Journal of Process Control 22 (2012) 397–403

Contents lists available at SciVerse ScienceDirect

Journal of Process Control

journal homepage: www.elsevier.com/locate/jprocont

Nonparametric profile monitoring in multi-dimensional data spaces

Ying-Chao Hung[a,*], Wen-Chi Tsai[a], Su-Fen Yang[a], Shih-Chung Chuang[b], Yi-Kuan Tseng[c]

[a] Department of Statistics, National Chengchi University, No. 64, Sec. 2, ZhiNan Rd., Wenshan District, Taipei 11605, Taiwan
[b] Department of Industrial Engineering and Engineering Management, National Tsing Hua University, No. 101, Sec. 2, Kuang-Fu Rd., Hsinchu 30013, Taiwan
[c] Graduate Institute of Statistics, National Central University, No. 300, Jhongda Rd., Jhongli, Taoyuan County 32049, Taiwan

ARTICLE INFO

Article history:
Received 18 January 2011
Received in revised form
28 November 2011
Accepted 19 December 2011
Available online 10 January 2012

Keywords:
Nonparametric profile monitoring
Support Vector Regression
Block bootstrap
Confidence region

ABSTRACT

Profile monitoring has received increasingly attention in a wide range of applications in statistical process control (SPC). In this work, we propose a framework for monitoring nonparametric profiles in multi-dimensional data spaces. The framework has the following important features: (i) a flexible and computationally efficient smoothing technique, called Support Vector Regression, is employed to describe the relationship between the response variable and the explanatory variables; (ii) the usual structural assumptions on the residuals are not required; and (iii) the dependence structure for the within-profile observations is appropriately accommodated. Finally, real AIDS data collected from hospitals in Taiwan are used to illustrate and evaluate our proposed framework.

1. Introduction

In many industrial and medical applications, the quality of a process (or product) can be characterized by a functional relationship between a quality measurement and explanatory variables. Under such circumstances, statistical process control (SPC) aims on monitoring data (called profiles) that represent such a functional relationship, instead of on monitoring a single quality measurement. In this study, we are interested in monitoring profiles for which the underlying functional relationship cannot be reasonably described by a parametric model. An important feature for such data is that they are usually collected over time (i.e. time-dependent). Therefore, it is natural to assume that the within-profile observations are correlated.

Profile monitoring has received increasingly attention over the last decade. A nice overview and extensive discussions can be found in [34,35]. As described in [34], SPC is generally divided into two phases. In Phase I, data are cleaned and the obtained in-control (IC) data are used to estimate certain parameters of the process. In Phase II, the estimated IC parameters are used to detect/monitor changes in the profiles. We highlight some relative works in the following. The early research on profile monitoring mostly focused on simple linear models (see [14,17,23,37] and references there in). Afterwards, correlation within linear profiles [13] and methods based on multiple and polynomial regression models were also explored [15,38]. Recently, methods for nonlinear (NL) profile monitoring have become popular in a wide area of applications due to their model flexibility. For example, three general approaches to NL profile monitoring with dose–response applications were given in [32,33]; the principal component analysis was used for monitoring roundness profiles of manufactured items in [4]; a wavelet-based procedure was used for process fault detection in [19]; a dimension-reduction method was used for monitoring nonlinear profiles in [7]; nonparametric regression methods were used in [39]; just to name a few.

Traditional nonlinear profile monitoring methods have an unrealistic assumption that the within-profile data are independent. On the other hand, methods that incorporate the correlation structure into analysis are rather limited. Two exceptional works can be found in [12,26], of which the former incorporates both the within-profile correlation and the correlation structure of errors via nonlinear mixed (parametric) models, and the later describes the within-profile correlation via nonparametric mixed-effects models. These two approaches are novel from a theoretical viewpoint, but both require a certain amount of computation for real implementation. Our goal here is to provide a practical and easy-to-implement framework for monitoring nonparametric profiles, especially for data in multi-dimensional spaces. Specifically, based on the observed in-control (IC) profiles we wish to establish an adequate confidence region for the underlying functional

* Corresponding author. Tel.: +886 2 29387115; fax: +886 2 29398024.
E-mail addresses: hungy@nccu.edu.tw (Y.-C. Hung), wctsai@nccu.edu.tw (W.-C. Tsai), yang@nccu.edu.tw (S.-F. Yang), scchuang@ie.nthu.edu.tw (S.-C. Chuang), tsengyk@ncu.edu.tw (Y.-K. Tseng).

0959-1524/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jprocont.2011.12.009

relationship without requiring strong model assumptions. This confidence region can then serve as a control chart for Phase II process monitoring.

The proposed framework is mainly divided into five steps. In Step 1, an adequate Support Vector Regression (SVR) model is fitted to each of IC profiles. In Step 2, the moving block bootstrap method (MBB) is used to generate correlated samples for each IC profile. In Step 3, the SVR model is fitted to each of the bootstrap sample and its corresponding "surface depth" is calculated. In Step 4, the SVR models with smaller surface depths are removed and, the resulting confidence regions of all profiles are pooled so as to obtain a simultaneous confidence region for the underlying functional relationship. In Step 5, a future profile is monitored online based on the obtained simultaneous confidence region and a simple data matching process. The remaining of this study is organized as follows. In Section 2, the framework for nonparametric profile monitoring in multi-dimensional data spaces is introduced. In Section 3, the proposed framework is illustrated by using real AIDS data collected from hospitals in Taiwan. In addition, numerical results show that our proposed framework performs well by testing a moderate number of in-control (IC) and out-of-control (OC) profiles. Some concluding remarks are drawn in Section 4.

2. Nonparametric profile monitoring in multi-dimensional data spaces

Here we consider a general formulation for the problem with multi-dimensional data structures. Suppose there are M independent profiles obtained from a typical in-control (IC) process and the ith profile has n_i observations, $i = 1, \ldots, M$. Let y_{ij} be the measurement of interest for the jth observation in the ith profile, with $\mathbf{x}_{ij} = (x_{ij}^1, \ldots, x_{ij}^K)$ be the vector of K corresponding explanatory variables, $j = 1, \ldots, n_i$ for each $i = 1, \ldots, M$. With this assumption, it is clear that the data range for each profile can be different. Suppose the underlying IC model is denoted by

$$y_{ij} = f(\mathbf{x}_{ij}) + \varepsilon_{ij}, \quad i = 1, \ldots, M, \ j = 1, \ldots, n_i, \tag{1}$$

where $f(\cdot)$ is a general function with some degree of smoothness, and ε_{ij} are associated error terms from some unknown distribution. It should be mentioned that for time-series profiles, index j of the ith profile then corresponds to a specific point in time t_{ij} at which y_{ij} and the corresponding explanatory variables $x_{ij}^1, \ldots, x_{ij}^K$ are observed. Further, for modeling flexibility we do not place any structural assumptions on the errors ε_{ij} (such as i.i.d. or normal distribution). This relaxed assumption is suitable for many real-life data that present the characteristics of non-normality and correlation over time.

Our goal here is to propose an easy-to-implement framework for constructing an overall confidence region for the functional relationship $f(\mathbf{x})$ based on the observed IC profiles. The proposed framework is sequential and outlined as follows: (i) an adequate Support Vector Regression (SVR) model is fitted to each profile; (ii) the block bootstrap method is used to resample from each profile data; (iii) the confidence region for each profile is constructed based on the bootstrap percentiles of "surface depths"; (iv) the simultaneous confidence region for the desired function f is obtained by pooling the confidence regions for all IC profiles; and (v) an online monitoring scheme is introduced by utilizing the obtained simultaneous confidence region and a data matching process.

2.1. Fitting Support Vector Regression models

In order to simplify the notations, we denote the data of a particular profile by $\{(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)\}$, where $\mathbf{x}_j = (x_j^1, \ldots, x_j^K)$ is a K-dimensional vector of explanatory variables and $y_j = y(\mathbf{x}_j)$ is the corresponding output measure of interest. The first step of our guide is to obtain a smooth function $\hat{f}(\mathbf{x})$ that represents well the relationship between y_j and $\mathbf{x}_j$ for each profile so that the underlying function $f(x)$ can be estimated accordingly in later stages. Note that one of the most popular and powerful techniques in nonparametric regression originates from the framework of statistical learning theory, called Support Vector Regression (SVR) [1,5,8,27,30]. The ideas of SVR are summarized as follows. In ε-SVR, the goal is to find a function $f(\mathbf{x})$ that has at most ε deviation from the actually observed outputs y_j for all the training data, and at the same time, is as flat as possible. For simple linear functions $f(\mathbf{x}) = \langle \omega, \mathbf{x} \rangle + b$, this corresponds to finding the solution of the following optimization problem:

$$\begin{aligned} &\textit{minimize} \quad \frac{1}{2}\|\omega\|^2 + C\sum_{j=1}^{n}(\xi_j + \xi_j^*) \\ &\text{subject to} \quad \begin{cases} y_j - \langle \omega, \mathbf{x}_j \rangle - b \le \varepsilon + \xi_j \\ \langle \omega, \mathbf{x}_j \rangle + b - y_j \le \varepsilon + \xi_j^* \\ \xi_j, \xi_j^* \ge 0 \end{cases} \end{aligned} \tag{2}$$

Note that the l_2-norm $\|\omega\|^2$ takes into account the flatness of function $f(\mathbf{x})$, $\sum_{j=1}^{n}(\xi_j + \xi_j^*)$ is the amount up to which deviations lager than ε are tolerated, and $C > 0$ is the trade off between both [31]. To achieve nonlinearity, the SVR algorithm finds the optimal solution of f in a high dimensional feature space (or Hilbert space) $\mathcal{H}$ using a mapping $\Phi : D \to \mathcal{H}$. With this mapping, it is shown that the optimization solution depends on the data merely through inner products in $\mathcal{H}$, that is, on functions of the form $\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$. Hence, a computationally cheaper way is to use a kernel function $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$ instead of $\Phi(\cdot)$ explicitly.

A popular choice of the kernel function is the Gaussian kernel (or radial basis function), which has the form that

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{ \frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2} \right\}. \tag{3}$$

From the viewpoint of implementation, the Gaussian kernel has the following two advantages: (i) it can easily handle nonlinear models by mapping data into infinite-dimensional spaces; and (ii) it has relatively low complexity for model selection (since the model has only two unknown parameters C and σ^2). Therefore, it is particularly suitable for exploring high-dimensional data structures. In practice, (C, σ^2) can be chosen by performing a grid search in R_+^2 and utilizing the idea of cross-validation (CV) so as to minimize the "mean square prediction error" [10]. Specifically, one can superimpose a reasonable number of grids over a pre-selected region $(0, a] \times (0, b]$ in R_+^2. Let us denote the set of grids by $\mathcal{G}$, then for each grid point $g \in \mathcal{G}$, the leave-one-out CV score is defined as the following mean squared error of prediction (MSEP):

$$\text{MSEP}(g) = \frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{f}_{(-j)}(\mathbf{x}_j))^2 \tag{4}$$

where $\hat{f}_{(-j)}$ is the obtained SVR model (with (C, σ^2) chosen to be g) by removing the jth observation $(y_j, \mathbf{x}_j)$. After calculating the MSEP for each grid point $g \in \mathcal{G}$, the optimal choice of (C, σ^2) is given by

$$g^* = \underset{g \in \mathcal{G}}{\arg\min}\, \text{MSEP}(g). \tag{5}$$

Remark: For other alternative approaches of choosing (C, σ^2) in SVR with the Gaussian kernel, the readers can refer to [11,16].

2.2. Block bootstrap resampling for dependent data

Suppose now an adequate SVR model $\hat{f}_i(\mathbf{x})$ for each profile i is obtained, the residuals are then given by $e_{ij} = y_{ij} - \hat{f}_i(\mathbf{x}_{ij})$, where

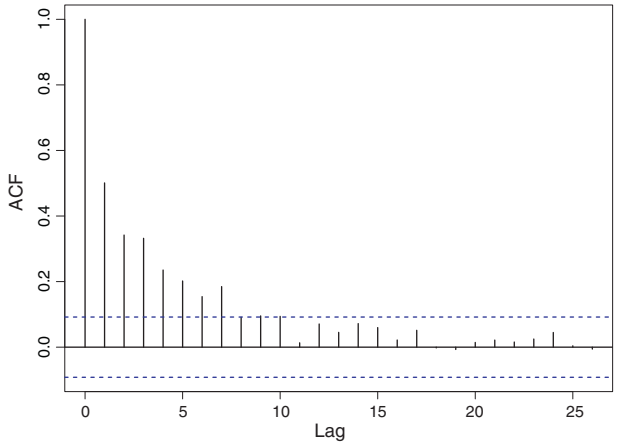

Fig. 1. An illustration of ACF for a time series with size $n = 460$.

$j = 1, \ldots, n_i$. The next step is to construct the confidence region for $f_i(\mathbf{x})$ within the data range of each profile i based on all e_{ij}. This can be done naturally by a *bootstrap resampling* procedure. However, since we assume residuals $e_{i1}, \ldots, e_{in_i}$ can be dependent for each profile i, it is important that the resampling should be carried out in a way that the dependence structure is well captured. We next introduce a simple and popular bootstrap method for dependent data, called *block bootstrap*. Its basic idea is introduced as follows.

In the block bootstrap, data are divided into several blocks so that the original dependence structure within a block is preserved. A popular version for this type of resampling method is the *moving block bootstrap* (MBB), for which overlapping blocks of the same length are drawn randomly with replacement. It was shown in extensively studies that the MBB outperforms other methods based on subsequent values presenting high correlation in relatively short periods of observations (see [24] and references therein). Further, it does not require specific assumptions on the structure of the data generating process. For the MBB with a block length l, the residuals $e_{i1}, \ldots, e_{in_i}$ of each profile i can be divided into $n_i - l + 1$ blocks, viz, with block 1 being $\{e_{i1}, \ldots, e_{il}\}$, block 2 being $\{e_{i2}, \ldots, e_{i(l+1)}\}$, etc. Therefore, the bootstrap sample for each profile i is generated by the following mechanism:

$$\{y^*_{ij}, \mathbf{x}^*_{ij}\} = \{\hat{f}_i(x_{ij}) + e_{i(t+k)}, \mathbf{x}_{ij}\} \tag{6}$$

where $j = 1, 2, \ldots, n_i$, t is generated independently from a uniform random variable on $\{1, 2, \ldots, n_i - l + 1\}$, and $k = 1, 2, \ldots, l$.

It is noted that the accuracy of the MBB is sensitive to the choice of block length l. In general, blocks of a shorter length can achieve a better approximation of the underlying distribution, but on the other hand they may destroy the structure of medium-range (or long-range) dependence. There are different ways suggested in literature for choosing the optimal block length [2,9,18,20,21]. Among all the methods, two essential ones are called the plug-in methods and the empirical criterion-based methods. However, these two types of methods may not be practical in the sense that the plug-in methods often require a huge amount of work in order to obtain the theoretical expression for the optimal block length, while the criterion-based methods often require a certain amount of computation. In this work, we introduce a simple method for choosing a "suitable" block length for time series data. The idea is based on examining the diagnostic plot of the sample autocorrelation function (ACF). To illustrate, let us look at the ACF of a time series (with sample size $n = 460$) shown in Fig. 1. As can be seen from Fig. 1, the value of ACF has a sharp decay right after lag 7 and becomes insignificant beyond that lag. In this case, the suggested smallest block length is $l = 8$ (since the block length is often chosen to be substantially longer than the dependence length). However, it should be mentioned here, the block length chosen by observing the ACF can be unexpectedly large (e.g. for long-range-dependent data). To overcome this problem, one can utilize another criterion suggested in [20], viz., choosing the block length $l \le \sqrt{n}$. As a result, we suggest the following guideline for choosing the best block length:

Let k^* be the smallest lag such that the ACF is not significant for all lag $k \ge k^*$. The suggested block length is $l^* = \min\{k^*, [\sqrt{n}]\}$, where n is the sample size and the bracket $[\sqrt{n}]$ denotes the greatest integer less than or equal to $\sqrt{n}$.

Remark: Since the glue points break the property of stationarity, the data generated by the MBB are non-stationary. However, this problem can be solved by using a method called *stationary bootstrap* (if the original data are stationary), for which the block length l is randomly selected from a geometric distribution [25]. For other methods that can be used to resample from dependent data (such as subsampling, sieve bootstrap, local bootstrap, wild bootstrap, and Markov bootstrap), the readers can refer to [3,6].

2.3. Simultaneous confidence region based on bootstrap surface depths

To construct the confidence region for the underlying function $f(\mathbf{x})$, we first establish the bootstrap percentile confidence region for each profile and then glue all the resulting confidence regions together. Suppose for each profile i, the bootstrap sampling process is repeated N times so that a collection of N resulting SVR "surfaces" $C_i^N(\mathbf{x}) = \{\hat{f}_i^1(\mathbf{x}), \ldots, \hat{f}_i^N(\mathbf{x})\}$ is obtained, where $\mathbf{x} \in \mathcal{X}_i = \{(x^1, \ldots, x^K) : \min_j\{x_{ij}^k\} \le x^k \le \max_j\{x_{ij}^k\} \text{ for all } k = 1, \ldots, K\}$. The confidence region for the ith profile is obtained by ranking the "surface depths" for all $\hat{f}_i^1(\mathbf{x}), \ldots, \hat{f}_i^N(\mathbf{x})$ in $C_i^N(\mathbf{x})$. Its basic idea, which is motivated by the concept of "curve depth" in a two-dimensional data space [36], is described as follows.

In a two-dimensional data space (x, y), the *curve distance* for a particular bootstrap curve $\hat{f}_i^j(x) \in C_i^N(x)$ with respect to the baseline curve $\hat{f}_i(x)$ (obtained from the original profile data) is defined as

$$d_{ij} = d(\hat{f}_i(x), \hat{f}_i^j(x)) = \int_{\mathcal{X}_i} |\hat{f}_i^j(x) - \hat{f}_i(x)| dx \tag{7}$$

or

$$d_{ij} = d(\hat{f}_i(x), \hat{f}_i^j(x)) = \int_{\mathcal{X}_i} (\hat{f}_i^j(x) - \hat{f}_i(x))^2 dx. \tag{8}$$

The corresponding *curve depth* is then defined as $D_{ij} = (1 + d_{ij})^{-1}$. With this definition, the smaller the curve depth is, the further the curve is located from the benchmark curve. For high-dimensional data spaces, a similar measure called *surface distance* can be defined as

$$s_{ij} = d(\hat{f}_i(\mathbf{x}), \hat{f}_i^j(\mathbf{x})) = \int_{\mathcal{X}_i} |\hat{f}_i^j(\mathbf{x}) - \hat{f}_i(\mathbf{x})| d\mathbf{x} \tag{9}$$

or

$$s_{ij} = d(\hat{f}_i(\mathbf{x}), \hat{f}_i^j(\mathbf{x})) = \int_{\mathcal{X}_i} (\hat{f}_i^j(\mathbf{x}) - \hat{f}_i(\mathbf{x}))^2 d\mathbf{x}. \tag{10}$$

The corresponding *surface depth* is then defined as $S_{ij} = (1 + s_{ij})^{-1}$. The surface depth has a similar interpretation to the curve depth – the smaller the depth is, the further the surface is located from the benchmark surface. To obtain the bootstrap percentile confidence region for the ith profile, we can exclude $100\alpha\%$ surfaces with the lowest depths from the collection $C_i^N(\mathbf{x})$.

Now we describe the detailed steps for constructing the overall confidence region of $f(\mathbf{x})$. Let $S_{i(1)} \le S_{i(2)} \le \cdots \le S_{i(N)}$ be the sorted surface depths S_{ij} for all the bootstrap surfaces in $C_i^N(\mathbf{x})$. For any given $0 < \alpha < 1$, define the collection $C_{i,1-\alpha}^N(\mathbf{x}) = \{\hat{f}_i^{(j)}(\mathbf{x}) : \alpha N \le j \le$

$N\}$, where $\mathbf{x} \in \mathcal{D}_i$. The $100(1-\alpha)\%$ bootstrap percentile confidence region for the ith profile is then given by

$$B^N_{i,1-\alpha}(\mathbf{x}) = \{(\mathbf{x}, y) : \text{For each fixed } \mathbf{x}, \min_j \hat{f}^j_i(\mathbf{x}) \le y \le \max_j \hat{f}^j_i(\mathbf{x})\}, \quad (11)$$

where $\hat{f}^j_i(\mathbf{x}) \in C^N_{i,1-\alpha}(\mathbf{x})$ and $\mathbf{x} \in \mathcal{D}_i$, $i = 1, \ldots, M$. Since we assume all the profiles are independent, it is natural to pool all the confidence regions $B^N_{1,1-\alpha}(\mathbf{x})$, $B^N_{2,1-\alpha}(\mathbf{x})$, $\ldots$, $B^N_{M,1-\alpha}(\mathbf{x})$ so as to obtain the confidence region of $f(\mathbf{x})$ over the entire data space. Thus, the simultaneous confidence region of $f(\mathbf{x})$ is given by

$$B_{1-\alpha}(\mathbf{x}) = \{(\mathbf{x}, y) : \text{For each fixed } \mathbf{x}, \min_{i,j} \hat{f}^j_i(\mathbf{x}) \le y \le \max_{i,j} \hat{f}^j_i(\mathbf{x})\}, \quad (12)$$

where $\hat{f}^j_i(\mathbf{x}) \in C^N_{i,1-\alpha}(\mathbf{x})$ and $\mathbf{x} \in \bigcup_{i=1}^M \mathcal{D}_i$.

Remark: Since the consideration of squared distance is more sensitive to surfaces lying outwardly with respect to the baseline surface (i.e. outliers), the confidence region based on the surface distance defined by Eq. (10) is, intuitively, thinner than that based on the surface distance defined by Eq. (9).

Remark: It is noted that the resulting confidence region $B^N_{i,1-\alpha}(\mathbf{x})$ represents a volume in a $(K+1)$-dimensional space. In practice, it can be approximated by a set of "fine grids" superimposed over the data space.

2.4. Online profile monitoring

Note that the confidence region $B_{1-\alpha}(\mathbf{x})$ can serve as a control chart for online monitoring of nonparametric profiles in Phase II. To do this, for a particular new profile we need to obtain an SVR model at any point in time t based on the current observations. If the obtained SVR model, denoted by $\hat{f}(\mathbf{x}_t)$, falls completely within the confidence region (i.e. $\hat{f}(\mathbf{x}_t) \subset B_{1-\alpha}(\mathbf{x})$ for all $\mathbf{x}_t \in \bigcup_{i=1}^M \mathcal{D}_i$), then the profile is considered as "in-control" at time t. Otherwise, it is considered as "out-of-control". However, such an online monitoring scheme can induce huge amounts of computation since a new grid search for (C, σ^2) is necessary for obtaining the desired SVR model (see Section 2.1) when a new data point is observed and included in analysis. To overcome this computational issue, we suggest choosing (C, σ^2) based on the concept of "data matching". Its basic idea is introduced as follows.

Let $(y_1, \mathbf{x}_1), \ldots, (y_{n(t)}, \mathbf{x}_{n(t)})$ be the observed profile data at any given point in time t. The IC profile that "best matches" the observed data can be simply given by

$$i^*(t) = arg \min_{i=1,\ldots,M} \sum_{j=1}^{n(t)} (y_j - \hat{f}_i(\mathbf{x}_j))^2, \quad (13)$$

where $\hat{f}_i$ is the SVR model for the ith IC profile obtained from the earlier stage. Let us denote the best choice of (C, σ^2) (based on a grid search) for establishing $\hat{f}_{i^*(t)}$ by $(C_{i^*(t)}, \sigma^2_{i^*(t)})$. Since the $i^*(t)$th IC profile best matches the observed data based on Eq. (13), $(C_{i^*(t)}, \sigma^2_{i^*(t)})$ can serve as a good surrogate for fitting the desired SVR model $\hat{f}(\mathbf{x}_t)$. It should be mentioned here that, the proposed data matching procedure for finding $(C_{i^*(t)}, \sigma^2_{i^*(t)})$ is computationally much cheaper than performing consecutive grid searches over time.

2.5. Algorithm for profile monitoring

For implementation purpose, we summarize the proposed framework of nonparametric profile monitoring in the following algorithm. Note that Step 1–4 basically constitute Phase I of the monitoring scheme, while Step 5 focuses on Phase II studies.

Step 1: Obtain the SVR model $\hat{f}_i(\mathbf{x})$ for each profile i based on the procedure introduced in Section 2.1. If possible, remove potential outliers that may be caused by measurement error.

Step 2: Generate N bootstrap samples $\{y^*_{ij}, \mathbf{x}^*_{ij}\}$ for each of retained profile i based on the procedure introduced in Section 2.2.

Step 3: For each profile i, obtain the SVR model $\hat{f}^j_i(\mathbf{x})$ based on each generated bootstrap sample and compute its surface depth S_{ij} with respect to the benchmark surface $\hat{f}_i(\mathbf{x})$. Obtain the sorted surface depths $S_{i(j)}$ and identify the collection of surfaces $C^N_{i,1-\alpha}(\mathbf{x})$ for a given $0 < \alpha < 1$.

Step 4: Construct the $100(1-\alpha)\%$ bootstrap percentile confidence region $B^N_{i,1-\alpha}(\mathbf{x})$ for each profile i by using Eq. (11). Obtain the simultaneous confidence region $B_{1-\alpha}(\mathbf{x})$ for $f(\mathbf{x})$ by using Eq. (12).

Step 5: Monitor a future profile online based on the simultaneous confidence region obtained in Step 4. At any given point in time t, perform a data matching procedure based on Eq. (13) so that the desired SVR model $\hat{f}(\mathbf{x}_t)$ can be established by choosing $(C, \sigma^2) = (C_{i^*(t)}, \sigma^2_{i^*(t)})$. If $\hat{f}(\mathbf{x}_t) \subset B_{1-\alpha}(\mathbf{x})$ for all $\mathbf{x}_t \in \bigcup_{i=1}^M \mathcal{D}_i$, then the profile is characterized as "in-control" at time t; otherwise it is characterized as "out-of-control".

3. A real example

In this section, we illustrate the proposed framework on real AIDS data collected from hospitals in Taiwan. We first introduce the data set, some numerical results are presented afterward.

3.1. Introduction to the data set

The AIDS cohort data, which were collected between January 1990 and January 2003, include the information of clinical, biochemical, serologic, and histologic parameters of 1054 HIV-infected patients in Taiwan. All patients were advised to return to hospital every three or four months for a follow-up diagnosis. The primary goal of collecting such a data set is to evaluate the efficacy of the highly active antiretroviral therapy (HAART), which consists of at least three anti-HIV drugs (see [29] for a detailed description of this data set). To illustrate our proposed framework, here we select 4 important variables from the data set for analysis, of which one is treated as the dependent variable (Y) and the others are treated as explanatory variables (X_1, X_2, X_3). These selected variables are described as follows.

Y: The CD4 cell count (per cubic millimeter of blood) in log scale.
X_1: The measurement time in years.
X_2: The ratio of CD4 cell count to CD8 cell count.
X_3: The CD4 cell count in percentage.

We divide the patients into two groups. The patients who took the therapy HAART are categorized as the "in-control" (IC) profiles. On the other hand, the patients who did not take the therapy HAART are categorized as the "out-of-control" (OC) profiles. Fig. 2 summarizes all the paired relationships of 20 randomly selected IC profiles (patients) between Y and X_1, X_2, X_3. As can be seen from Fig. 2, there exists a very clear functional relationship between Y and the explanatory variables X_2 and X_3. It should be pointed here that, due to a fairly large number of censored observations, the confidence region for the underlying relationship between Y and $X_1, \ldots, X_3$ will be constructed based on merely these 20 IC profiles. The rest of the profile data are left for testing the performance of the proposed framework shown later in Section 3.3.

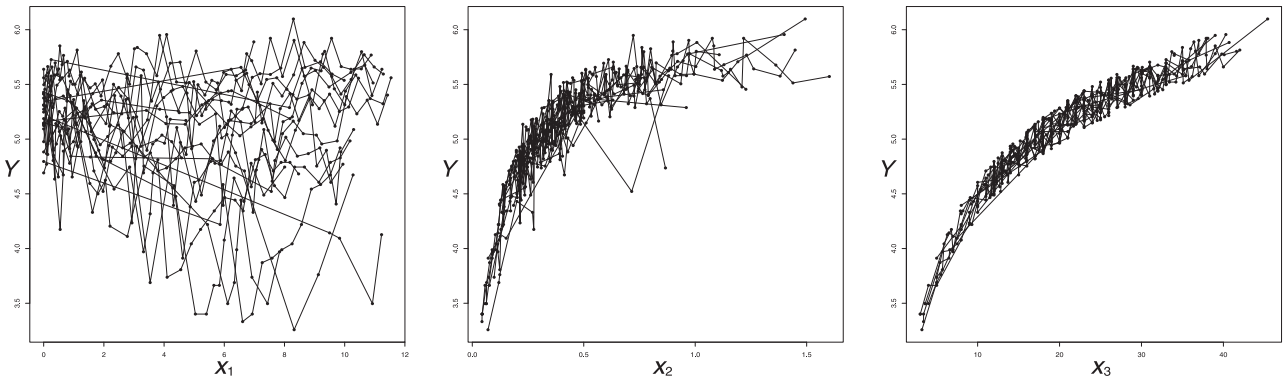

Fig. 2. The paired scatter diagrams for the 20 selected IC profiles.

3.2. Construction of the simultaneous confidence region

We start with fitting the SVR model for each of the 20 IC profiles. For each fitted SVR model $\hat{f}_i(\mathbf{x})$, the optimal choice of (C, σ^2) is based on a grid search over a pre-specified region in R_2^+. To illustrate, the contour plot of the MSEP for the 11th IC profile over the region $(0, 1.4] \times (0, 0.2]$ is shown in Fig. 3. As can be seen, the optimal choice of (C, σ^2) is around (0.43, 0.15), for which the fitted SVR model has the minimum value of MSEP. For other IC profiles, the best fitted SVR models are selected in a similar fashion.

We next examine the plot of ACF for the residuals based on the best fitted $\hat{f}_i(\mathbf{x})$ for each of the 20 IC profiles. The goal of this step is to choose a suitable block size so as to conduct an adequate bootstrap sampling procedure for each IC profile. Fig. 4 shows the resulting ACF for the 9th and 19th IC profile. As can be seen from Fig. 4, the ACF for the 9th and the 19th IC profile becomes insignificant after lag 3 and lag 4, respectively. Therefore, the suggested block sizes for these two profiles are 4 and 5, respectively. After examining all the diagnostic plots, the suggested block sizes for the remaining 18 IC profiles are: 3 for the 3rd IC profile, 2 for the 13th IC profile, and 1 for all the other IC profiles.

Remark: It is noted that a small sample size can easily result in "insignificance" of the ACF (for this example $n_i \leq 48$ for all $i = 1, \ldots, 20$). This might be the reason why a rather small block size is suggested for each of the 20 IC profiles.

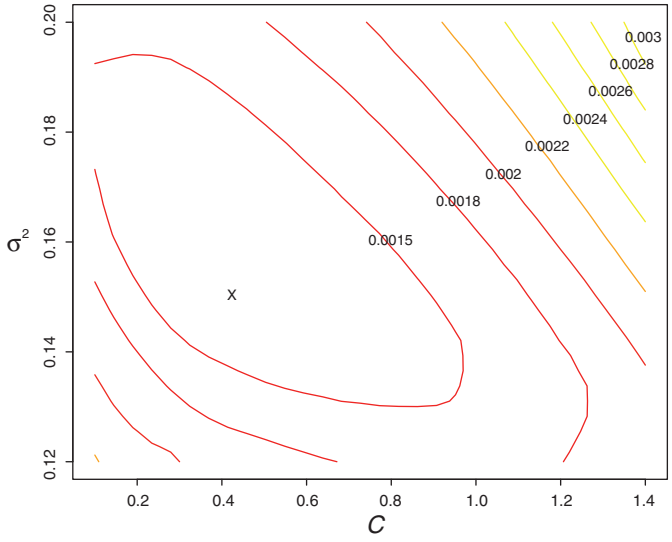

Fig. 3. An illustration of the MSEP contour lines with respect to the choice of (C, σ^2) in the SVR model fitted to the 11th IC profile.

Based on the obtained block sizes, the MBB method is then used to resample from the residuals of each profile. Thus, for each generated bootstrap sample we can obtain a fitted SVR model that represents the relationship between Y and X_1, X_2, X_3. To construct the 95% bootstrap percentile confidence region $B^N_{i,0.95}(\mathbf{x})$ for each profile i, the MBB method is repeated 10^4 times (i.e. $N = 10^4$) so that 10^4 SVR models (i.e. $\hat{f}_i^j(\mathbf{x}), j = 1, \ldots, 10^4$) are constructed. By deleting 5% of models based on the sorted surface depths (where s_{ij} are computed using Eq. (10)) for each IC profile, the 95% simultaneous confidence region $B_{0.95}(\mathbf{x})$ is then obtained by Eq. (12).

For visualization purpose, the resulting confidence region $B_{0.95}(\mathbf{x})$ is projected respectively onto the axes of the three explanatory variables X_1, X_2, and X_3. The result is given in Fig. 5. As can be seen from Fig. 5, the confidence region projected onto the first explanatory variable X_1 reveals to be wider (in general) than that projected onto the axis of X_2 and X_3. The result clearly agrees with Fig. 2 in which a rather larger data variability in X_1 is presented.

3.3. Performance evaluation

In this section we evaluate the effectiveness of the proposed framework by comparing with two benchmark methods in terms of Type I and Type II errors. The first method to be compared is the nonparametric regression approach introduced by Zou et al. [39], wherein a standard Gaussian kernel function is selected to construct the local linear smoother and the error terms e_{ij} are assumed to be i.i.d. normal random variables (thus the within-profile correlation is not taken into account). The second method to be compared is the nonlinear mixed (NLM) models introduced by Jensen and Birch [12], wherein a logistic model is selected and correlation within the profile is also incorporated. To conduct such a comparison, 50 untested IC and OC profiles (patients) with more than five observations are selected from the original data set. The resulting Type I and Type II errors are given in Table 1.

Let $\hat{f}(\mathbf{x})$ be the obtained SVR model for a particular profile to be tested. The Type I and Type II errors are defined as

$$\text{Type I error} = P(\hat{f}(\mathbf{x}) \not\subset B_{0.95}(\mathbf{x}) | \text{ the profile is IC}) \quad (14)$$

and

$$\text{Type II error} = P(\hat{f}(\mathbf{x}) \subset B_{0.95}(\mathbf{x}) | \text{ the profile is OC}), \quad (15)$$

Table 1
The Type I and Type II errors of classifying the IC and OC profiles for three different methods.

Errors	Our method	The method by Zou et al.	The method by Jensen and Birch
Type I error	0.12	0.54	0.00
Type II error	0.40	0.32	0.92

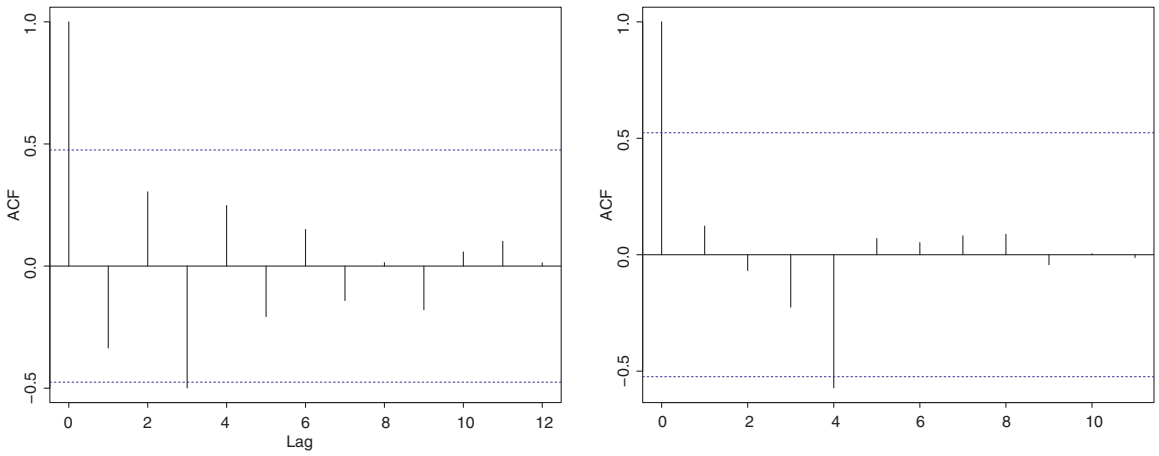

Fig. 4. The ACF of the residuals obtained by fitting the SVR model to the 9th (left panel) and the 19th IC profile (right panel).

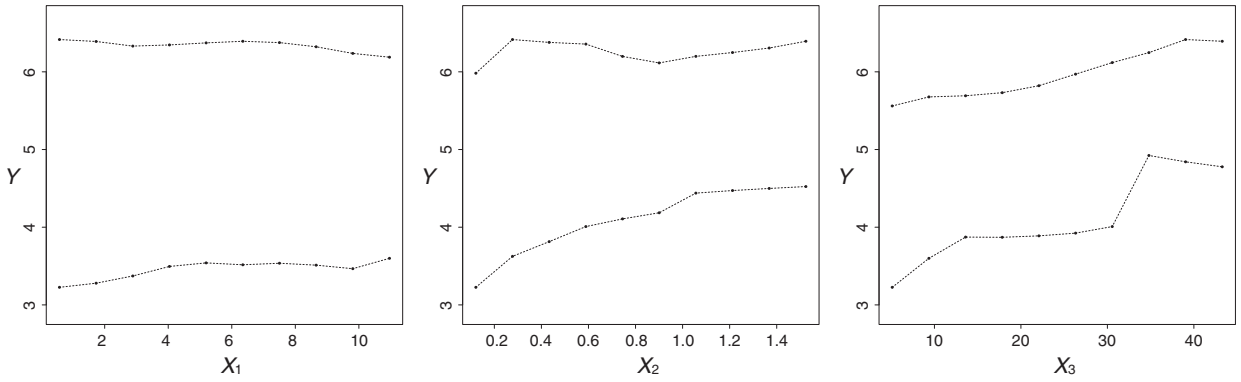

Fig. 5. The resulting confidence region $B_{0.95}(\mathbf{x})$ projected onto the axes of X_1, X_2, and X_3.

which can be simply estimated by the proportion of the obtained IC/OC SVR models falling outside/inside the confidence region. The numerical results in Table 1 reveal several interesting findings. First, the method by Jensen and Birch perfectly classifies the IC profiles (Type I error = 0%) but misclassifies almost all the OC profiles (Type II error = 92%). This may be due to the fact that the method incorporates inadequate correlation structures into analysis so that a rather "conservative" (large) confidence region is obtained. On the other hand, the method by Zou et al. misclassifies more than a half of the IC profiles (Type I error = 54%) but performs relatively well in classifying the OC profiles (Type II error = 32%). This may be due to the fact that the method totally ignores the correlation structure within the profile so that a rather "tight" (small) confidence region is obtained. Finally, it is worth noting that our method classifies well the IC profiles (Type I error = 12%) and performs fairly well in classifying the OC profiles (Type II error = 40%). Although the number of tested profiles is not particularly large, the result strongly supports that our proposed method is effective in classifying both the IC and OC profiles.

4. Conclusions

We proposed an easy-to-implement framework for monitoring nonparametric profiles in multi-dimensional data spaces. The framework is sequential and mainly comprised of five steps. In Step 1, an adequate support vector regression (SVR) model is fitted to each IC profile. This modeling technique has the advantages that (i) it does not require any structural assumptions on data (i.e. it is data-driven); (ii) it can easily handle the data in high-dimensional spaces; and (iii) it is computationally efficient. In Step 2, the moving block bootstrap (MBB) method is used to generate dependent samples for each profile. Such a sampling technique shares the same intuition with the mixed models introduced in [12,22,28] – it incorporates fairly well the correlation structure of errors within each IC profile. In Step 3, the SVR model is fitted to each of the bootstrap sample and its corresponding surface depth is calculated. In Step 4, the SVR models with smaller surface depths are removed and, the resulting confidence regions of all profiles are pooled so as to obtain an overall confidence region for the underlying functional relationship. In Step 5, an online monitoring scheme is introduced based on the obtained overall confidence region and a data matching process. Numerical results show that, compared to other two benchmark methods, our proposed framework is effective in classifying both the in-control and out-of-control profiles.

Here we highlight some potential problems for future research studies. First, in practice any nonparametric modeling technique with adequately chosen tuning parameters can be applied in the Step 1 of our proposed framework. However, one needs to take into account the computational cost (especially when the data dimension or the number of profiles becomes large) and how to best compare the confidence regions obtained from different modeling techniques. Second, it is possible to develop the online control chart for monitoring nonparametric profiles in real time by extending the ideas introduced in this work. For example, instead of monitoring the functional relationship, one can possibly establish a control chart for monitoring the residuals. However, how to best incorporate the dependence structure of the observed residuals into the development of such a control chart needs to be further investigated. Finally, incorporating common-cause variation between profiles into the control chart scheme is sometimes

necessary in real applications. This is obviously a more challenging task for purely nonparametric monitoring techniques since one needs to take into account both the intra-profile and inter-profile correlations.

References

[1] B.E. Boser, I.M. Guyon, V.N. Vapnik, A training algorithm for optimal margin classifiers, in: The 5th Annual ACM Workshop on COLT, 1992, pp. 144–152.
[2] P. Bühlmann, H.R. Künsch, Block length selection in the bootstrap for time series, Comput. Stat. Data Anal. 31 (1999) 295–310.
[3] M.R. Chernick, Bootstrap Methods, A Practitioner's Guide, Wiley Series in Probability and Statistics, 1999.
[4] B.M. Colosimo, M. Pacella, On the use of principal component analysis to identify systematic patterns in roundness profiles, Qual. Reliab. Eng. Int. 23 (2007) 707–725.
[5] C. Cortes, V.N. Vapnik, Support vector networks, Mach. Learn. 20 (1995) 273–297.
[6] A.C. Davison, D. Hinkley, Bootstrap Methods and Their Application, 8th ed., Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge, 2006.
[7] Y. Ding, L. Zeng, S. Zhou, Phase I analysis for monitoring nonlinear profiles in manufacturing processes, J. Qual. Technol. 38 (2006) 199–216.
[8] I.M. Guyon, B.E. Boser, V.N. Vapnik, Automatic capacity tuning of very large VC-dimension classifiers, Adv. Neural Inform. Process. Syst. 5 (1993) 147–155.
[9] P. Hall, J.L. Horowitz, B.-Y. Jing, On blocking rules for the bootstrap with dependent data, Biometrika 82 (1995) 561–574.
[10] C.W. Hsu, C.C. Chang, C.J. Lin, A practical guide to support vector classification, Technical Report CWH03a, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, 2003.
[11] C.M. Huang, Y.J. Lee, D.K.J. Lin, S.Y. Huang, Model selection for support vector machines via uniform design, Comput. Stat. Data Anal. 52 (2007) 335–346.
[12] W.A. Jensen, J.B. Birch, Profile monitoring via nonlinear mixed models, J. Qual. Technol. 41 (2009) 18–34.
[13] W.A. Jensen, J.B. Birch, W.H. Woodall, Monitoring correlation within linear profiles using mixed models, J. Qual. Technol. 40 (2008) 167–183.
[14] L. Kang, S.L. Albin, On-line monitoring when the process yields a linear profile, J. Qual. Technol. 32 (2000) 418–426.
[15] R.B. Kazemzadeh, R. Noorossana, A. Amiri, Phase I monitoring of polynomial profiles, Comm. Stat. Theor. Methods 37 (2008) 1671–1686.
[16] S.S. Keerthi, C.-J. Lin, Asymptotic behaviors of support vector machines with Gaussian kernel, Neural Comput. 15 (2003) 1667–1689.
[17] K. Kim, M.A. Mahmoud, W.H. Woodall, On the monitoring of linear profiles, J. Qual. Technol. 35 (2003) 317–328.
[18] H.R. Künsch, The jackknife and the bootstrap for general stationary observations, Ann. Stat. 17 (1989) 1217–1241.
[19] E.K. Lada, J.-C. Lu, J.R. Wilson, A wavelet-based procedure for process fault detection, IEEE Trans. Semicond. Manuf. 15 (2002) 79–90.
[20] S.N. Lahiri, Theoretical comparisons of block bootstrap methods, Ann. Stat. 27 (1999) 386–404.
[21] S.N. Lahiri, K. Furukawa, Y.-D. Lee, A nonparametric plug-in rule for selecting optimal block lengths for block bootstrap methods, Stat. Methodol. 4 (2007) 292–321.
[22] R.C. Little, G.A. Milliken, W.W. Stroup, R.D. Wolfinger, SAS System for Mixed Models, SAS Institute Inc., Cary, NC, 1996.
[23] M.A. Mahmoud, W.H. Woodall, Phase I analysis of linear profiles with calibration applications, Technometrics 46 (2004) 380–391.
[24] S. Mignani, R. Rose, Markov chain monte carlo in statistical mechanics: the problem of accuracy, Technometrics 43 (2001) 347–355.
[25] D.N. Politis, J.P. Romano, The stationary bootstrap, J. Am. Stat. Assoc. 89 (1994) 1303–1313.
[26] P. Qiu, C. Zou, Z. Wang, Nonparametric profile monitoring by mixed effects modeling, Technometrics 52 (2010) 265–277.
[27] B. Schölkopf, Support Vector Learning, R. Oldenbourg Verlag, Munich, 1997.
[28] O. Schabenberger, F.J. Pierce, Contemporary Statistical Models for the Plant and Soil Sciences, CRC Press, Boca Raton, Florida, 2002.
[29] K.N. Shu, Y.K. Tseng, A semiparametric extended hazard model with time-dependent covariates, in: Proceedings of the Joint Statistical Meetings, 2009, pp. 831-843.
[30] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[31] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[32] J.D. Williams, J.B. Birch, W.H. Woodall, N.M. Ferry, Statistical monitoring of heteroscedastic dose–response profiles from high-throughput screening, J. Agric. Biol. Environ. Stat. 12 (2007) 216–235.
[33] J.D. Williams, W.H. Woodall, J.B. Birch, Statistical Monitoring of Nonlinear Product and Process Quality Profiles, Qual. Reliab. Eng. Int. 23 (2007) 925–941.
[34] W.H. Woodall, D.J. Spitzner, D.C. Montgomery, S. Gupta, Using control charts to monitor process and product quality profiles, J. Qual. Technol. 36 (2004) 309–320.
[35] W.H. Woodall, Current research on profile monitoring, Revista Producão 17 (2007) 420–425.
[36] A.B. Yeh, Bootstrap percentile confidence bands based on the concept of curve depth, Commun. Stat. Simulat. Comput. 25 (1996) 905–922.
[37] C. Zou, Y. Zhang, Z. Wang, Control chart based on change-point model for monitoring linear profiles, IIE Trans. 38 (2006) 1093–1103.
[38] C. Zou, F. Tsung, Z. Wang, Monitoring general linear profiles using multivariate EWMA schemes, Technometrics 49 (2007) 395–408.
[39] C. Zou, F. Tsung, Z. Wang, Monitoring profiles based on nonparametric regression methods, Technometrics 50 (2008) 512–526.