t-1}^T g_{t-1}(I_{t-1} - H_{t-1}) & (I_{t-1} - H_{t-1})k_{t-1}g_{t-1} \\ r_{t-1}^T(I_{t-1} + k_{t-1}r_{t-1}g_{t-1} - g_{t-1}) & (1 - r_{t-1}^T k_{t-1})g_{t-1} \end{bmatrix},$$

where $r_t = K_{\lambda_t,t} k_t$.

Appendix C: Roughness Minimization Modeling Estimator $\hat{\theta}_t$

Proof. First, from the property of Kronecker product we know that $\theta^T(A \otimes B)\theta = \mathrm{tr}(A\Theta^T B\Theta)$ if $\theta = \mathrm{vec}(\Theta)$. The penalization term can be reduced to

$$\begin{aligned} \theta^T R\theta &= \theta^T\left(I_t \otimes R_s + \lambda_t D_t^T D_t \otimes B_s^T B_s + \lambda_t D_t^T D_t \otimes R_s\right)\theta \\ &= \mathrm{tr}\left(\Theta^T R_s \Theta + D_t\Theta^T\left(B_s^T B_s + R_s\right)\Theta D_t^T\right) \\ &= \sum_{i=1}^{t}\left(\theta_i R_s\theta_i + (\theta_{i+1} - \theta_i)^T\left(B_s^T B_s + R_s\right)(\theta_{i+1} - \theta_i)\right). \end{aligned}$$

Finally $\hat{\theta}_t$ can be solved by

$$\begin{aligned} \hat{\theta}_t &= \underset{\theta_t}{\mathrm{argmin}} \sum_{i=1}^{t}\left((\theta_i R_s\theta_i + (\theta_{i+1} - \theta_i)^T\left(B_s^T B_s + R_s\right) \times (\theta_{i+1} - \theta_i)) + \|y_t - B_s\theta_t - a_t\|^2\right) \\ &= \underset{\theta_t}{\mathrm{argmin}}\, \lambda_t(\theta_t - \theta_{t-1})^T\left(B_s^T B_s + R_s\right)(\theta_t - \theta_{t-1}) \\ &\quad + \theta_t^T R_s\theta_t + \|y_t - B_s\theta_t - a_t\|^2 \qquad \text{(C.1)} \\ &= \underset{\theta_t}{\mathrm{argmin}}(1 + \lambda_t)\theta_t^T\left(B_s^T B_s + R_s\right)\theta_t \\ &\quad - 2\theta_t^T\left(\lambda_t B_s^T B_s\theta_{t-1} + \lambda_t R_s\theta_{t-1} + B_s^T(y_t - S_t)\right) \\ &= \left(\frac{\lambda_t}{1+\lambda_t}\theta_{t-1} + \frac{1}{1+\lambda_t}\left(B_s^T B_s + R\right)^{-1}B_s^T(y_t - a_t)\right) \\ &= (1 - \tilde{\lambda}_t)\theta_{t-1} + \tilde{\lambda}_t\left(B_s^T B_s + R\right)^{-1}B_s^T(y_t - a_t). \end{aligned}$$

The first equation holds since $\theta_1, \ldots, \theta_{t-1}$ is fixed, only the last term of the summation ($i = t$) is considered. Finally, we know that $\hat{y}_t = B_s\theta_t = (1 - \tilde{\lambda}_t)\hat{y}_{t-1} + \tilde{\lambda}_t H_s(y_t - a_t)$ because $H_s = B_s(B_s^T B_s + R_s)^{-1}B_s^T$. □

Appendix D: Equivalency of Equation (8) To Weighted Lasso Formulation

Proof. According to Appendix A, we have solved θ_t by fixing other variables as $\hat{\theta}_t = \frac{\lambda_t}{1+\lambda_t}\theta_{t-1} + \frac{1}{1+\lambda_t}(B_s^T B_s + R)^{-1}B_s^T(y_t - a_t)$. Then, by plugging it into (C.1), and considering the terms that only contain $y_t - a_t$, we have

$$\begin{aligned} &\lambda_t(\theta_t - \theta_{t-1})^T\left(B_s^T B_s + R_s\right)(\theta_t - \theta_{t-1}) \\ &= \tilde{\lambda}_t(1 - \tilde{\lambda}_t)\left((y_t - a_t)^T B_s\left(B_s^T B_s + R\right)^{-1} - \theta_{t-1}^T\right) \\ &\quad \times \left(B_s^T B_s + R_s\right)\left(\left(B_s^T B_s + R\right)^{-1}B_s^T(y_t - a_t) - \theta_{t-1}\right) \\ &= \tilde{\lambda}_t(1 - \tilde{\lambda}_t)\left((y_t - a_t)^T H_s(y_t - a_t) - 2(y_t - a_t)^T B_s\theta_{t-1}\right) \\ &\quad + C_0\ \theta_t^T R_s\theta_t = \tilde{\lambda}_t^2(y_t - a_t)^T B_s\left(B_s^T B_s + R_s\right)^{-1} \\ &\quad \times R_s\left(B_s^T B_s + R\right)^{-1}B_s^T(y_t - a_t) + 2\tilde{\lambda}_t(1 - \tilde{\lambda}_t) \\ &\quad \times (y_t - a_t)^T B_s\left(B_s^T B_s + R_s\right)^{-1}R_s\theta_{t-1} + C_1 \\ &\quad \times \|y_t - B_s\theta_t - a_t\|^2 = \|(I - \tilde{\lambda}_t H_s)(y_t - a_t) \\ &\quad - (1 - \tilde{\lambda}_t)\hat{y}_{t-1}\|^2 = (y_t - a_t)^T(I - \tilde{\lambda}_t H_s)^2(y_t - a_t) \\ &\quad - 2(1 - \tilde{\lambda}_t)(y_t - a_t)^T(I - \tilde{\lambda}_t H_s)\hat{y}_{t-1} + C_2. \end{aligned}$$

C_0, C_1, C_2 are the constant terms that do not include a_t. Finally, by only taking consideration of the quadratic and linear term of $y_t - a_t$. Equation (8) becomes

$$\begin{aligned} &\|y_t - B_s\theta_t - a_t\|^2 + \lambda_t(\theta_t - \theta_{t-1})^T\left(B_s^T B_s + R_s\right)(\theta_t - \theta_{t-1}) \\ &\quad + \theta_t^T R_s\theta_t + \gamma\|\theta_{a,t}\|_1 \\ &= (y_t - a_t)^T Q(y_t - a_t) + (y_t - a_t)^T P + \gamma\|\theta_{a,t}\|_1. \qquad \text{(D.1)} \end{aligned}$$

In which

$$\begin{aligned} Q &= \tilde{\lambda}_t(1 - \tilde{\lambda}_t)H_s + \tilde{\lambda}_t^2 B_s\left(B_s^T B_s + R_s\right)^{-1} \\ &\quad \times R_s\left(B_s^T B_s + R\right)^{-1}B_s^T + (I - \tilde{\lambda}_t H_s)^2 \\ &= (\tilde{\lambda}_t H_s - \tilde{\lambda}_t^2 H_s) + \tilde{\lambda}_t^2\left(H_s - H_s^2\right) + I - 2\tilde{\lambda}_t H_s + \tilde{\lambda}_t^2 H_s^2 \\ &= I - \tilde{\lambda}_t H_s. \end{aligned}$$

The second "=" holds because $B_s(B_s^T B_s + R_s)^{-1}R_s(B_s^T B_s + R)^{-1}B_s^T = H_s - H_s^2$ and

$$\begin{aligned} P &= 2\tilde{\lambda}_t(1 - \tilde{\lambda}_t)B_s\left(B_s^T B_s + R_s\right)^{-1}R_s\theta_{t-1} \\ &\quad + 2B_s\theta_{t-1} - 2(1 - \tilde{\lambda}_t)(I - \tilde{\lambda}_t H_s)\hat{y}_{t-1} \\ &= 2\tilde{\lambda}_t(1 - \tilde{\lambda}_t)B_s\left(\left(B_s^T B_s + R_s\right)^{-1}R_s - I\right)\theta_{t-1} \\ &\quad - 2(1 - \tilde{\lambda}_t)(I - \tilde{\lambda}_t H_s)\hat{y}_{t-1} = -2\tilde{\lambda}_t(1 - \tilde{\lambda}_t) \\ &\quad \times H_s B_s\hat{y}_{t-1} - 2(1 - \tilde{\lambda}_t)(I - \tilde{\lambda}_t H_s)\hat{y}_{t-1} = -2(1 - \tilde{\lambda}_t)\hat{y}_{t-1}. \end{aligned}$$

The third "=" holds because $B_s((B_s^T B_s + R_s)^{-1}R_s - I)\theta_{t-1} = -B_s(B_s^T B_s + R_s)^{-1}B_s^T B_s\theta_{t-1} = -H_s B_s\theta_{t-1} = -H_s\hat{y}_{t-1}$.

Finally, plugging P, Q, and $a_t = B_{as}\theta_{a,t}$ into (D.1), we will have (10). □

Appendix E: Convexity of $f(\theta_a) = (y_t - B_{as}\theta_{a,t})^T(I - \tilde{\lambda}_t H_s)(y_t - B_{as}\theta_{a,t}) - 2(1 - \tilde{\lambda}_t)(y_t - B_{as}\theta_{a,t})^T y_{t-1}$

To prove $f(\theta_a)$ is convex, we only need to show that $I - \tilde{\lambda}_t H_s$ is a positive semidefinite matrix, in which $\tilde{\lambda}_t = \frac{1}{1+\lambda_t} \in (0, 1)$.

We first show that H_s is positive semidefinite matrix. $H_s = B_s(B_s^T B_s + \lambda_s R_s)^{-1}B_s^T$. Since $(B_s^T B_s + \lambda_s R_s)^{-1}$ is a positive definite matrix, we know H_s is also a positive definite matrix.

We then show that $I - H_s$ is positive semidefinite matrix by $I - H_s = (I - H_s)^2 + \tilde{\lambda}_t B_s(B_s^T B_s + \lambda_s R_s)^{-1}R_s(B_s^T B_s + \lambda_s R_s)^{-1}B_s^T$, and both terms are positive semidefinite matrices.

We then know $I - \tilde{\lambda}_t H_s = \tilde{\lambda}_t(I - H_s) + (1 - \tilde{\lambda}_t)I$ is also a positive definite matrix.

Appendix F: Lipschitz Continuity of $f(\cdot)$

$f(\cdot)$ satisfies $\|\nabla f(\alpha) - \nabla f(\beta)\| \le L\|\alpha - \beta\|$ for any $\alpha, \beta \in R$ with $L = 2\|B_{as}\|_2^2$

We first proved that $\|I - \tilde{\lambda}_t H_s\|_2 \le 1$. Notice that $\|X\|_2$ refers to the spectrum norm of matrix X. From the definition of the spectrum norm, we know that $\|X\|_2 = \sqrt{\lambda_{\max}(X^T X)}$. Consequently, $\|I - \tilde{\lambda}_t H\|_2 = \sqrt{\lambda_{\max}[(I - \tilde{\lambda}_t H)^2]} = \lambda_{\max}(I - \tilde{\lambda}_t H) = 1 - \lambda_{\min}(\tilde{\lambda}_t H) \le 1$. For any $\tilde{\lambda}_t \in (0, 1)$.

We then know from Appendix D that $\nabla f(\alpha) = -2B_{as}^T(I - \tilde{\lambda}_t H_s)(y_t - B_{as}\alpha) + 2(1 - \tilde{\lambda}_t)B_{as}^T y_{t-1}$ and $\|\nabla f(\alpha) - \nabla f(\beta)\| = \|2B_{as}^T(I - \tilde{\lambda}_t H_s)B_{as}(\alpha - \beta)\| \le \|2B_{as}^T(I - \tilde{\lambda}_t H_s)B_{as}\|_2 \cdot \|\alpha - \beta\| \le L\|\alpha - \beta\|$, in which $L = 2\|B_{as}\|_2^2$. The last equation holds because $\|2B_{as}^T(I - \tilde{\lambda}_t H_s)B_{as}\|_2 \le \|2B_{as}^T\|_2\|I - \tilde{\lambda}_t H_s\|_2\|B_{as}\|_2 \le \|2B_{as}^T\|_2\|B_{as}\|_2 = 2\|B_{as}\|_2^2$.

Appendix G: Solution of $\theta_{a,t}^{(k)}$ in Proximal Gradient Algorithm

It is not hard to show that the proximal gradient method for (10) given by

$$\theta_{a,t}^{(k)} = \underset{\theta_{a,t}}{\operatorname{argmin}} \left\{ f\left(\theta_{a,t}^{(k-1)}\right) + \left\langle \theta_{a,t} - \theta_{a,t}^{(k-1)}, \nabla f\left(\theta_{a,t}^{(k-1)}\right)\right\rangle + \frac{L}{2}\left\|\theta_{a,t} - \theta_{a,t}^{(k-1)}\right\|^2 + \gamma\|\theta_{a,t}\|_1 \right\}$$

has a closed-form solution in each iteration k and can be solved. Since $f(\theta_a) = (y_t - B_{as}\theta_{a,t})^T(I - \tilde{\lambda}_t H_s)(y_t - B_{as}\theta_{a,t}) - 2(1 - \tilde{\lambda}_t)(y_t - B_{as}\theta_{a,t})^T y_{t-1}$.

We know that

$$\begin{aligned}\nabla f(\theta_{a,t}^{(k-1)}) &= -2B_{as}^T(I - \tilde{\lambda}_t H_s)\left(y_t - B_{as}\theta_{a,t}^{(k-1)}\right) \\ &\quad + 2(1 - \tilde{\lambda}_t)B_{as}^T y_{t-1} = -2B_{as}^T\left(y_t - B_{as}\theta_{a,t}^{(k-1)}\right) \\ &\quad + 2B_{as}^T\left((1 - \tilde{\lambda}_t)y_{t-1} + \tilde{\lambda}_t H_s\left(y_t - B_{as}\theta_{a,t}^{(k-1)}\right)\right) \\ &= -2B_{as}^T\left(y_t - B_{as}\theta_{a,t}^{(k-1)} - \mu_t^{(k)}\right).\end{aligned}$$

The last equation holds because of (9).

$$\begin{aligned}\theta_{a,t}^{(k)} &= \underset{\theta_{a,t}}{\operatorname{argmin}} \left\{ \left\langle \theta_{a,t} - \theta_{a,t}^{(k-1)}, \nabla f\left(\theta_{a,t}^{(k-1)}\right)\right\rangle \right. \\ &\quad \left. + \frac{L}{2}\left\|\theta_{a,t} - \theta_{a,t}^{(k-1)}\right\|^2 + \gamma\|\theta_{a,t}\|_1 \right\} \\ &= \underset{\theta_{a,t}}{\operatorname{argmin}} \frac{L}{2}\left\|\theta_{a,t} - \theta_{a,t}^{(k-1)} - \frac{2}{L}B_{as}^T\left(y_t - B_{as}\theta_{a,t}^{(k-1)} - \mu_t^{(k)}\right)\right\|^2 \\ &\quad + \gamma\|\theta_{a,t}\|_1 \Bigg\}.\end{aligned}$$

We know that this can be solved by the soft-thresholding operator as follow: $\theta_{a,t}^{(k)} = S_{\frac{\gamma}{L}}(\theta_{a,t}^{(k-1)} + \frac{2}{L}B_{as}^T(y - B_{as}\theta_{a,t}^{(k-1)} - \mu_t^{(k)}))$, which is exactly (11).

Appendix H: The Limit of The Temporal Projection Matrix For The Static Background

Proposition H.1. The temporal projection matrix H_t in (7) and (5) becomes the average projection matrix $H_t \to \frac{1}{n}1_n 1_n^T$ when $\lambda_t \to \infty$ and $c \to 0$, respectively, where

$$1_n = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

is the column vector of 1.

Proof. To prove this, we look at the following two lemmas □

Lemma H.1. For the roughness minimization model: $H_t = (I + \lambda_t D_t^T D_t)^{-1} \to \frac{1}{n}1_n 1_n^T$ when $\lambda \to \infty$.

Suppose the eigendecomposition of $D^T D$ yields, $D^T D = U\Lambda U^{-1}$. It has been proved in that $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ with $\lambda_i = -2 + 2\cos((i-1)\pi/n)$. This gives that there is only one eigenvalue that equals to 0 as $\lambda_1 = 0$, and $\lambda_i \neq 0$, when $i \ge 2$.

$(I + \lambda\Lambda)^{-1} = \operatorname{diag}(\frac{1}{1+\lambda\lambda_1}, \cdots, \frac{1}{1+\lambda\lambda_n}) \to \operatorname{diag}(1, 0 \ldots 0)$

$$\begin{aligned}H_t &= (I + \lambda_t U\Lambda U^{-1})^{-1} = U^T(I + \lambda_t\Lambda)^{-1}U \\ &\to U^T \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} U = uu^T,\end{aligned}$$

in which u is the first eigenvector of $D^T D$, which corresponds to eigenvalue 0. It is not hard to show that $u = \frac{1}{\sqrt{n}}1_n$, in which

$$1_n = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

because $D^T Du = \frac{1}{\sqrt{n}}D^T D1_n = 0$. Therefore, $H = \frac{1}{\sqrt{n}}1_n\frac{1}{\sqrt{n}}1_n^T = \frac{1}{n}1_n 1_n^T$.

Lemma H.2. For the kernel model: $H_t = K_t(K_t + \lambda_t I)^{-1} \propto 1_n 1_n^T$ when $c \to 0$.

When $c \to \infty$, sine $\kappa(i, j) = \exp(-\frac{(i-j)^2}{2c^2})$,
Therefore, $K_t = 1_n 1_n^T$ and

$$H_t = K_t(K_t + \lambda_t I)^{-1}.$$

From Sherman–Morrison formula we know that

$$\begin{aligned}(K_t + \lambda_t I)^{-1} &= \left(1_n 1_n^T + \lambda_t I\right)^{-1} = \frac{1}{\lambda_t}\left(I + \frac{1}{\lambda_t}1_n 1_n^T\right)^{-1} \\ &= \frac{1}{\lambda_t}\left(I - \frac{1}{\lambda_t}\frac{1_n 1_n^T}{1 + 1_n^T 1_n}\right) \\ &= \frac{1}{\lambda_t}\left(I - \frac{1}{\lambda_t(n+1)}1_n 1_n^T\right).\end{aligned}$$

This gives

$$H = K_t(K_t + \lambda_t I)^{-1} = 1_n 1_n^T\frac{1}{\lambda_t}\left(I - \frac{1}{\lambda_t(n+1)}1_n 1_n^T\right) \propto 1_n 1_n^T.$$

Supplementary Materials

We also added one simulation scenario such that the functional mean is static and doesn't change over time. In this case, we compared the Average Run Length of the proposed method and other benchmark methods, and demonstrated that the proposed method still performs the best due to the accurate estimation of functional mean and sparse clustered background.

References

Babaud, J., Witkin, A. P., Baudin, M., and Duda, R. O. (1986), "Uniqueness of the Gaussian Kernel for Scale-Space Filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8, 26–33. [185]

Bakshi, B. R. (1998), "Multiscale PCA With Application to Multivariate Statistical Process Monitoring," *AIChE Journal*, 44, 1596–1610. [183]

Berlinet, A., and Thomas-Agnan, C. (2011), *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, New York: Springer Science & Business Media. [185]

Chang, S. I., and Yadama, S. (2010), "Statistical Process Control for Monitoring Non-Linear Profiles Using Wavelet Filtering and b-Spline Approximation," *International Journal of Production Research*, 48, 1049–1068. [183]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499. [187]

Hawkins, D. M. (1993), "Regression Adjustment for Variables in Multivariate Quality Control," *Journal of Quality Technology*, 25, 170–182. [187]

Jia, H., Murphey, Y. L., Shi, J., and Chang, T.-S. (2004), "An Intelligent Real-Time Vision System for Surface Defect Detection," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004* (Vol. 3), IEEE, pp. 239–242. [192]

Jin, N., Zhou, S., and Chang, T.-S. (2004), "Identification of Impacting Factors of Surface Defects in Hot Rolling Processes Using Multi-Level Regression Analysis," *Transactions of NAMRI/SME*, 32, 557–564. [181]

Kalpakjian, S., Schmid, S. R., and Kok, C.-W. (2008), *Manufacturing Processesfor Engineering Materials*, Pearson Education India. [191]

Lei, Y., Zhang, Z., and Jin, J. (2010), "Automatic Tonnage Monitoring for Missing Part Detection in Multi-Operation Forging Processes," *Journal of Manufacturing Science and Engineering*, 132, 051010. [181,193,194]

Liu, K., Mei, Y., and Shi, J. (2014), "An Adaptive Sampling Strategy for Online High-Dimensional Process Monitoring," *Technometrics*, 57, 305–319. [183,193]

Liu, R. Y. (1995), "Control Charts for Multivariate Processes," *Journal of the American Statistical Association*, 90, 1380–1387. [183]

Mei, Y. (2010), "Efficient Scalable Schemes for Monitoring a Large Number of Data Streams," *Biometrika*, 97, 419–433. [183,189]

Otsu, N. (1975), "A Threshold Selection Method From Gray-Level Histograms," *Automatica*, 11, 23–27. [188]

Patankar, S. (1980), *Numerical Heat Transfer and Fluid Flow*, Boca Raton, FL: CRC Press. [188]

Paynabar, K., and Jin, J. (2011), "Characterization of Non-Linear Profiles Variations Using Mixed-Effect Models and Wavelets," *IIE Transactions*, 43, 275–290. [183]

Paynabar, K., Jin, J., and Pacella, M. (2013), "Monitoring and Diagnosis of Multichannel Nonlinear Profile Variations Using Uncorrelated Multilinear Principal Component Analysis," *IIE Transactions*, 45, 1235–1247. [193]

Paynabar, K., Qiu, P., and Zou, C. (2015), "A Change Point Approach for Phase-I Analysis in Multivariate Profile Monitoring and Diagnosis," *Technometrics*, 58, 191–204. [183,193,194]

Qiu, P., and Xiang, D. (2014), "Univariate Dynamic Screening System: An Approach for Identifying Individuals With Irregular Longitudinal Behavior," *Technometrics*, 56, 248–260. [183]

Qiu, P., Zou, C., and Wang, Z. (2010), "Nonparametric Profile Monitoring by Mixed Effects Modeling," *Technometrics*, 52, 265–277. [183]

Ruppert, D. (2012), "Selecting the Number of Knots for Penalized Splines," *Journal of Computational and Graphical Statistics*, 11, 735–757. [183]

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001), "A Generalized Representer Theorem," in *Computational Learning Theory*, eds. D. Helmbold and B. Williamson, Berlin/Heidelberg: Springer, pp. 416–426. [185]

Xiang, D., Qiu, P., and Pu, X. (2013), "Nonparametric Regression Analysis of Multivariate Longitudinal Data," *Statistica Sinica*, 23, 769–789. [183]

Xiao, L., Li, Y., and Ruppert, D. (2013), "Fast Bivariate p-Splines: The Sandwich Smoother," *Journal of the Royal Statistical Society*, Series B, 75, 577–599. [183]

Xie, Y., Huang, J., and Willett, R. (2013), "Change-Point Detection for High-Dimensional Time Series With Missing Data," *IEEE Journal of Selected Topics in Signal Processing*, 7, 12–27. [193]

Yan, H., Paynabar, K., and Shi, J. (2015a), "Anomaly Detection in Images With Smooth Background Via Smooth-Sparse Decomposition," *Technometrics*, 59, 102–114. [182,183,184]

Yan, H., Paynabar, K., and Shi, J. (2015b), "Image-Based Process Monitoring Using Low-Rank Tensor Decomposition," *IEEE Transactions on Automation Science and Engineering*, 12, 216–227. [181,183,187]

Zou, C., Jiang, W., and Tsung, F. (2011), "A Lasso-Based Diagnostic Framework for Multivariate Statistical Process Control," *Technometrics*, 53, 297–309. [189]

Zou, C., Jiang, W., Wang, Z., and Zi, X. (2015), "An Efficient Online Monitoring Method for High-Dimensional Data Streams," *Technometrics*, 57, 374–387. [183]

Zou, C., and Qiu, P. (2009), "Multivariate Statistical Process Control Using Lasso," *Journal of the American Statistical Association*, 104, 1586–1596. [187]

Zou, C., Qiu, P., and Hawkins, D. (2009), "Nonparametric Control Chart for Monitoring Profiles Using Change Point Formulation and Adaptive Smoothing," *Statistica Sinica*, 19, 1337–1357. [183]

Zou, C., Tsung, F., and Wang, Z. (2008), "Monitoring Profiles Based on Nonparametric Regression Methods," *Technometrics*, 50, 512–526. [183]