Taylor & Francis
Taylor & Francis Group

Statistics
A Journal of Theoretical and Applied Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gsta20

Outlier detection in non-parametric profile monitoring

Tao Wang, Yunlong Wang & Qingpei Zang

To cite this article: Tao Wang, Yunlong Wang & Qingpei Zang (2022) Outlier detection in non-parametric profile monitoring, Statistics, 56:4, 805-822, DOI: 10.1080/02331888.2022.2085707

To link to this article: https://doi.org/10.1080/02331888.2022.2085707

STATISTICS
2022, VOL. 56, NO. 4, 805–822
https://doi.org/10.1080/02331888.2022.2085707

Taylor & Francis
Taylor & Francis Group

Outlier detection in non-parametric profile monitoring

Tao Wang[a], Yunlong Wang[b] and Qingpei Zang[a]

[a]School of Mathematics and Statistics, Huaiyin Normal University, Huaian, People's Republic of China; [b]School of Data and Artificial Intelligence, Dongbei University of Finance and Economics, Dalian, People's Republic of China

ABSTRACT

The robustness of the various profile monitoring methods that can perform poorly in the presence of outliers has been heavily debated. To contribute to this strand of the literature, we propose a two-stage outlier detection scheme for non-parametric profile monitoring that can control type-I error rates as well as identify outlier profiles. In the first stage, we define an outlier detection measure by using a non-parametric test statistic and extend the well-known least trimmed squares algorithm to find a clean profile set. Then, in the second stage, the detection of outlying profiles is deemed as a hypothesis testing problem, where the thresholding rule is obtained from the asymptotic distribution of the proposed measure. Furthermore, to enhance efficiency, a one-step refinement algorithm is proposed to determine whether a data point is a real outlying profile. Simulation studies show that the proposed procedure can control type-I error rates, while maintaining reasonably high outlier detection power values. Finally, we apply the proposed approach to a real data analysis to demonstrate the effectiveness of our method. Our approach responds to the fact that in the profile monitoring of statistical process control problems, the dimensionality and complexity of the relationship between the response and explanatory variables can increase the possibility that a profile is outlying and such outlying profiles may influence the data analysis markedly.

ARTICLE HISTORY
Received 27 September 2020
Accepted 31 May 2022

KEYWORDS
Asymptotic normality; nonparametric regression; nonparametric test statistic; outlier detection; profile monitoring

2020 MATHEMATICS SUBJECT CLASSIFICATIONS
62G86; 62H30; 62G10

1. Introduction

In certain industrial applications, the quality of a process can be characterized by the functional relationship between the response and explanatory variables, which is called profile or functional data. In the field of statistical process control, profile monitoring uses statistical methods to check the stability of functional relationships over time (e.g., [1–7]). In the first-stage analysis of profiles, the presence of outliers may seriously hinder modelling the functional curve and accordingly the properties of control charts [8]. Therefore, in the profile monitoring process, any outliers among a set of complex profiles should be identified and eliminated from the corresponding profile data set.

CONTACT Tao Wang skywangtao2007@126.com Mathematical and Statistics, Huaiyin Normal University, 100 Changjiang West Road, Huaian 223300, People's Republic of China

The robustness of the various profile monitoring methods that can perform poorly in the presence of outliers has been heavily debated and many statisticians have studied the problem of detecting outlying profile data. Zhang and Albin treated profiles as high-dimensional vectors and proposed the χ^2 control chart method to detect outlying profiles [9]. Yu et al. [10] proposed an outlier detection procedure based on functional principal component analysis and demonstrated its use for profile monitoring; however, because this procedure uses single-case diagnostics, it may suffer from masking effects in the presence of multiple outliers [11]. Zou et al. [12] pointed out that the χ^2 control chart method may fail to detect multiple true outlying profiles and proposed an effective procedure from the viewpoint of penalized regression. Li et al. [13] treated the binary profiles of interest as an integrated high-dimensional vector, from the viewpoint of penalized likelihood, and developed outlier detection procedures based on the group LASSO method and directional information. Ren et al. [14] used projections coupled with the high-breakdown mean function estimator to identify outliers for functional data sets. Abdel-Salam et al. proposed a semi-parametric mixed model procedure for determining unusual profiles in the first-stage analysis [15]. Gomaa and Birch [16] pointed out that the first-stage profile monitoring method of [15] only works for linear mixed models and proposed a new semi-parametric method for modelling complicated non-linear data. Ebrahimi et al. [17] proposed an integrated monitoring and diagnostics approach for large-scale streaming data using principal component analysis.

Before we proceed, we show a simple example to illustrate the outliers in non-linear profiles. The profiles are generated as $y_{ij} = f_a(x_j) + \varepsilon_{ij}$, where $f_a(x_j) = 10 - 20ae^{-ax_j}\sin(\sqrt{4-a^2}x_j)/\sqrt{(4-a^2)} + 10e^{-ax_j}\cos(\sqrt{4-a^2}x_j)$, $\varepsilon_{ij} \sim N(0, \sigma^2)$, and $\sigma = 10$. The 50 profiles presented in Figure 1 (48 non-outlying profiles in gray and two outlying profiles in black) are characterized by the relationship between a response variable, Y, and the explanatory variables, X. For each profile, the explanatory variables can have 100 values, namely, $x_j = 0.08, 0.16, \ldots, 8, j = 1, \ldots, 100$. In Figure 1, the non-outlying profiles have $a = 0.5$ and the outlying profiles have $a = 1.0$; hence, the corresponding regression functions are $f_{0.5}(x)$ and $f_{1.0}(x)$, respectively.

However, while the above-mentioned literature and simple example focus on univariate profile monitoring problems, the explanatory variable may be multivariate in practical applications. Further, as the dimensionality of the explanatory variables increases, the profile data become more complex. These factors may increase both the chance of a profile being outlying as well as its potential impact on the data analysis. Based on the foregoing, this work proposes a non-parametric outlier detection procedure that can control type-I error rates as well as identify outlier profiles in multivariate statistical process control problems. We search for a clean subset that is presumably free of outlying profiles and then measure the outlyingness by using a robust statistic based on the clean subset [18].

First, we define the least trimmed kernel distance (LTKD) by using a non-parametric specification test statistic and then adapt the fast LTKD algorithm by minimizing the kernel-based measure to obtain a clean subset. Second, determining whether a profile data point is an outlier can be regarded as a multiple hypothesis testing problem. We therefore prove the asymptotic distribution of the kernel-based measure and then determine the threshold rule for identifying outlying profiles preliminarily. Furthermore, a one-step refinement algorithm is proposed to control the type-I error more accurately.

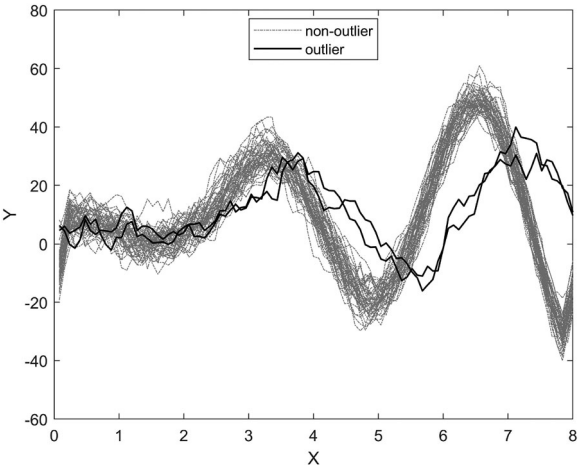

Figure 1. Non-outlying profiles (gray) and outlying profiles (black).

By appropriately choosing parameters and constructing kernel-based non-parametric statistics, our proposed procedure can control type-I error rates and enhance the outlier detection power value. Moreover, the simulation results show that our procedure performs better than other methods in various settings.

The remainder of this article is organized as follows. In Section 2, we first formulate the problem of outliers in profile monitoring and then introduce our proposed outlier detection methodology. The simulation studies of the proposed method are presented in Section 3. In Section 4, we apply the proposed approach to analyze a real data set. We conclude in Section 5. Some of the technical proofs are provided in the Appendix.

2. Methodology

2.1. Model settings

Suppose the observed data set $\{(y_{i1}, \mathrm{x}_{i1}), \ldots, (y_{in}, \mathrm{x}_{in})\}_{i=1}^{m}$ consists of m profiles, where y_{ij} $(j = 1, \ldots, n)$ is the jth response of the ith profile and $\mathrm{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ denotes the corresponding p-dimensional explanatory variable. In the spirit of non-parametric regression, the relationship between the response and explanatory variables can be connected by a Borel measurable function $g(\cdot)$, such that $g_i(\mathrm{x}) = E(y_{ij} \mid \mathrm{x}_{ij} = \mathrm{x})$, $\mathrm{x} \in \mathbb{R}^p$. Owing to many uncertain factors, unexpected and unusual changes may exist in the observed process; hence, some profiles may be outlying. Without loss of generality, we assume the observed m profiles are generated from the following model:

$$y_{ij} = \begin{cases} g_0(\mathrm{x}_{ij}) + \zeta_{ij}, & j = 1, \ldots, n, \text{if } i \notin \mathcal{O}; \\ g_i(\mathrm{x}_{ij}) + \zeta_{ij}, & j = 1, \ldots, n, \text{if } i \in \mathcal{O}, \end{cases} \tag{1}$$

where $\mathcal{O}$ denotes the outlying profile set (a subset of $\{1, \ldots, m\}$), the functional relationship between the response and explanatory variables, $g_0(\cdot)$ and $g_i(\cdot), i \in \mathcal{O}$, respectively denote the non-outlier and outlier profiles, and the error terms ζ_{ij} are independently and identically distributed as $N(0, \sigma^2)$. Any profile with $g_i(\cdot) \neq g_0(\cdot)$ is considered as

outlying in the data set. Suppose that among the m profiles, there are m_0 outlying profiles ($m_0 = |\mathcal{O}|$, where $|\mathcal{O}|$ denotes the cardinality of $\mathcal{O}$). Therefore, it is reasonable to assume that $m_0 < m/2$ since the number of outliers cannot exceed half the whole data set; otherwise, those outliers cannot be considered as suspicious profiles.

2.2. Problem formulation

Under the above settings, suppose the observed m profiles $\{(y_{i1}, \mathrm{x}_{i1}), \ldots, (y_{in}, \mathrm{x}_{in})\}_{i=1}^{m}$ are generated from model (1). The aim is to detect the outlying profiles whose regression function curve is significantly different from that of the non-outlying profiles. Here, to determine whether the ith profile data point is an outlier, it can be regarded as an m hypothesis testing problem with the following null and alternative hypotheses:

$$\mathcal{H}_{0,i} : g_i(\mathrm{x}_{ij}) = g_0(\mathrm{x}_{ij}), \quad \text{vs.} \quad \mathcal{H}_{1,i} : g_i(\mathrm{x}_{ij}) \neq g_0(\mathrm{x}_{ij}) \tag{2}$$

for $i = 1, \ldots, m, j = 1, \ldots, n$.

To this end, a straightforward idea is to construct a test statistic and then perform the hypothesis testing process. Hence, the following non-parametric outlier detection measure is considered:

$$D_{in}(\zeta) = \frac{1}{n(n-1)} \sum_{k=1}^{n} \sum_{l=1, k\neq l}^{n} K_w(\mathrm{x}_{ik} - \mathrm{x}_{il}) \zeta_{ik} \zeta_{il}, \quad \text{for } i = 1, \ldots, m, \tag{3}$$

where $K_w(\cdot) = K(\cdot/w)/w^p$ is a kernel function (the Gaussian kernel is chosen in this work) and $w > 0$ is the bandwidth, which depends on n. The choice of the optimal bandwidth is discussed in Section 2.3. $\zeta_{ik} = y_{ik} - g_i(\mathrm{x}_{ik})$ denotes the bias values between y_{ik} and $g_i(\mathrm{x}_{ik})$. Thus, if the ith profile data point is an outlier, the corresponding measure $D_{in}^2(\zeta)$ is expected to be large.

Under the null hypothesis of Equation (2), similar to the result of [19], we define $\sigma^2(\mathrm{x}) = E(\zeta_{ij}^2 \mid \mathrm{x}_{ij} = \mathrm{x})$, where $p(\mathrm{x})$ is the density function of x_{ij}. Under certain mild conditions, the detection measure $D_{in}(\zeta)$ is asymptotically normal; that is, as $w \to 0$ and $nw \to \infty$:

$$nw^{p/2} D_{in}(\zeta) \xrightarrow{\mathrm{d}} N(0, \Sigma), \tag{4}$$

where $\Sigma = 2 \int K^2(u)\, \mathrm{d}u \cdot \int [\sigma^2(\mathrm{x})]^2 p^2(\mathrm{x})\, \mathrm{dx}$ is the asymptotic variance of $nw^{p/2} D_{in}(\zeta)$. The detailed proof is in the Appendix.

A consistent estimation of $g(\cdot)$ is needed to determine the threshold for the outlying measures. For this, we use the Nadaraya–Watson estimator introduced by Nadaraya [20] and Watson [21]. The ith regression function $g_i(\mathrm{x})$ can be estimated by

$$\hat{g}_i(\mathrm{x}) = E(y_{ij} \mid \mathrm{x}_{ij} = \mathrm{x}) = \frac{\sum_{j=1}^{n} y_{ij} K_w(\mathrm{x}_{ij} - \mathrm{x})}{\sum_{j=1}^{n} K_w(\mathrm{x}_{ij} - \mathrm{x})}, \tag{5}$$

where $K_w(\cdot)$ is the kernel function as in Equation (3). However, the reliable regression function estimator of $g(\cdot)$ may break down in the presence of outliers in the profile dataset. Thus, it is crucial to obtain a reliable estimator $\hat{g}(\mathrm{x}_{ij})$ based on a clean profile dataset.

2.3. Choice of the optimal bandwidth

In this subsection, we discuss the choice of the optimal bandwidth. For each of the m profiles, by using the leave-one-out cross-validation method [22], we obtain the Nadaraya–Watson kernel estimator of the ith regression function by omitting the jth variable, which is defined as

$$\hat{g}_{i,-j}(\mathrm{x}_{ij}) = \frac{\sum_{k\neq j} y_{ik} K_w(\mathrm{x}_{ik} - \mathrm{x}_{ij})}{\sum_{k\neq j} K_w(\mathrm{x}_{ik} - \mathrm{x}_{ij})}, \quad i = 1, \ldots, m,\ j, k = 1, \ldots, n. \tag{6}$$

To choose an appropriate bandwidth, we can use the optimal bandwidth selection methods in non-parametric regression [23–25], Specifically, we minimize the cross-validation score

$$\hat{e}_i(w) = \frac{1}{n}\sum_{j=1}^{n}(y_{ij} - \hat{g}_{i,-j}(\mathrm{x}_{ij}))^2,$$

and the optimal bandwidth of the ith profile is denoted as

$$\hat{w}_i = \arg\min(\hat{e}_i(w)).$$

Finally, we choose the median of the m bandwidths as the final bandwidth, which is defined as

$$\hat{w} = \mathrm{median}\{\hat{w}_1, \ldots, \hat{w}_m\}.$$

Then, the choice of the optimal bandwidth is based on the data.

2.4. Kernel-based outlier detection procedure

Let $\mathcal{H} = \{H \subset \{1, \ldots, m\} : |H| = h\}$ be the collection of all the subsets of size h. Then, the kernel-based outlier detection measure can be constructed as

$$D_{in}(\hat{\zeta}_H) = \frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{l=1,k\neq l}^{n} K_w(\mathrm{x}_{ik} - \mathrm{x}_{il})\hat{\zeta}_{ikH}\hat{\zeta}_{ilH}, \tag{7}$$

for $i = 1, \ldots, m$, where $K_w(\cdot)$ is the kernel function as in Equation (3), $\hat{\zeta}_{ikH} = y_{ik} - \hat{g}_{iH}(\mathrm{x}_{ik})$ is the estimated bias value, and $\hat{g}_{iH}(\mathrm{x}_{ij})$ denotes the consistently estimated regression function based on profiles with the index subset $H \in \mathcal{H}$.

Intuitively, if H is a clean subset that contains no outliers, $\hat{g}_{iH}(\mathrm{x}_{ij})$ and $\hat{\zeta}_{ik}$ are consistent estimations of $g_i(\mathrm{x}_{ij})$ and ζ_{ik}, respectively; then, by applying Slutsky's theorem, we have

$$D_{in}(\hat{\zeta}_H) \xrightarrow{\mathrm{d}} D_{in}(\zeta), \quad \text{as } n \to \infty. \tag{8}$$

The ith profile is considered to be outlying when $D_{in}(\hat{\zeta}_H)$ is above a threshold. To this end, we first search for a clean subset H that minimizes the sum of the kernel-based outlier detection measures.

Definition 2.1: The LTKD measure $D_i(\hat{\zeta}_{H_{ltkd}})$ is defined by the clean subset

$$H_{ltkd} = \arg\min \sum_{i=1}^{h} D^2_{(i)n}(\hat{\zeta}_{H_{ltkd}}), \tag{9}$$

where $D^2_{(1)n}(\hat{\zeta}_{H_{ltkd}}) \leq \cdots \leq D^2_{(h)n}(\hat{\zeta}_{H_{ltkd}})$ are h-ordered low values among $\{D^2_{in}(\hat{\zeta}_{H_{ltkd}}), i = 1, \ldots, m\}$.

Next, we develop the fast LTKD algorithm to obtain the clean subset H_{ltkd}. Proposition 2.1 guarantees that the objective function decreases.

Proposition 2.1: *Suppose H_1 is a subset of $\{1, \ldots, m\}$ with $|H_1| = h$. Let $\hat{\zeta}_{ikH_1} = y_{ik} - \hat{g}_{iH_1}(\mathrm{x}_{ik})$ based on H_1 and calculate $D_{in}(\hat{\zeta}_{H_1})$ for $i = 1, \ldots, m$. If we take H_2 such that $\{D^2_{in}(\hat{\zeta}_{H_1}) : i \in H_2\} = \{D^2_{(1)n}(\hat{\zeta}_{H_1}), \ldots, D^2_{(h)n}(\hat{\zeta}_{H_1})\}$ and calculate $D_{in}(\hat{\zeta}_{H_2})$ based on H_2, we have $\sum_{i\in H_2} D^2_{in}(\hat{\zeta}_{H_2}) \leq \sum_{i\in H_1} D^2_{in}(\hat{\zeta}_{H_1})$. Hence, the equality holds if and only if $H_1 = H_2$.*

The proof of this proposition is similar to that of [26]. As H_2 corresponds to the h smallest measures, we have $\sum_{i\in H_2} D^2_{in}(\hat{\zeta}_{H_1}) \leq \sum_{i\in H_1} D^2_{i}(\hat{\zeta}_{H_1})$ and $\sum_{i\in H_2} D^2_{in}(\hat{\zeta}_{H_2}) \leq \sum_{i\in H_2} D^2_{in}(\hat{\zeta}_{H_1})$; moreover, owing to the measures of these h profiles that minimize $\sum_{i\in H_2} D^2_{in}(\hat{\zeta}_{H_2})$, we have $\sum_{i\in H_2} D^2_{in}(\hat{\zeta}_{H_2}) \leq \sum_{i\in H_1} D^2_{in}(\hat{\zeta}_{H_1})$. Hence, to obtain the optimal clean subset H, the fast least trimmed squares algorithm in [26] can be applied by replacing the squared errors with $D^2_{in}(\hat{\zeta})$. This can be named the C-step, which concentrates on the h profiles with the smallest $D^2_{in}(\hat{\zeta})$ to obtain the clean subset.

Moreover, to eliminate the bias values of the test that result from estimating the clean subset, we apply a segment step to the profiles in advance. Specifically, we randomly divide each profile data point $\{(y_{ij}, \mathrm{x}_{ij}),\ i = 1, 2, \ldots, m,\ j = 1, \ldots, n\}$ into two parts, with one labelled as $\{(y_{ij}, \mathrm{x}_{ij}),\ i = 1, \ldots, m,\ j \in \text{Part I}\}$ and the other labelled as $\{(y_{ij}, \mathrm{x}_{ij}),\ i = 1, \ldots, m,\ j \in \text{Part II}\}$, where Part I is a subset of $\{1, \ldots, n\}$ with size $[n/2]$ and Part II is the complement of Part I on $\{1, \ldots, n\}$. In this paper, we use the observed data set $\{(y_{ij}, \mathrm{x}_{ij}),\ i = 1, \ldots, m,\ j \in \text{Part I}\}$ to search for a clean subset. $\{(y_{ij}, \mathrm{x}_{ij}),\ i = 1, \ldots, m,\ j \in \text{Part II}\}$ is used to perform the hypothesis testing process and then detect the outlying profiles.

The following algorithm is used to find the optimal clean subset. It starts from a random initial subset H_{int}. The detailed steps are as follows:

Algorithm 2.1: The LTKD algorithm

(1) Let H_{int} be an initial subset of $\{1, \ldots, m\}$, with $|H_{int}| = 2$;
(2) Estimate $\hat{g}_{iH_{int}}(\mathrm{x}_{ij})$ and compute $D_{in}(\hat{\zeta}_{H_{int}})$ for $i = 1, \ldots, m$ based on data set $\{(y_{ij}, \mathrm{x}_{ij}) : i \in H_{int}, j \in \text{Part I}\}$;
(3) Sort the measures as $D^2_{\pi(1)n}(\hat{\zeta}_{H_{int}}) \leq \ldots \leq D^2_{\pi(m)n}(\hat{\zeta}_{H_{int}})$, where π is a permutation of $\{1, \ldots, m\}$; then, take $H_1 = \{\pi(1), \ldots, \pi(h)\}$ of size $h = [m]/2 + 1$;
(4) From H_1, repeatedly apply the iteration process in Proposition 2.1 t times. The corresponding subsets are $H_1, H_2, \ldots, H_t$ until convergence, that is $\sum_{i\in H_t} D^2_{in}(\hat{\zeta}_{H_t}) = \sum_{i\in H_{t-1}} D^2_{in}(\hat{\zeta}_{H_{t-1}})$. The subset H_t is the ultimate clean subset, denoted by $H_{ltkd} = H_t$.

Remarks 1: (1) The LTKD algorithm is computationally efficient and stable. Further, the iterative process is convergent, which is guaranteed by Proposition 2.1. Indeed, in our actual calculation, the algorithm typically converges no more than 10 times.
(2) The initial subset H_{int}, which is randomly selected from $\{1, 2, \ldots, m\}$, is used to estimate the regression function. Proposition 2.1 ensures that the initial subset does not affect the convergence of the results; hence, we simply take the size to be 2 in our algorithm.

After obtaining the clean subset H_{ltkd} by using Algorithm 2.1, we estimate the regression function $\hat{g}_{iH_{ltkd}}(\mathrm{x}_{ij})$ based on the clean profile data set $\{(y_{ij}, \mathrm{x}_{ij}) : i \in H_{ltkd}, j \in \text{Part II}\}$, and calculate the outlier detection measures as

$$D_{in}(\hat{\zeta}_{H_{ltkd}}) = \frac{1}{n(n-1)} \sum_{k=1}^{n} \sum_{l=1, k\neq l}^{n} K_w(\mathrm{x}_{ik} - \mathrm{x}_{il}) \hat{\zeta}_{ikH_{ltkd}} \hat{\zeta}_{ilH_{ltkd}}, \tag{10}$$

where $\hat{\zeta}_{ikH_{ltkd}} = y_{ik} - \hat{g}_i(\mathrm{x}_{ik})$ is the bias values between the ith response and estimated regression function values.

Next, we develop a threshold rule to determine whether an individual profile data point is outlying. As the asymptotic behaviour of $D_{in}(\hat{\zeta}_{H_{ltkd}})$ should be considered, we make the following assumptions.

(A1): The kernel function $K_w(\cdot)$ is a continuous, non-negative, symmetric, bounded function from $\mathbb{R}^p$ that integrates to 1.

(A2): $\{(y_{i1}, \mathrm{x}_{i1}), \ldots, (y_{in}, \mathrm{x}_{in})\}$ is a random sample from a probability distribution $F(y, \mathrm{x})$ on $\mathbb{R} \times \mathbb{R}^p$, and $p(\mathrm{x})$ is the density function of $\mathrm{x}_{i\cdot}$, whose first-order derivatives are uniform. The conditional expectation $E(y_{i\cdot} \mid \mathrm{x}_{i\cdot})$ is continuously differentiable and bounded.

(A3): In each profile, the number of sampling points n is assumed to tend to infinite.

Assumption (A1) is the most commonly used in non-parametric regression studies, while Assumptions (A2) and (A3) are essentially the same as those in Zheng (1996); hence, they are sufficient conditions for obtaining asymptotic results.

Proposition 2.2: *Assume that conditions* (A1)–(A3) *hold if* $w \to 0$ *and* $nw^p \to \infty$*; then, under the null hypothesis in Equation* (2)*, for any* $i \in \{1, 2, \ldots, m\}$*, we have*

$$nw^{p/2} D_{in}(\hat{\zeta}_{H_{ltkd}}) \xrightarrow{\mathrm{d}} N(0, \hat{\Sigma}), \tag{11}$$

where $\hat{\Sigma} = \frac{2}{n(n-1)} \sum_{k=1}^{n} \sum_{l=1, k\neq l}^{n} K_w^2(\mathrm{x}_{ik} - \mathrm{x}_{il}) \hat{\zeta}_{ik}^2 \hat{\zeta}_{il}^2$ *is the consistent estimator of* Σ*. The proofs of Propositions* 2.2 *and* 2.3 *are provided in the appendix.*

The standardized version of the outlier detection measure $D_{in}(\hat{\zeta}_{H_{ltkd}})$ is written as

$$T_{in}(\hat{\zeta}_{H_{ltkd}}) = \sqrt{\frac{n-1}{n}} \cdot \frac{nw^{p/2} D_{in}(\hat{\zeta}_{H_{ltkd}})}{\sqrt{\hat{\Sigma}}}, \tag{12}$$

and the asymptotic distribution of $T_{in}(\hat{\zeta}_{H_{ltkd}})$ is stated as follows.

Proposition 2.3: *When* $w \to 0$, $nw^p \to \infty$, *and Assumptions* (A1)–(A3) *hold, then under the null hypothesis in Equation* (2) *for* $i \in \{1, 2, \ldots, m\}$, *we have*

$$T_{in}(\hat{\zeta}_{H_{ltkd}}) \xrightarrow{d} N(0, 1). \tag{13}$$

Thus, based on the m profile data $\{(y_{ij}, \mathrm{x}_{ij}) : i \in \{1, \ldots, m\},\ j \in \text{Part II}\}$, *we calculate the standardized versions of the outlier detection measures* $T_{in}(\hat{\zeta}_{H_{ltkd}})$ *for* $i = 1, \ldots, m$. *According to the asymptotic distribution in Equation* (13), *the ith profile is deemed as an outlier if*

$$|T_{in}(\hat{\zeta}_{H_{ltkd}})| > z_\alpha, \tag{14}$$

where α *is the given significance level and* z_α *is the upper* α*th quantile of the* $N(0, 1)$ *distribution.*

Hence, by using the threshold rule in (13), we can obtain a relatively reliable non-outlier subset H_c (larger than H_{ltkd}) of normal profiles for which $|T_{in}(\hat{\zeta}_{H_{ltkd}})| \le z_\alpha$. However, its supplement, $\{1, \ldots, m\} \setminus H_c$, may also contain some non-outlying profiles. Therefore, a refinement scheme is often used in practice to enhance detection efficiency [27]. Thus, after obtaining H_c, we refine the outlier detection rule by using a one-step algorithm.

2.5. Refinement scheme

In this section, based on the profile data of $\{(y_{ij}, \mathrm{x}_{ij}),\ i \in H_c,\ j \in \text{Part II}\}$, we re-estimate the regression function and calculate the ith kernel-based outlier detection measure:

$$D_{in}(\hat{\zeta}_{H_c}) = \sum_{k=1}^{n} \sum_{l=1, k \neq l}^{n} K_w(\mathrm{x}_{ik} - \mathrm{x}_{il}) \hat{\zeta}_{ikH_c} \hat{\zeta}_{ilH_c}, \tag{15}$$

for $i = 1, \ldots, m$, where $K_w(\cdot)$, $\hat{\zeta}_{ik}$, and $\hat{\zeta}_{il}$ have similar definitions to in Equation (9).

The standardized version of $D_{in}(\hat{\zeta}_{H_c})$ can be defined as

$$T_{in}(\hat{\zeta}_{H_c}) = \sqrt{\frac{n-1}{n}} \cdot \frac{nw^{p/2} D_{in}(\hat{\zeta}_{H_c})}{\sqrt{\hat{\sigma}^2}}. \tag{16}$$

According to the asymptotic distribution results in Section 2.4, under certain assumptions, it can be proven that $T_{in}(\hat{\zeta}_{H_c})$ still follows the distribution of $N(0, 1)$ under the null hypothesis. Thus, we refine the outlier detection rule to determine the outlying profiles using an appropriately chosen $\alpha' = \alpha/2$. The ith profile is deemed as outlying if

$$|T_{in}(\hat{\zeta}_{H_c})| > z_{\alpha'}, \tag{17}$$

where $z_{\alpha'}$ is the upper α'th quantile of the $N(0, 1)$ distribution.

The one-step refinement process for detecting outliers is summarized in Algorithm 2.2.

Algorithm 2.2: The refined LTKD algorithm

(5) Apply Algorithm 2.1 to obtain a clean subset H_{ltkd}; then, compute $T_{in}(\hat{\zeta}_{H_{ltkd}})$ based on H_{ltkd};

(6) Set the significance level α, apply the threshold rule in (14), and obtain the relatively reliable non-outlier subset H_c;
(7) Compute $T_{in}(\hat{\zeta}_{H_c})$ based on the profiles in H_c;
(8) Apply the threshold rule in Equation (17) at an appropriately chosen α'; if $|\hat{T}_{in}(H_C)| > z_{\alpha'}$, the ith profile is identified as an outlier.

In this refined LTKD algorithm, steps 6–8 could be iterated many times; that is, based on the relatively reliable set obtained in the previous step, we re-estimate the regression function and determine the threshold rule for detecting outliers. The simulation results show that we can achieve good performance by using the one-step iteration. Thus, to reduce the computational cost, we perform the iteration process only once in this work.

3. Numerical simulation

In this section, we report the numerical simulations conducted to illustrate the performance of our proposed R-LTKD method for detecting outliers. To assess its performance, we choose the type-I error rate and power value as the evaluation indicators and compare our method with other methods. All the simulations are calculated by using Matlab R2019b and the results are obtained through 1000 replications.

In our simulation studies, we generate m profiles that consist of $m - m_0$ non-outlying profiles and m_0 outlying profiles from the following additive model:

$$\begin{aligned} y_{ij} &= g_a(\mathrm{x}_{ij}) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0,1); \\ g_a(\mathrm{x}_{ij}) &= g_1(x_{ij1}) + g_2(x_{ij2}) + g_3(x_{ij3}) + g_4(x_{ij4}), \end{aligned} \tag{18}$$

where $g_1(x) = ax$, $g_2(x) = (2ax-1)^2$, $g_3(x) = \frac{\sin(2\pi ax)}{2-\sin(2\pi ax)}$, and $g_4(x) = 0.1\sin(2\pi ax) + 0.2\cos(2\pi ax) + 0.3\sin(2\pi ax)^2 + 0.4\cos(2\pi ax)^3 + 0.5\sin(2\pi ax)^3$. x_{ij} is the p ($p = 4$)-dimensional covariate of the ith profile and the covariates are generated based on the following two settings:

(i) Autoregressive (AR) structures: The covariates are simulated from $N_p(0, \Sigma)$, where $\Sigma = (\rho^{|k-l|})_{p\times p}$ for $k, l = 1, \ldots, p$, $\rho = 0.5$ is chosen;
(ii) Moving average (MA) structures: $\mathrm{x}_{ij} = \Sigma_{k=1}^{L} \eta_{ik} z_{ij(k+l-1)} / (\Sigma_{=1}^{L} \eta_{ik}^2)^{1/2}$ for $j = 1, \ldots, n$ and $l = 1, \ldots, p$, where η_{ik}, $\{z_{ijk}\}$ independently follow $U(0,1)$ and $N(0,1)$, respectively and $L = [p^{1/2}]$.

Suppose that among the $m - m_0$ non-outlying profiles, m_1 profiles are incorrectly identified as outliers, and among these m_0 outlying profiles, m_2 profiles are correctly identified. The type-I error rate is denoted as $100m_1/(m - m_0)\%$ and the power value is denoted as $100m_2/m_0\%$. The type-I error rate reflects the swamping probability and the power value reflects the masking probability. Specifically, we aim to both improve the outlier detection power value (to alleviate the masking effect) and achieve the designed type-I error rate (to show that the swamping effect is not serious).

We compare the proposed R-LTKD method with a number of other methods. The first of these other approaches is the APC method proposed by Ebrahimi et al. [17]. Since the IC data set is unknown, we use the LTKD method to choose an optimal subset as the IC data

Table 1. Average percentage of type-I error rates when using the R-LTKD procedure under different choices of m_0, m, and α when $n = 200$ and $a = 1.1$ (standard deviations of the type-I error rates are in parentheses).

		$m_0/m = 0.1$			$m_0/m = 0.2$		
	m	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$
AR	50	0.9 (0.9)	4.9 (1.8)	10.1 (2.5)	0.9 (0.9)	4.9 (1.8)	9.9 (2.8)
	100	0.9 (0.9)	4.9 (1.8)	10.1 (2.5)	0.9 (0.9)	4.9 (1.8)	9.9 (2.8)
	200	1.0 (0.5)	4.9 (1.5)	10.1 (1.7)	1.0 (0.6)	4.7 (1.6)	9.9 (1.8)
	500	1.0 (0.5)	4.9 (1.3)	9.9 (1.1)	1.0 (0.6)	4.7 (1.5)	9.9 (1.1)
MA	50	0.9 (0.9)	4.9 (1.8)	10.1 (2.5)	0.9 (0.9)	4.9 (1.8)	9.9 (2.8)
	100	0.9 (0.8)	5.0 (1.8)	10.0 (2.6)	0.9 (0.9)	4.8 (1.9)	10.0 (2.7)
	200	0.9 (0.6)	4.9 (1.4)	10.0 (1.6)	1.0 (0.6)	5.0 (1.2)	10.0 (1.8)
	500	1.0 (0.4)	5.0 (0.7)	10.0 (1.1)	1.0 (0.4)	5.0 (0.8)	10.0 (1.2)

Table 2. Average percentage of type-I error rates ($\tilde{\alpha}\%$) and power values ($\tilde{\beta}\%$) when using the R-LTKD procedure under different choices of m_0 and a when $m = 100$, $n = 200$, and $\alpha = 0.05$.

		$a = 0.7$		$a = 0.9$		$a = 1.1$		$a = 1.3$	
	m_0/m	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}$	$\tilde{\beta}$
AR	0.10	4.5	96.2	4.9	100.0	5.0	100.0	5.0	100.0
	0.20	4.5	94.2	4.8	99.7	4.9	100.0	4.9	100.0
	0.30	4.3	90.8	4.8	99.2	4.9	100.0	5.0	100.0
	0.40	4.4	87.6	4.8	96.6	4.8	98.1	4.8	99.7
MA	0.10	4.5	95.4	4.8	100.0	5.0	100.0	4.9	100.0
	0.20	4.3	95.0	4.9	100.0	4.8	100.0	4.9	100.0
	0.30	4.3	92.6	4.8	99.3	4.7	99.8	4.8	100.0
	0.40	4.2	89.4	4.7	98.6	4.7	99.8	4.8	100.0

set and then perform their adaptive PC monitoring method to detect outliers. In addition, we compare our R-LTKD method with outlier detection methods for high-dimensional data. Here, we regard the m profile responses as vectors in the n-dimensional space, among which m_0 profiles are outliers. Other outlier detection methods for high-dimensional data include the R-MDP method of [28], PPOD method of [12], and χ^2 control chart method of [9]. The non-outlying profiles have $a = 0.5$ and we generate m_0 outlying profiles using different values of a. The bandwidth w is chosen based on the rule in Section 2.3. To determine the threshold rule, we choose the significance level of $\alpha = 0.05$.

Table 1 reports the average type-I error rates when using the proposed R-LTKD procedure based on the simulated data. The results show that as m and m_0 increase, the empirically found type-I error rates are close to the nominal values in most cases. Table 2 reports the average percentage of the type-I error rates and power values when $\alpha = 0.05$ and $m = 100$. The results show that as the rate of outlying profiles increases from 0.1 to 0.4, the corresponding type-I error rates and power values decrease; meanwhile, when the perturbation parameter a increases from 0.7 to 1.3, the power values are closer to 100% and the type-I error rates are controlled near 5% in most cases.

Finally, as shown in Table 3, the proposed R-LTKD method can identify the true outlying profiles in most cases and the type-I error rates are well controlled at a nominal level. By contrast, while the other methods can also achieve high power values (close to 100%), the corresponding type-I error rates are unsatisfactory. For example, the χ^2 control chart and PPOD methods incorrectly identify many of the normal profiles as outlying ones (the

Table 3. Comparison of the type-I error rates ($\tilde{\alpha}$%) and power values ($\tilde{\beta}$%) when using different methods under different choices of a when $m = 100$, $m_0 = 10$, $n = 200$, and $\alpha = 0.05$.

		R-LTKD		R-MDP		APC		PPOD		χ^2 control chart	
	a	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}$	$\tilde{\beta}$
AR	0.5	4.6	n/a	0.9	n/a	81.2	n/a	28.2	n/a	28.4	n/a
	0.7	4.4	96.4	0.0	99.2	14.6	98.7	28.2	99.8	30.3	99.3
	0.9	5.0	99.8	0.0	100.0	3.2	100.0	28.3	100.0	33.5	100.0
	1.1	5.0	100.0	0.0	100.0	3.1	100.0	28.7	100.0	37.4	100.0
	1.3	4.9	100.0	0.0	100.0	3.0	100.0	28.2	100.0	41.6	100.0
MA	0.5	5.0	n/a	0.9	n/a	81.5	n/a	28.1	n/a	28.2	n/a
	0.7	4.8	96.6	0.0	99.5	13.7	99.0	28.1	99.2	30.3	99.4
	0.9	4.3	99.7	0.0	100.0	3.1	100.0	28.1	100.0	33.7	100.0
	1.1	5.0	100.0	0.0	100.0	3.0	100.0	28.3	100.0	36.8	100.0
	1.3	4.9	100.0	0.0	100.0	2.9	100.0	27.8	100.0	40.1	100.0

type-I error rate is above α). Further, while the results of the APC and R-MDP methods are slightly better, the former still cannot control the type-I error rate well when $a = 0.5$. Indeed, the type-I error rates obtained by using these two methods tend to be extremely small and fail to reach the targeted rates in most cases. Note that when $a = 0.5$, i.e., there are no outlying profiles, the power value is not applicable (n/a).

To compare the detection levels of the methods, we set the percentage of the outlying profiles to increase from 5% to 40%. Figure 2 shows the simulation results obtained by using the various methods. Two covariance settings are considered, namely, the AR and MA cases. Figures (a) and (b) report the change in the type-I error rates and power values of the AR settings, while Figures (c) and (d) report the change in the type-I error rates and power values of the MA settings. Figures (a) and (c) show that the type-I error rates are well controlled by using the proposed R-LTKD procedure irrespective of the extent to which the proportion of non-normal data changes. By contrast, the results of the other four methods are significantly greater or less than the nominal significance level of $\alpha = 0.05$ and none of them can control the type-I error rates well. Figures (b) and (d) show that when the proportion of outliers is less than 25%, all the methods can detect outliers with a 100% power value, although the power values of the R-MDP and APC methods decrease significantly. Therefore, the results of the type-I error rates and power values confirm that the proposed method performs better than the other methods.

To illustrate how the one-step refinement algorithm affects the proposed R-LTKD procedure, we conduct further simulations. Figure 3 reports the type-I error rates and power values when the perturbation parameter changes from 0.5 to 1.5. Figures (e) and (f) report the type-I error rates and power values of the AR settings, while Figures (g) and (h) report the results when the covariate data are generated in the MA settings. As shown in Figure 2, as the disturbance parameters gradually increase, the type-I error rates tend to be stable. The rate obtained by using the LTKD method is about 0.1 compared with about 0.05 for R-LTKD, close to the nominal significance level of $\alpha = 0.05$. In addition, as the perturbation parameters increase, the power values also rise until they tend toward 1. The power value of the R-LTKD method is slightly lower than that of LTKD, but the result is satisfactory for controlling the type-I error rate. This shows that the one-step refinement procedure can greatly reduce the type-I error rates, whereas the power values do

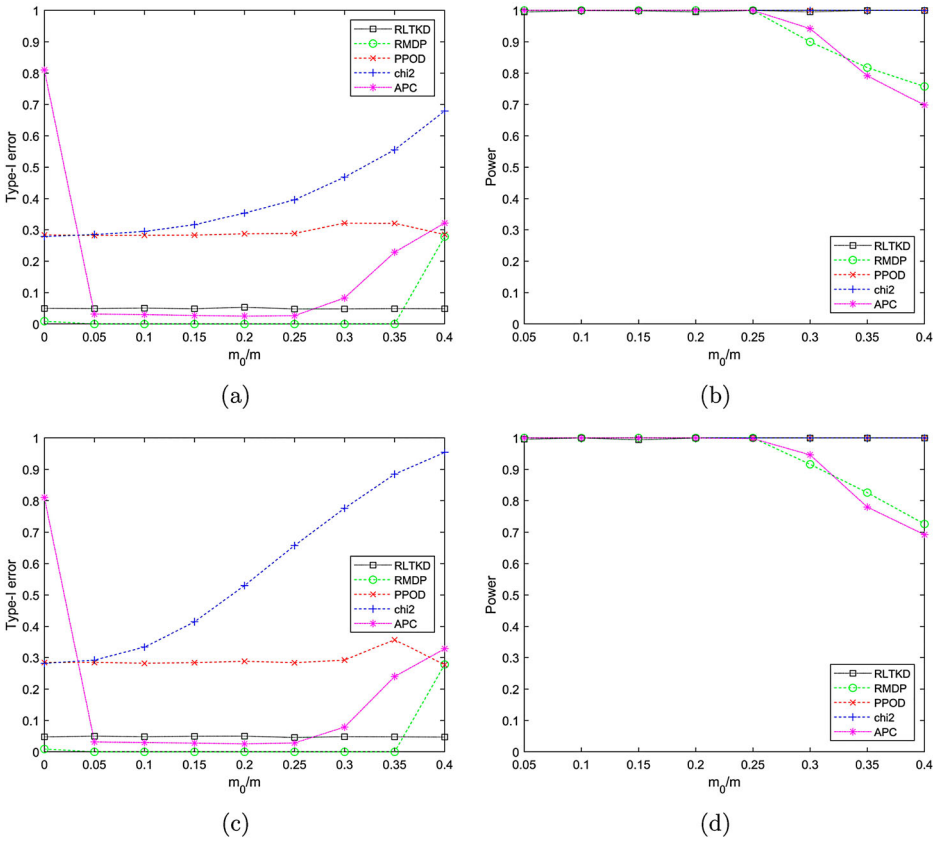

Figure 2. The change in the type-I error rates and power values (%) over time when using the various procedures when $m = 100$, $n = 200$, and $\alpha = 0.05$ with an increase in the number of outliers.

not decrease significantly. Therefore, the proposed R-LTKD method can control the type-I error rates and obtain the highest power values.

4. Real data example

By way of a real-world illustration, we apply our proposed R-LTKD approach to analyze the air quality data of Chinese cities. The data analyzed can be downloaded from China's air quality online monitoring and analysis platform (http://www.aqistudy.cn/historydata). This data set includes the air quality monitoring data of most cities in China. It contains six air monitoring indicators, namely, SO2, NO2, PM10, PM2.5, O3, CO, and an air quality index (AQI). We choose data from January 1, 2019 to July 1, 2019, collected at hourly intervals. Thus, we obtain 168 data points for each city.

Here, the AQI data are deemed as the response variable $\mathrm{y}_i = (y_{i1}, \ldots, y_{in})^\top$, where y_{ij} is the AQI value of the ith city at time j, and the predictor variables over the n time points are denoted as an $n \times p$ matrix of $\mathrm{X}_i = (x_{ijk})_{n\times p}$, for $i = 1, \ldots, m(= 349)$, $j = 1, \ldots, n(= 168)$, $k = 1, \ldots, p(= 6)$, where x_{ijk} denotes the kth predictor for city i at time j. The weekly air quality data of some of the 349 Chinese cities will inevitably be non-normal. For example, the air over some cities is seriously polluted, meaning that the monitored air quality

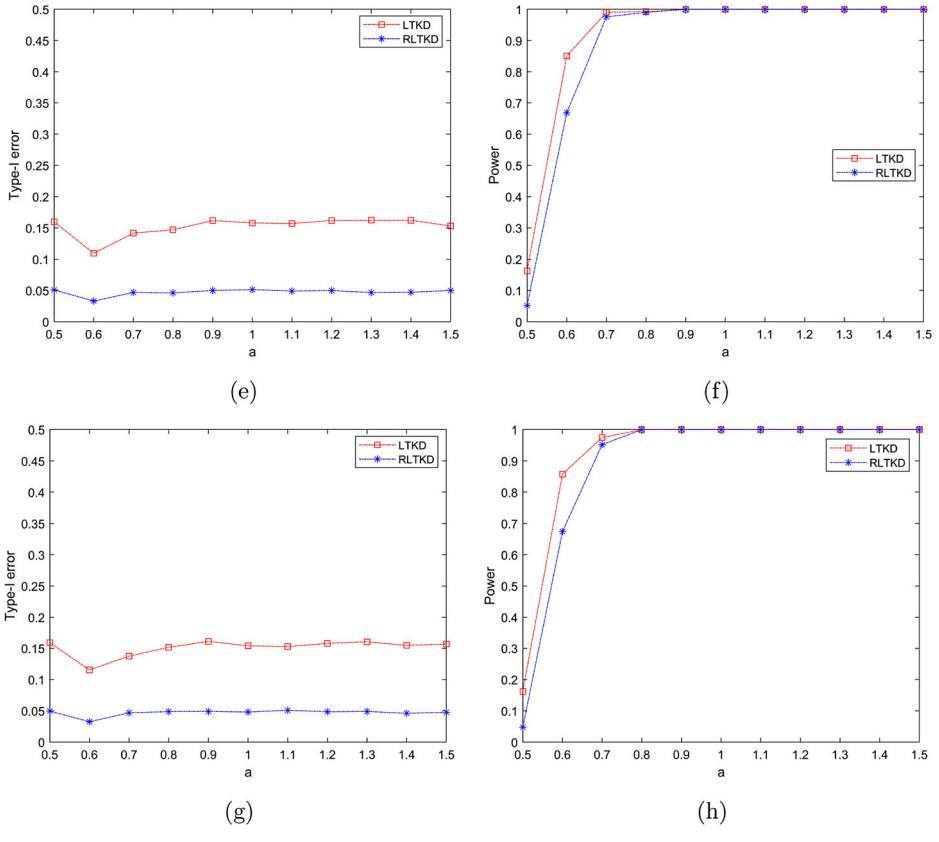

Figure 3. Comparison of the LTKD and R-LTKD methods in terms of their type-I error rates and power values (%) with an increase in a, when $m = 100$, $m_0 = 40$, $n = 200$, and $\alpha = 0.05$.

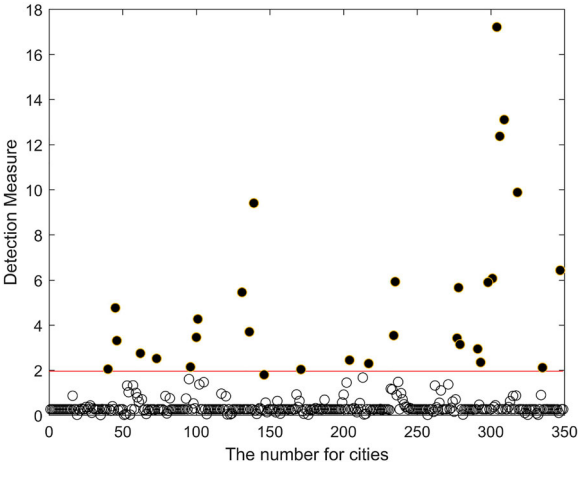

Figure 4. The detection measures of the R-LTKD procedure for the 349 Chinese cities.

data is very poor. Thus, we apply the proposed R-LTKD algorithm to the processed city air quality data with $\alpha = 0.05$. In a real data analysis, because we cannot determine how many data points are non-normal in advance, we cannot calculate the power value as in

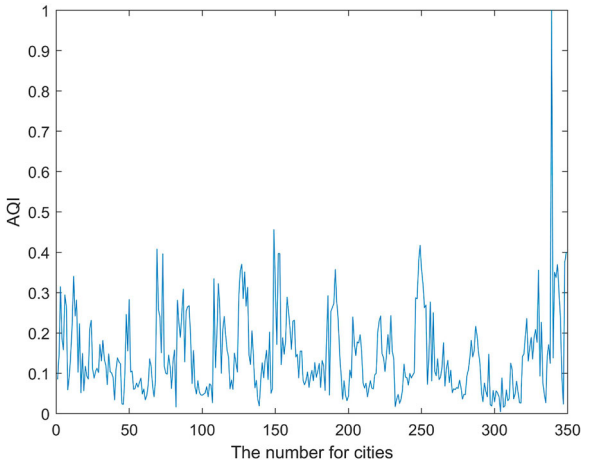

Figure 5. The AQI values of the 349 Chinese cities.

the numerical simulation to examine whether our method is effective. However, we can use the proposed R-LTKD method to calculate the outlier detection measures $T_{in}(\hat{\zeta}_{H_c})$ based on Equation (16), and then determine the outliers according to a given threshold.

For the monitored 349 cities, Figure 4 presents the absolute value of the outlier detection measures; the red line is the threshold value. Those detection measures above the red line are identified as unusual cities, comprising about 20% of the 349 cities. In addition, in Figure 4, we select the first 30 cities with high power values and mark them as black solid points. We then compare them with the first 30 values of the AQI in Figure 5, which shows that the two trends are generally consistent overall.

5. Concluding remarks

This research studies the outlier detection problem in a general set of profile data and proposes a two-stage non-parametric methodology for detecting outlying profiles. In the profile monitoring of statistical process control problems, the dimensionality and complexity of the relationship between the response and explanatory variables can increase the possibility that a profile is outlying and such outlying profiles may influence the data analysis markedly. In this article, we therefore define a non-parametric outlier detection measure and then adapt the fast LTKD algorithm by minimizing the kernel-based measure to obtain a clean subset that contains only non-outlying profiles. The threshold rule is determined based on the asymptotic distribution of the kernel-based statistic and we then use the multiple hypothesis testing process to identify outlying profiles. In addition, a one-step refinement algorithm is proposed, which allows us to control the type-I error rates more accurately. Overall, this procedure provides a new characterization using the R-LTKD algorithm that is effective at detecting outlying profiles with multiple explanatory variables, as the simulation experiments show. In particular, the proposed method can alleviate masking effects to a great extent when using an optimal choice of bandwidth. Further, the method can control type-I error rates and generate higher power values for detecting outlying profiles.

Acknowledgements

The authors are grateful to the editor, the associate editor and two anonymous referees for their comments that have greatly improved this paper. The authors also gratefully acknowledge Professor Zhaojun Wang, Professor Changliang Zou of Nankai University and Professor Zhongyi Zhu of Fudan University for their guidance and suggestions on this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by National Nature Science Foundation of China [11926347, 11926350, 11971197], Natural Science Foundation of the Higher Education Institutions of Jiangsu of China [21KJD110003], Jiangsu Government Scholarship Program [2019-43], and Qinglan Project of Jiangsu Province of China [2022].

References

[1] Mahmoud MA, Woodall WH. Phase I analysis of linear profiles with calibration applications. Technometrics. 2004;46(4):380–391.
[2] Woodall WH, Spitzner DJ, Montgomery DC, et al. Using control charts to monitor process and product quality profiles. J Qual Technol. 2004;36(3):309–320.
[3] Mahmoud MA. Phase I analysis of multiple linear regression profiles. Comm Statist Simulation Comput. 2008;37(10):2106–2130.
[4] Zou C, Qiu P. Multivariate statistical process control using lasso. J Amer Statist Assoc. 2009;104(488):1586–1596.
[5] Wang K, Jiang W. High-dimensional process monitoring and fault isolation via variable selection. J Qual Technol. 2009;41:247–258.
[6] Paynabar K, Zou C, Qiu P. A change-point approach for phase-i analysis in multivariate profile monitoring and diagnosis. Technometrics. 2016;58(2):191–204.
[7] Ren H, Chen N, Wang Z. Phase-II monitoring in multichannel profile observations. J Qual Technol. 2019;51(4):338–352.
[8] Qiu P, Zou C, Wang Z. Nonparametric profile monitoring by mixed effects modeling. Technometrics. 2010;52(3):265–277.
[9] Zhang H, Albin S. Detecting outliers in complex profiles using a $\chi 2$ control chart method. IIE Trans. 2009;41(4):335–345.
[10] Yu G, Zou C, Wang Z. Outlier detection in functional observations with applications to profile monitoring. Technometrics. 2012;54(3):308–318.
[11] Barnett V, Lewis T. Outliers in statistical data. 3rd ed. New York: Wiley; 1994.
[12] Zou C, Tseng S-T, Wang Z. Outlier detection in general profiles using penalized regression method. IIE Trans. 2014;46(2):106–117.
[13] Li Z, Shang Y, He Z. Phase I outlier detection in profiles with binary data based on penalized likelihood. Qual Reliab Eng. 2019;35(1):1–13.
[14] Ren H, Chen N, Zou C. Projection-based outlier detection in functional data. Biometrika. 2017;104(2):411–423.
[15] Abdel-Salam ASG, Birch JB, Jensen WA. A semiparametric mixed model approach to phase I profile monitoring. Qual Reliab Eng Int. 2013;29(4):555–569.
[16] Gomaa AS, Birch JB. A semiparametric nonlinear mixed model approach to phase I profile monitoring. Comm Statist Simulation Comput. 2019;48(6):1677–1693.
[17] Ebrahimi S, Ranjan C, Paynabar K. Monitoring and root-cause diagnostics of high-dimensional data streams. J Qual Technol. 2020;54(1):20–43.

[18] Hadi A, Simonoff JS. Procedures for the identification of multiple outliers in linear-models. J Am Stat Assoc. 1993;88(424):1264–1272.
[19] Zhang JX. A consistent test of functional form via nonparametric estimation techniques. J Econom. 1996;75(2):263–289.
[20] Nadaraya EA. On estimating regression. Theory Probab Appl. 1964;9(1):157–159.
[21] Watson GS. Smooth regression analysis. Sankhyā. 1964;26(4):359–372.
[22] Volpe V, Manzoni S, Marani M, et al. Leave-one-out cross-validation. Berlin: Springer; 2011.
[23] Rice J. Bandwidth choice for nonparametric regression. Ann Statist. 1984;12(4):1215–1230.
[24] Hardle W, Marron JS. Optimal bandwidth selection in nonparametric regression function estimation. Ann Statist. 1985;13(4):1465–1481.
[25] Vieu P. Nonparametric regression: optimal local bandwidth choice. J R Stat Soc Ser B Methodol. 1991;53(2):453–464.
[26] Rousseeuw PJ, Van Driessen K. Computing lts regression for large data sets. Data Min Knowl Discov. 2006;12(1):29–45.
[27] Cerioli A. Multivariate outlier detection with high-breakdown estimators. J Am Stat Assoc. 2010;105(489):147–156.
[28] Ro K, Zou C, Wang Z, et al. Outlier detection for high-dimensional data. Biometrika. 2015;102(3):589–599.

Appendix

Proof of Proposition 2.2: $g_i(\mathrm{x}_{ij})$ can be deemed as belonging to a parametric family of known functions $f((\mathrm{x}_{ij},\theta))$. Let $\tilde{g}_i(\mathrm{x}_{ij})$ be the fitting function with the smallest sum of squared errors. Then, the parameter form can be written as $\tilde{g}_i(\mathrm{x}_{ik}) = f(\mathrm{x}_{ik},\theta_0)$, where $\theta_0 = \arg\min E(y_{ik} - f(\mathrm{x}_{ik},\theta)^2)$. Thus, $\tilde{g}_i(\mathrm{x}_{ik})$ is the optimal estimation regression function of $g_i(\mathrm{x}_{ik})$.

First, as we assume when building the model, regression $g(\cdot)$ can be deemed as belonging to a parametric family. Then, we let $\zeta_{ik} = y_{ik} - f(\mathrm{x}_{ik},\hat{\theta})$, $\varepsilon_{ik} = y_{ik} - f(\mathrm{x}_{ik},\theta_0)$, where $f(\mathrm{x}_{ik},\hat{\theta})$ is a parametric estimator of $g(\mathrm{x}_i)$. Therefore, $D_{in}(\hat{\zeta}_{H_{ltkd}})$ $(i = 1,\ldots,m)$ can be decomposed into three parts as follows:

$$
\begin{aligned}
D_{in}(\hat{\zeta}_{H_{ltkd}}) &= \frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{l=1,k\neq l}^{n}\frac{1}{w^p}K\left(\frac{\mathrm{x}_{ik}-\mathrm{x}_{il}}{w}\right)\zeta_{ik}\zeta_{il} \\
&= \frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{l=1,k\neq l}^{n}\frac{1}{w^p}K\left(\frac{\mathrm{x}_{ik}-\mathrm{x}_{il}}{w}\right)[y_{ik} - f(\mathrm{x}_{ik},\hat{\theta})]\cdot[y_{il} - f(\mathrm{x}_{il},\hat{\theta})] \\
&= \frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{l=1,k\neq l}^{n}\frac{1}{w^p}K\left(\frac{\mathrm{x}_{ik}-\mathrm{x}_{il}}{w}\right)[y_{ik} - f(\mathrm{x}_{ik},\theta_0) + f(\mathrm{x}_{ik},\theta_0) - f(\mathrm{x}_{ik},\hat{\theta})] \\
&\quad\cdot[y_{ik} - f(\mathrm{x}_{ik},\hat{\theta}) + f(\mathrm{x}_{ik},\hat{\theta}) - f(\mathrm{x}_{il},\theta_0)] \\
&= \frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{l=1,k\neq l}^{n}\frac{1}{w^p}K\left(\frac{\mathrm{x}_{ik}-\mathrm{x}_{il}}{w}\right)[\varepsilon_{ik} - (f(\mathrm{x}_{ik},\hat{\theta}) - f(\mathrm{x}_{ik},\theta_0))] \\
&\quad\cdot[\varepsilon_{il} - (f(\mathrm{x}_{il},\hat{\theta}) - f(\mathrm{x}_{il},\theta_0))] \\
&= \frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{l=1,k\neq l}^{n}\frac{1}{w^p}K\left(\frac{\mathrm{x}_{ik}-\mathrm{x}_{il}}{w}\right)\varepsilon_{ik}\varepsilon_{il} \\
&\quad - 2\left\{\frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{l=1,k\neq l}^{n}\frac{1}{w^p}K\left(\frac{\mathrm{x}_{ik}-\mathrm{x}_{il}}{w}\right)\cdot\varepsilon_{ik}[f(\mathrm{x}_{il},\theta_0) - f(\mathrm{x}_{il},\hat{\theta})]\right\}
\end{aligned}
$$

$$+\frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{l=1,k\neq l}^{n}\frac{1}{w^p}K\left(\frac{\mathrm{x}_{ik}-\mathrm{x}_{il}}{w}\right)[(f(\mathrm{x}_{ik},\hat{\theta})-f(\mathrm{x}_{ik},\theta_0))]$$

$$\cdot[(f(\mathrm{x}_{il},\hat{\theta})-f(\mathrm{x}_{il},\theta_0))]$$

$$:= D_{in1}-2D_{in2}+D_{in3}.$$

Next, we show that $nw^{p/2}D_{in1}$ is asymptotically normally distributed as well as that $nw^{p/2}D_{in2} = o_p(1)$ and $nw^{p/2}D_{in3} = o_p(1)$ under certain conditions.

Step 1. Using Lemma 3.3 of [19], under certain mild conditions, we use the theory of U-statistics to show that $nw^{p/2}D_{in1}$ is asymptotically normally distributed; that is,

$$nw^{p/2}D_{in1}\xrightarrow{\mathrm{d}}N(0,\Sigma),$$

where $\Sigma = 2\int K^2(u)\,\mathrm{d}u\cdot\int[\sigma^2(\mathrm{x})]^2p^2(\mathrm{x})\,\mathrm{dx}$, which can be estimated by

$$\hat{\Sigma}=\frac{2}{n(n-1)}\sum_{k=1}^{n}\sum_{l=1,k\neq l}^{n}\frac{1}{w^p}K^2\left(\frac{\mathrm{x}_{ik}-\mathrm{x}_{il}}{w}\right)\zeta_{ik}^2\zeta_{il}^2.$$

Step 2. Next, we show that $nw^{p/2}D_{in2} = o_p(1)$ and $nw^{p/2}D_{in3} = o_p(1)$ under the null. ■

Proof: (i) Note that

$$D_{in2}=\frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{l=1,k\neq l}^{n}\frac{1}{w^p}K\left(\frac{\mathrm{x}_{ik}-\mathrm{x}_{il}}{w}\right)\cdot\varepsilon_{ik}[f(\mathrm{x}_{il},\theta_0)-f(\mathrm{x}_{il},\hat{\theta})]$$

$$=\frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{l=1,k\neq l}^{n}\frac{1}{w^p}K\left(\frac{\mathrm{x}_{ik}-\mathrm{x}_{il}}{w}\right)\varepsilon_{ik}M_n(\mathrm{x}_{il}),$$

where $M_n(\mathrm{x}_{il}) = f(\mathrm{x}_{il},\theta_0)-f(\mathrm{x}_{il},\hat{\theta})$ is continuously differentiable, $\|M(x)\|\le b(x)$ for $x\in R^p$, and $E[(b^2(x))]<\infty$:

$$D_{in2}=\frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{l=1,k\neq l}^{n}\frac{1}{w^p}K\left(\frac{\mathrm{x}_{ik}-\mathrm{x}_{il}}{w}\right)\cdot\varepsilon_{ik}[f(\mathrm{x}_{il},\theta_0)-f(\mathrm{x}_{il},\hat{\theta})]$$

$$=\frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{l=1,k\neq l}^{n}\frac{1}{w^p}K\left(\frac{\mathrm{x}_{ik}-\mathrm{x}_{il}}{w}\right)\cdot\varepsilon_{ik}\left[\frac{\partial f(x_{il},\theta_0)}{\partial\theta'}(\hat{\theta}-\theta_0)\right]$$

$$+\left[(\hat{\theta}-\theta_0)'\frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{l=1,k\neq l}^{n}\frac{1}{w^p}K\left(\frac{\mathrm{x}_{ik}-\mathrm{x}_{il}}{w}\right)\cdot\varepsilon_{ik}\frac{\partial^2 f(x_{il},\theta_0)}{\partial\theta\partial\theta'}(\hat{\theta}-\theta_0)\right]$$

$$=S_1(\hat{\theta}-\theta_0)+(\hat{\theta}-\theta_0)'S_2(\hat{\theta}-\theta_0).$$

Applying Lemmata 3.3(b) and 3.3(c) of [19], we have

$$S_1=O_p(1/\sqrt{n}),\quad \hat{\theta}-\theta_0=O_p(1/\sqrt{n}),\quad\text{and}\quad S_2=O_p(1);$$

therefore,

$$D_{in2}=O_p(1/n),$$

when $w\to 0$ and $nw^p\to 0$,

$$nw^{p/2}D_{in2}=O_p(w^{p/2})\xrightarrow{\mathrm{p}}0.$$

(ii) Similarly,

$$D_{in3} = \frac{1}{n(n-1)} \sum_{k=1}^{n} \sum_{l=1,k\neq l}^{n} \frac{1}{w^p} K\left(\frac{\mathrm{x}_{ik} - \mathrm{x}_{il}}{w}\right) [(f(\mathrm{x}_{ik}, \hat{\theta}) - f(\mathrm{x}_{ik}, \theta_0))] \cdot [(f(\mathrm{x}_{il}, \hat{\theta}) - f(\mathrm{x}_{il}, \theta_0))],$$

$$= (\hat{\theta} - \theta_0)' \frac{1}{n(n-1)} \sum_{k=1}^{n} \sum_{l=1,k\neq l}^{n} \frac{1}{w^p} K\left(\frac{\mathrm{x}_{ik} - \mathrm{x}_{il}}{w}\right) \cdot \frac{\partial f(x_{il}, \bar{\theta}_1)}{\partial \theta} \frac{\partial f(x_{il}, \bar{\theta}_2)}{\partial \theta'} (\hat{\theta} - \theta_0)$$

$$= (\hat{\theta} - \theta_0)' S_3 (\hat{\theta} - \theta_0),$$

where $\bar{\theta}_1$ and $\bar{\theta}_2$ depend on x_{il} and x_{ik}, respectively, and $\bar{\theta}_1, \bar{\theta}_2$ are between $\hat{\theta}$ and θ_0. Similar to the proof of D_{i2}, we have

$$S_3 = O_p(1), \quad \text{and} \quad D_{i3} = O_p(1/n).$$

Thus, we have

$$nw^{p/2} D_{in3} = O_p(w^{p/2}) \xrightarrow{\mathrm{p}} 0 \quad \text{as } w \to 0, \text{ and } nw^p \to 0.$$

Step 3. Finally, we show that $\hat{\Sigma}$ is a consistent estimator of Σ. ■

Proof: By using Lemmata 3.1–3.4 of [19], we can show that

$$\hat{\Sigma} = \frac{2}{n(n-1)} \sum_{k=1}^{n} \sum_{l=1,k\neq l}^{n} \frac{1}{w^p} K^2\left(\frac{\mathrm{x}_{ik} - \mathrm{x}_{il}}{w}\right) \varepsilon_{ik}^2 \varepsilon_{il}^2 + o_p(1)$$

$$\equiv 2S_4 + o_p(1).$$

Because $S_4 = \Sigma/2 + o_p(1)$, we have $\hat{\Sigma} = 2S_4 + o_p(1) = \sigma^2 + o_p(1)$. ■

Proof of Proposition 2.3: Because $T_{in}(\hat{\zeta}_{H_{ltkd}})$ is a standardized version of the test statistics $D_{in}(\hat{\zeta}_{H_{ltkd}})$, we first prove the asymptotic distribution of $D_{in}(\hat{\zeta}_{H_{ltkd}})$. Then, by applying Slutsky's theorem and Proposition 2.2, we can obtain the asymptotic distribution of $T_{in}(\hat{\zeta}_{H_{ltkd}})$. ■