Outlier Detection in Functional Observations with Applications to Profile Monitoring

Guan Yu, Changliang Zou and Zhaojun Wang[*]

LPMC and Department of Statistics, School of Mathematical Sciences,

Nankai University, Tianjin, China

Abstract

The presence of outliers has seriously adverse effects on the modeling and forecasting of functional data. Therefore, outlier detection, aiming at identifying abnormal functional curves from a dataset, is quite important. This paper proposes a new testing procedure based on functional principal component analysis. Under mild conditions, the null distribution of the test statistic is shown to be asymptotically pivotal with a well-known asymptotic distribution. The simulation results demonstrate good finite-sample performance of the asymptotic test and detection procedure. Finally, by illustrating the connection between profile monitoring in statistical process control and outlier detection in functional data, we apply the proposed approach to a real-data example from a manufacturing processes and show that it performs quite well in detecting outlying profiles.

Keywords: Asymptotic test; Functional data analysis; Functional principal component analysis; Statistical process control

1 Introduction

In many data analysis tasks, outlier detection plays an important role in modeling, inference and even data processing because outliers could adversely lead to model misspecification,

---

[*]Corresponding author. Email: zjwang@nankai.edu.cn

biased parameter estimation and poor predictions. The original outlier detection methods were arbitrary but now, principled and systematic techniques are used, developed from the contexts of statistics and computer science. By the type of data, the popular methods can be divided between univariate methods, proposed in earlier works in this field, and multivariate methods that usually form most of the current body of research. Another fundamental taxonomy of outlier detection methods is between parametric methods and nonparametric methods that are model-free (see Barnett and Lewis 1994). Statistical parametric methods either assume a known underlying distribution of the observations (e.g., Rousseeuw and Leory 1987) or, at least, they are based on statistical estimates of unknown distribution parameters (Hadi 1992). These methods flag observations deviating from the model assumptions. Within the class of nonparametric outlier detection methods, local distance measures are often used and such methods are usually capable of handling large databases (e.g., Fawcett and Provost 1997).

The motivation of this work originates from the so-called profile monitoring problem in the context of statistical process control (SPC). In some recent SPC applications, a manufacturing process or product is characterized by a profile, i.e., responses as a function of one or more explanatory variables. In particular, the profile is often some function varying over a covariate which is often time, but may also be spatial location, wavelength, etc. The aim of profile monitoring is for checking the stability of this functional/ curve relationship over time (see Woodall et al. 2004; Zou et al. 2007). In the SPC of profile problem, a critical step is to identify any outlying profiles among a set of complex profiles and to remove them from the the reference dataset. The presence of outliers has seriously adverse effects on the modeling of functional curve and accordingly on the properties of control charts (Qiu et al. 2010). See Section 4 for the details of examples.

Naturally, this problem can be regarded as outlier detection in functional data analysis (FDA). In recent years, FDA has been enjoying increased popularity due to its applicability to problems which are difficult to cast into a framework of scalar or vector observations. It deals with the case in which the data are repeated measurements of the same subject densely taken over an ordered grid of points belonging to an interval of finite length (Ramsay and Silverman 2005). Thus, for each subject, we observe a function and, though the recording points are really discrete like the curve observations in the foregoing profile monitoring, we may regard the entire function as being continuously observed.

2

Although much research has been carried out into the important problem of outlier detection in univariate/multivariate samples and regression problems, far less work has been done in FDA. Among others, Hyndman and Ullah (2007) used a method based on robust principal components analysis and the integrated squared error from a linear model; Febrero et al. (2008) considered functional outlier detection using functional depth, the likelihood ratio test and smoothed bootstrapping. In this paper, we propose a new method that uses functional principal components analysis. Our method is fast to compute and efficient in detecting the outliers in functional data. Specially, the asymptotic distribution of our proposed test statistic is derived and the threshold value which divides anomalous and non-anomalous data is based on this asymptotic distribution.

The remainder of this paper is organized as follows: our proposed methodology is described in detail in Section 2. Its numerical performance is thoroughly investigated and compared with several other approaches in Section 3. In Section 4, we demonstrate the method using a real-data applications in profile monitoring from manufacturing industries. Several remarks draw the paper to its conclusion in Section 5. Technical details are provided in the Appendix. Some other technical details, including proofs of some theorems, are provided in another appendix, which is available online as supplementary materials.

2 Methodology

2.1 Problem and notation

Consider a functional observations set $\{X_i(t), i = 1, \ldots, N\}$. Without loss of generality, we assume that $t \in \mathcal{T} = [a, b], -\infty < a < b < \infty$. Moreover, the observations $X_i(t)$ are assumed to be independent and we want to test whether there are outliers in the dataset. An exact definition of an outlier often depends on hidden assumptions regarding the data structure and the applied detection method. Yet, some definitions are regarded general enough to cope with various types of data and methods. Hawkins (1980) defines an outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. Analogously, Barnett and Lewis (1994) indicate that an outlier, is one that appears to deviate markedly from other members of the sample in which it occurs. By a similar fashion, we define outliers in a functional dataset as the observations whose

means are significantly different from the others'. Accordingly, we want to test the null hypothesis

$$H_0 : EX_1(t) = EX_2(t) = \cdots = EX_N(t), \quad t \in \mathcal{T}$$

against the alternative

$H_1 :$ There is a subset $\mathcal{A}_N$ of $\{1, \ldots, N\}$ such that $EX_k(t) = EX_l(t)$

for each $k, l \notin \mathcal{A}_N$ while $EX_k(t) \neq EX_l(t)$ for each $k \in \mathcal{A}_N$ and $l \notin \mathcal{A}_N$,

where $\mathcal{A}_N$ is the outlier set.

To be more specific, under the null hypothesis H_0, the functional observations can be modeled as independent realizations of an underlying stochastic process

$$X_i(t) = \mu_0(t) + Y_i(t), \quad i = 1, \ldots, N, \tag{1}$$

where $\mu_0(t)$ is the mean function of the stochastic process and $Y_i(t)$ is the stochastic error with $EY_i(t) = 0$. We do not specify the value of the common mean $\mu_0(t)$ in the hypothesis H_0 since this is the most common case in practice. By the preceding assumption, under H_1, the observations follow the model

$$X_i(t) = \begin{cases} \mu_i(t) + Y_i(t), & i \in \mathcal{A}_N, \\ \mu_0(t) + Y_i(t), & i \notin \mathcal{A}_N. \end{cases}$$

A straightforward approach to identify outliers in functional data is to apply the parametric/nonparametric multivariate outlier detection procedures. However, the infinite-dimensional nature of functional variation implies that in many situations, the number of grid points is larger than the number of subjects. It is well known that the most usual multivariate statistical methods suffers from "the curse of dimensionality", and thus these methods are not applicable (or at least not effective) in which the number of variables is larger than the number of individuals in the sample. Hence, it is usually important to perform dimension reduction in FDA. Functional principal component analysis (PCA) is a fundamental technique to extract a few major and typical features from functional data. Since our proposed test statistic will be constructed based on the functional PCA, we firstly briefly review it and introduce some necessary notations.

4

Let $c(t,s) = E\{Y(t)Y(s)\}$ denote the covariance function of $Y(\cdot)$. Denote λ_k and $v_k(\cdot)$ as the eigenvalues and eigenfunctions of the covariance operator $c(t,s)$ respectively, i.e., they are defined by

$$\int_a^b c(t,s)v_k(s)ds = \lambda_k v_k(t), \quad t \in \mathcal{T}, \quad k = 1, 2, \dots \tag{2}$$

In the classic FDA, $c(t,s)$ is estimated by

$$\hat{c}(t,s) = \frac{1}{N} \sum_{1 \le i \le N} \{X_i(t) - \bar{X}_N(t)\}\{X_i(s) - \bar{X}_N(s)\},$$

where $\bar{X}_N(t) = \frac{1}{N}\sum_{i=1}^N X_i(t)$. The corresponding estimators of λ_k and $v_k(\cdot)$ are $\hat{\lambda}_k$ and $\hat{v}_k(\cdot)$, defined by

$$\int_a^b \hat{c}(t,s)\hat{v}_k(s)ds = \hat{\lambda}_k \hat{v}_k(t), \quad t \in \mathcal{T}, \quad k = 1, 2, \dots$$

Under some mild conditions (given in the Appendix), $\hat{c}(t,s)$, $\hat{\lambda}_k$ and $\hat{v}_k(\cdot)$ are consistent estimators of $c(t,s)$, λ_k and $v_k(\cdot)$ respectively.

In the functional data setting, some related testing problems have recently been studied by several authors. Hall and Van Keilegom (2007) proposed two-sample functional tests from discrete data, and Cuevas et al. (2004) developed the functional analysis of variance. Benko et al. (2009) developed a bootstrap test for checking whether the elements of the two decompositions are the same by using functional PCA. Other recent contributions to hypothesis testing in this field include Locantore et al. (1999) and Spitzner et al. (2003), etc.

A more closely related work is Berkes et al. (2009) which developed a methodology for the detection of change-point $i^\star$ in the mean of functional observation $\mu_i(t)$. It is assumed that there is possibly a change-point in the dataset and the goal is to test whether it occurs or not. Also, it has been shown how to locate the change-points if the null hypothesis is rejected. It is worth pointing out that both of Berkes et al. (2009) and Qiu et al. (2010) considered a change-point problem for functional data but their assumptions and techniques are totally different. The former is based on fixed sample change-point detection and functional PCA whereas the latter considers the on-line monitoring problem (sequential change-point detection) and employs nonparametric mixed-effect models. In this work, we

consider to use a similar idea to Berkes et al. (2009), functional PCA representation and reduction of data, but focus on the outlier scenario in which we do not have any partition of the data and the data may contain several outlying functional curves with possibly different means.

2.2 The outlier detection procedure

Denote $\Delta_i(t) = X_i(t) - \bar{X}_N(t), i = 1, \ldots, N$. Suppose there are no outliers in the functional observations, then the FDA model (1) holds and we can expect that the absolute value $|\Delta_i(t)|$ is small for all $1 \leq i \leq N$ and all $t \in \mathcal{T}$. Contrarily, if there are some outliers in the sample, $\max_{1 \leq i \leq N} |\Delta_i(t)|$ would become large due to the shift of the mean of the outliers. Therefore, our test can be constructed based on the set of curves $\{|\Delta_1(t)|, |\Delta_2(t)|, \ldots, |\Delta_N(t)|\}$. We must bear in mind that the observations considered in this paper are functional data which is in an infinite dimensional space. The covariance function would be difficult to interpret and does not give a fully comprehensible presentation of the structure of the variability in the observed data directly.

To this end, we use functional PCA to reduce the dimension and to construct a test by the projections of the functions $\Delta_i(t)$ on the principal components of the functional observations. The projections are all linear combination of $\{\hat{v}_k(t), k = 1, 2, \ldots\}$. The coefficients corresponding to the largest d eigenvalues are

$$\hat{\eta}_{ik} = \int_a^b \{X_i(t) - \bar{X}_N(t)\}\hat{v}_k(t)dt, \quad i = 1, \ldots, N, \quad k = 1, \ldots, d.$$

These coefficients are ideal indicators which reflect the difference between the ith sample and the sample mean. In particular, $\hat{\eta}_{ik}$ shows the deviation degree of the ith sample on the kth mode of variation. Therefore, we propose the following test statistics

$$S_{N,d} = \max_{1 \leq i \leq N} \sum_{1 \leq k \leq d} \frac{\hat{\eta}_{ik}^2}{\hat{\lambda}_k}. \tag{3}$$

This test statistic is similar to the commonly used test statistic for outlier detections in multivariate dataset $\{\mathbf{Z}_i\}_{i=1}^N$ of dimension d (Rousseeuw and Leory 1987):

$$D_{\max}^2 = \max_{1 \leq i \leq N} (\mathbf{Z}_i - \bar{\mathbf{Z}})^T \widehat{\mathbf{\Sigma}}^{-1} (\mathbf{Z}_i - \bar{\mathbf{Z}}),$$

6

where $\bar{\mathbf{Z}}$ and $\widehat{\boldsymbol{\Sigma}}$ are the sample mean vector and covariance matrix respectively. Since the matrix $\widehat{\boldsymbol{\Sigma}}$ may be not invertible when the dimension d is large, the effectiveness and applicability of $D_{\max}^2$ would be doubtful for the high dimensional data. However, our proposed statistic does not suffer from this problem since we can choose suitable d so that all the estimated eigenvalues $\hat{\lambda}_k$ are far away from zero.

With respect to the choice of the number of the principal components, there are several approaches proposed in the literature. The data-based method to choose d is available through the cross-validation score based on the one-curve-leave-out prediction error (Yao et al. 2005a, Rice and Silverman 1991). Though the data-based cross-validation method is very attractive, it requires expensive computation, especially when the number of observations along each curve K is large. A less computationally intensive approach is to choose d based on the traditional cumulative percentage variance method. In our simulation study, typically two or three principal components were required in order to capture 85% of the variation. The simulation results in Section 3 indicates that this method is not only convenient but also effective.

When a set of N curves is measured on a fine grid of K equally spaced points, the functional principal components problem can be solved by applying standard principal components analysis to the N by K matrix of observed data. Often the grid is sparse or the time-points are unequally spaced, although still common to all curves. In this case, we usually impose smoothness constraints on the principal components in several ways. One direct approach is to represent them using a set of smooth basis functions (Ramsay and Silverman 2005). This amounts to projecting the individual rows of the data matrix on to the basis and then performing principal component analysis on the basis coefficients. Alternatively, one can use the basis coefficients to estimate the individual curves, sample the curves on a fine grid and perform principal component analysis on the resulting "data". The discrete trajectories were converted to functional observations by the latter method and Fourier bases. In addition, the observations could be irregularly spaced and the numbers of observations along each curve are unequal. In such situations, James et al. (2000) presented a technique based on reduced rank mixed effects framework. Yao et al. (2005b) proposed a nonparametric method to perform functional principal components analysis. Both methods can be used to estimate the eigenvalues λ_k and eigenfunctions υ_k, and then our proposed outlier test procedure would be still applicable. This deserves some future study.

7

Based on the foregoing discussion, to identify the true outliers set $\mathcal{A}_N$, we suggest the following Stepwise Functional Outliers Detection (SFOD) procedure by using the test statistic $S_{N,d}$ in a retrospective fashion:

Step 0: Give a significance level α and set the estimated outliers set $\mathcal{O}_N = \emptyset$;

Step 1: Choose d so that the functional PCA explains 85% of the variance;

Step 2: Compute $S_{N,d}$ and choose some threshold value $l_{N,d}(\alpha)$. If $S_{N,d} < l_{N,d}(\alpha)$, we stop the procedure. Otherwise set

$$\mathcal{O}_N = \mathcal{O}_N \cup \{i : \sum_{1 \leq k \leq d} \frac{\hat{\eta}_{ik}^2}{\hat{\lambda}_k} = \max_j \sum_{1 \leq k \leq d} \frac{\hat{\eta}_{jk}^2}{\hat{\lambda}_k}\};$$

Step 3: Delete the sample in $\mathcal{O}_N$ from the data and go back to Step 1.

The procedure will be illustrated in the real-data application in Section 4. When the SFOD procedure stops, $\mathcal{O}_N$ is the estimated outlier set. In Step 1, the PCA is recalculated every time observations are deleted from the sample. In Step 2, the cut-off value or threshold $l_{N,d}(\alpha)$ usually plays an important role in dividing anomalous and non-anomalous data numerically (cf., Hawkins 1980). Therefore, the basis for the decision on outlier identification lies on finding a proper threshold value $l_{N,d}(\alpha)$. The choice of $l_{N,d}(\alpha)$ will be discussed in the next subsection. Since $\sum_{1 \leq k \leq d} \hat{\lambda}_k^{-1} \hat{\eta}_{ik}^2$ represents the difference between the ith curve and the mean curve, the curve with the largest $\sum_{1 \leq k \leq d} \hat{\lambda}_k^{-1} \hat{\eta}_{ik}^2$ is identified as an outlier when H_0 is rejected. We delete the estimated outliers before going back to Step 1 because those "outlying profiles" may contaminate the estimation of the mean and covariance matrix in the PCA. The contamination effect of the outliers will be further discussed in the next section.

2.3 Theoretical properties

Next, we give some asymptotic properties of $S_{N,d}$, which could shed some light on practical design of the testing procedure and justify the performance of the detection procedure to a certain degree as well. We state our theorems here, but their proofs and the technical conditions used are relegated to the Appendix. Theorem 1 below gives the asymptotic null distribution of $S_{N,d}$.

To establish the asymptotic distribution of the test statistic under H_0, the following technical conditions are needed:

(C1) The mean $\mu(t)$ is square integrable, i.e. is in $\mathcal{L}^2(\mathcal{T})$. The errors $Y_i(t)$ are independent and identically distributed (i.i.d.) mean zero Gaussian process. Their covariance function $c(t, s)$ is square integrable.

(C2) The eigenvalues λ_k defined in (2) satisfy, for some $d > 0$,

$$\lambda_1 > \lambda_2 > \ldots > \lambda_d > \lambda_{d+1}.$$

Remark 2 Note that except for the Gaussian assumption on the error process $Y_i(t)$, Conditions (C1) and (C2) are the same as the conditions in Berkes et al. (2009). These two conditions are sufficient to guarantee that $\hat{\lambda}_k$ and $\hat{v}_k(\cdot)$ are reasonable estimators of λ_k and $v_k(\cdot)$. The results of Bosq (2000) imply that, for each $k \leq d$,

$$\limsup_{N \to \infty} \left[N(E(\|\hat{c}_k v_k(t) - \hat{v}_k(t)\|^2)) \right] < \infty, \tag{4}$$

$$\limsup_{N \to \infty} \left[N(E(|\lambda_k - \hat{\lambda}_k|^2)) \right] < \infty, \tag{5}$$

where $\hat{c}_k = \text{sgn}\left\{ \int_a^b v_k(t)\hat{v}_k(t)dt \right\}$. Furthermore, the Condition (C1) implies the following expansions

$$c(t, s) = \sum_{1 \leq k < \infty} \lambda_k v_k(t) v_k(s), \quad Y_i(t) = \sum_{1 \leq k < \infty} \lambda_k^{1/2} \xi_{ik}(t) v_k(t), \tag{6}$$

where the sequences $\{\xi_{ik}, i = 1, \ldots, N, k = 1, 2, \ldots\}$ are i.i.d. normal random variables with mean 0 and unit variance. It's easy to check that the infinite sum in Eq.(6) converges in $\mathcal{L}^2(\mathcal{T} \times \mathcal{T})$ and $\mathcal{L}^2(\mathcal{T})$, all λ_k's are non-negative and the eigenfunctions $v_k(t), k = 1, 2, \ldots$, form an orthonormal basis in $\mathcal{L}^2(\mathcal{T})$.

Theorem 1 *Suppose that Conditions (C1)-(C2) hold. Then, under null hypothesis H_0, for each $x \in \mathbb{R}$, we have*

$$P\left\{ \frac{S_{N,d}}{2} - \log N - (d/2 - 1) \log \log N + \log \Gamma(d/2) \leq x \right\} \to e^{-e^{-x}}, \quad as \ \ N \to \infty. \tag{7}$$

9

The asymptotic null distribution of $S_{N,d}$ is independent of the nuisance parameters $\mu(t)$ and $c(t,s)$ and thus $S_{N,d}$ is asymptotically pivotal. By this theorem, we can obtain the approximate critical value of the test statistic $S_{N,d}$. Define

$$u_{N,d}(\alpha) = 2c_d(\alpha) + 2\log N + (d-2)\log\log N - 2\log\Gamma(d/2),$$

where $c_d(\alpha)$ is the upper α quantile of the double exponential distribution. Then, we propose the functional data outlier test (FDOT) with rejection region $\{S_{N,d} \geq u_{N,d}(\alpha)\}$. It is the basis for our SFOD procedure. This test has asymptotic significance level α. However, the test based on $u_{N,d}$ may perform poorly in the small-sample situations since the convergence in Theorem 1 is relatively slow. Empirically speaking, when N is not large enough, the approximation of (7) yields somewhat conservative results for small values of d. Alternatively, when N is small, we suggest to simulate the distribution of the following random variable:

$$G_{N,d} = \max_{1 \leq i \leq N} \sum_{k=1}^{d} (\xi_{ik} - \bar{\xi}_k)^2,$$

where $\{\xi_{ik} : i = 1, \ldots, N; k = 1, \ldots, d\}$ are i.i.d. $N(0,1)$ random variables and $\bar{\xi}_k = \frac{1}{N}\sum_{i=1}^{N}\xi_{ik}, k = 1, \ldots, d$. The upper α quantile of the distribution of $G_{N,d}$, denoted as $g_{N,d}(\alpha)$, is a good choice for the critical value of our proposed test with approximate significance level α. The reason is that the above random variable $G_{N,d}$ is a reasonable approximation to the $S_{N,d}$ as shown in the proof of the Theorem 1. Accordingly, $U_{N,d}$ and $G_{N,d}$ are two choices for $l_{N,d}$ in Section 2.2. For some N, d and three significance levels 10%, 5% and 1%, these two kinds of critical values are tabulated in Table 1. Based on our experience and simulation results (partly reported in the next section), we recommend to use the distribution of $G_{N,d}$ when $N \leq 100$, and otherwise to use the asymptotic distribution given in (7).

Furthermore, regarding to the behavior of the test under H_1, we have the following result:

Theorem 2 *Suppose that Conditions (C1)-(C2) and additional conditions (C3)-(C5) given in the Appendix hold. Then, for each $\alpha \in (0,1)$, under alternative hypothesis H_1, we have*

$$P_{H_1}\{S_{N,d} > u_{N,d}(\alpha)\} \to 1, \quad as \quad N \to \infty.$$

Theorem 2 says that our proposed test is consistent if the number of outliers m_N grows with the sample size N in the order specified in the Appendix. It guarantees that the

Table 1: The critical values based on $u_{N,d}$ and $g_{N,d}$ for various N, d and α.

N	α	$d=1$		$d=2$		$d=3$		$d=4$	
		$u_{N,d}$	$g_{N,d}$	$u_{N,d}$	$g_{N,d}$	$u_{N,d}$	$g_{N,d}$	$u_{N,d}$	$g_{N,d}$
	0.10	9.81	9.26	12.32	12.07	13.93	14.39	15.05	16.46
50	0.05	11.25	10.58	13.76	13.46	15.37	15.91	16.49	18.03
	0.01	14.51	13.65	17.02	16.57	18.63	19.14	19.75	21.61
	0.10	11.03	10.65	13.71	13.61	15.47	15.98	16.76	18.18
100	0.05	12.47	11.96	15.15	15.04	16.91	17.51	18.21	19.75
	0.01	15.73	15.05	18.41	18.23	20.17	20.87	21.46	23.26
	0.10	12.28	11.92	15.09	15.04	17.01	17.55	18.43	19.81
200	0.05	13.72	13.23	16.53	16.48	18.44	19.03	19.87	21.38
	0.01	16.98	16.37	19.79	19.67	21.71	22.42	23.13	24.91
	0.10	13.54	13.28	16.48	16.45	18.51	19.01	20.06	21.35
400	0.05	14.98	14.66	17.92	17.88	19.95	20.48	21.51	22.89
	0.01	18.24	17.65	21.18	21.21	23.21	23.88	24.76	26.32

functional outliers test is effective from certain theoretical points. In the following section, through the simulation under a variety cases, we will show that our proposed test performs well in a finite-sample setting.

3 Simulation studies

To see the performance of our proposed test and outlier detection procedure, we have conducted many simulation studies. Some of the results are reported here. In Section 3.1, we investigate the approximation of the level by two methods given in Section 2.3. A power analysis is conducted in Section 3.2 to evaluate the effectiveness of the proposed test and make comparisons with three related methods. In Section 3.3, we discuss how to alleviate the masking effect. Finally, we study the performance of our proposed SFOD procedure in Section 3.4.

3.1 Empirical size study

To study the empirical size, for simplicity and without loss of generality, the mean $\mu(t)$ was chosen to be 0 and the following three different cases of $Y(t)$ were considered:

Scenario 1 : Standard Brownian Motion (BM);

Scenario 2 : Standard Brownian Bridge (BB);

Scenario 3 : $\sin(2\pi t)Z_0 + 0.5Z_t$, where Z_0 and Z_t are independent standard normal variables.

All the three processes were realized on a grid of 200 equispaced points in $\mathcal{T} = [0,1]$. To simulate a standard Brownian Motion, we repeatedly generated independent Gaussian random variables with mean 0 and standard deviation $1/\sqrt{200}$. The value of the Brownian Motion at time $i/200$ is the first i increments. In order to simulate a standard Brownian Bridge $S(t)$, we firstly generate a standard Brownian Motion $B(t)$. Then, through the transformation $S(t) = B(t) - tB(1)$ we acquire a standard Brownian Bridge sample. Following the basis function method introduced in Ramsay and Silverman (2005), the discrete trajectories were converted to functional observations by 15 Fourier bases. Our simulation study found that our method is not affected much by the type of the basis or the number of basis functions. For an estimation problem, one can use cross-validation (CV), generalized cross-validation (GCV) or other model selection criteria to choose the number of Fourier bases. However, our simulation results show that these criteria tailored for estimation often do not produce an optimal test. This finding is not surprising because similar conclusions have been made in the nonparametric regression testing problem and other related contexts (cf., Hart 1997). Similar observations in the context of profile monitoring have also been made in Zou et al. (2008). The number of bases should be chosen to balance the size of the test and the detection ability to various outlying profiles. We find that with 5 to 15 Fourier bases, the level of the proposed test can be maintained within an acceptable range. To provide a better protection against local/oscillating functional changes, we use a relatively larger number of bases, say 15. In practice, the spline or local polynomial smoothing can be used as well. We have studied many cases with different sample sizes, but only report here the results of the above three models when the sample sizes N were chosen to be 50, 100, 200 and 400. In each scenario, the empirical size is computed based on 2,000 replications. In each replication, the number of the eigenfunctions d was chosen automatically by the cumulative percentage variance approach which found a suitable d explaining 85% of the variance.

Figure 1 shows the empirical sizes based on two kinds of critical values computed from the limiting distribution and simulated distribution of $G_{N,d}$. The nominal significance level α is
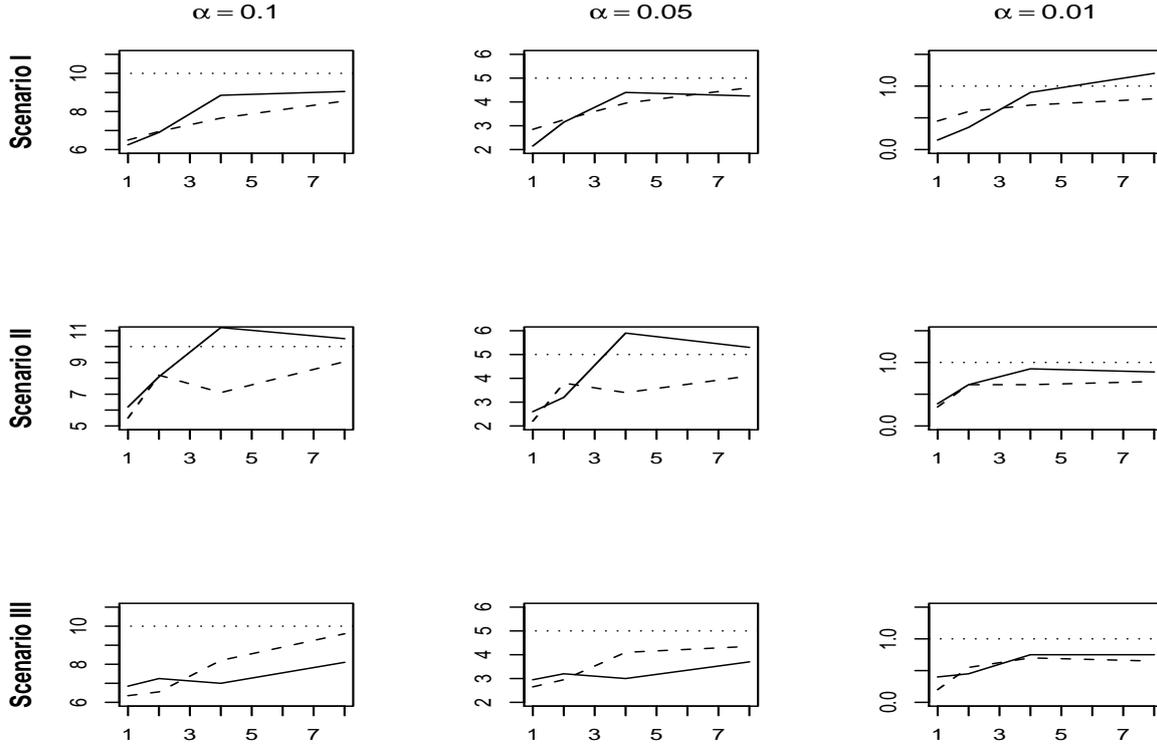
Figure 1: Empirical size of the test using $u_{N,d}$ (solid line) and $g_{N,d}$ (dashed line) along with the nominal size α (%; dotted line). The empirical size (y-axis) is plotted against $N/50$ (x-axis).

chosen to be 0.1, 0.05 and 0.01. The results indicate that the empirical sizes approximate the nominal significance levels as the sample size increases. Otherwise, the empirical sizes based on the simulated critical values $g_{N,d}$ are closer to the nominal levels when the sample size is relatively small, for example, $N \leqslant 100$. In a majority of cases, both two approximations of critical value seem to result in conservative tests. This partially stems from the fact that the number d of the eigenfunctions was not specified in our simulations (determined by cumulative percentage variance approach) whereas the asymptotic distribution of $S_{N,d}$ given in Theorem 1 was derived assuming that d was specified.

3.2 Power study

In this subsection, we study the power of different tests in rejecting null hypothesis that the means of all the function are the same. We consider the following three simulated data sets

13

generated by:

$$\text{Case (I)}: X_i(t) \sim 2\sin(2\pi t) + BM, \quad i \in \mathcal{A}_N; \quad X_i(t) \sim BM, \quad i \notin \mathcal{A}_N;$$

$$\text{Case (II)}: X_i(t) \sim 0.6e^t + BB, \quad i \in \mathcal{A}_N; \quad X_i(t) \sim BB, \quad i \notin \mathcal{A}_N;$$

$$\text{Case (III)}: X_i(t) \sim -3.8t + \sin(2\pi t)Z_0 + 0.5Z_t, \quad i \in \mathcal{A}_N;$$

$$X_i(t) \sim \sin(2\pi t)Z_0 + 0.5Z_t, \quad i \notin \mathcal{A}_N;$$

where $\mathcal{A}_N$ denotes the outlier set.

Denote $|\mathcal{A}_N| = [\rho N]$, where ρ is the ratio of the outliers in the sample and $[\rho N]$ denotes the greatest integer less than or equal to ρN. In our simulation study, we chose N to be 50, 100, 200 and 400. The ratio of the outliers ρ was chosen to be $2\%, 4\%$ and 6% for all the three cases. The locations of the outliers were chosen evenly. All the three processes were realized on a grid of 30 or 200 equispaced points in $\mathcal{T} = [0,1]$ and smoothed by 15 Fourier bases.

In order to assess the performance of FDOT, we compare it with some existing methods. A commonly used approach is the *multivariate outlier test* (MOT; Rousseeuw and Leory 1987) which uses the test statistic $D_{\max}^2$ introduced in Section 2.2. When the number of observations along each curve K is larger than the sample size N, we use the tapering estimator (Cai et al. 2010) to estimate the inverse of the covariance matrix. The tapering parameter was chosen to be seven as suggested by Cai et al. (2010). Another related approach is the test based on *h-modal functional depth* (Febrero et al. 2008; FDET). Its test statistic is:

$$R_N = -\min_{1 \leqslant i \leqslant N} \sum_{j=1}^{N} K\left(\frac{\| X_i(t) - X_j(t) \|}{h}\right),$$

where $K(\cdot)$ is the Gaussian kernel function, the bandwidth h is chosen to be the 15th percentile of the empirical distribution of $\{\| X_i(t) - X_j(t) \|, i, j = 1, ..., N\}$ and

$$\| X_i(t) - X_j(t) \| = \sup_{k=1,...,K} |X_i(t_k) - X_j(t_k)|.$$

In addition, since our proposed FDOT uses the functional PCA, it is also of interest to compare our method with some other tests based on functional PCA. A natural benchmark is the *functional data change-point test* (FDCT) proposed by Berkes et al. (2009). FDCT

uses the following test statistic:

$$Q_{N,d} = \frac{1}{N^2} \sum_{l=1}^{d} \hat{\lambda}_l^{-1} \sum_{j=1}^{N} (\sum_{1 \leqslant i \leqslant j} \hat{\eta}_{i,l} - \frac{j}{N} \sum_{1 \leqslant i \leqslant N} \hat{\eta}_{i,l})^2.$$

Berkes et al. (2009) demonstrated that the test has excellent finite sample performance for the functional data change-point problem.

For a fair comparison, all the critical values of the considered tests were computed by the simulated distributions of the test statistics under the true model to control their significance levels to be 0.05. Table 2 shows the power of different tests when $K = 30$ and 200 respectively with 2,000 replications for each case. In this table, the entries with asterisk indicate that the corresponding tests are not unbiased, say the power is less than 5%. The simulation results indicate that our proposed FDOT method outperforms the other approaches in most cases. The FDCT seems to be the worst one among all the tests. This can be well understood as FDCT was designed for the change-point problem. In order to have satisfactory power, it generally requires that the number of samples before and after the change-point are both large enough. However, in our functional data outlier test problem, we cannot assume that the outliers gather together in the sample. Moreover, the comparison between the simulation results of $K = 30$ and 200 indicates that our proposed FDOT procedure turns to be quite robust when the number of observations along each curve varies.

3.3   The masking effect

From Table 2, we note that for the fixed sample size, the powers of tests decrease to some extent as the number of the outliers increases. This tendency is not surprising since all the considered tests are suffering from the masking effect. When there are multiple outliers in the sample, the estimators of the mean curve and the covariance function will be skewed. Furthermore, with masking, removal of any single outlier might have little or no effect on estimates since other outliers remain. In such situations, robust estimators of the mean curve and covariance function are needed. Rousseeuw and van Zomeren (1990) proposed to use the minimum volume ellipsoid estimator of mean and covariance matrix for the multivariate outliers detection problem. However, our simulation study indicated that their robust estimation procedure was not effective in the present problem. It is often unstable and requires expensive computation. Some similar findings can be found in Yeh et al. (2009) in which

Table 2: Power (%) comparisons of different tests when $K = 30$ and 200

Model	ρ	N	$K = 30$				$K = 200$			
			FDOT	FDCT	MOT	FDET	FDOT	FDCT	MOT	FDET
Case I	2%	50	95.7	*	12.8	33.4	96.7	*	18.0	32.2
		100	99.3	8.20	65.4	32.2	99.6	7.60	15.7	30.2
		200	100	18.8	85.9	30.9	99.9	18.6	13.8	26.6
		400	100	48.6	94.1	27.4	100	47.7	7.45	29.4
	4%	50	99.7	*	8.30	31.2	99.5	*	13.5	31.7
		100	100	*	23.8	29.6	99.9	*	9.25	30.0
		200	99.9	11.0	32.3	26.6	100	10.5	7.10	27.9
		400	100	40.4	39.3	26.6	100	41.2	6.95	28.1
	6%	50	97.4	*	7.1	28.4	97.6	*	11.5	26.4
		100	96.7	*	14.2	25.0	96.2	*	6.90	27.1
		200	95.1	*	16.4	23.2	95.8	*	*	23.8
		400	94.7	*	19.2	20.7	95.3	*	6.00	23.2
Case II	2%	50	83.5	*	20.2	78.6	80.1	*	*	66.3
		100	91.7	7.40	75.8	79.9	89.7	5.25	*	72.2
		200	96.9	11.3	89.7	81.8	96.1	12.4	*	80.5
		400	99.5	32.3	96.2	87.8	99.4	29.8	6.50	85.4
	4%	50	86.8	*	8.05	81.7	81.2	*	5.45	76.7
		100	91.6	*	25.9	85.7	88.5	*	6.85	81.4
		200	96.0	5.90	35.6	85.4	94.7	7.60	8.55	85.3
		400	97.8	25.2	42.4	86.9	97.7	24.0	8.85	89.2
	6%	50	62.6	*	7.55	77.3	56.2	*	*	76.6
		100	62.5	*	15.1	77.9	59.1	*	*	78.0
		200	64.3	*	17.0	78.5	66.9	*	*	82.7
		400	67.9	*	19.4	79.2	71.9	*	15.8	86.2
Case III	2%	50	100	*	100	97.3	60.3	8.30	29.5	99.7
		100	100	*	99.9	98.6	61.1	18.5	29.3	99.2
		200	100	7.6	100	99.0	63.7	63.3	32.9	99.2
		400	100	76.8	100	99.3	65.6	66.2	13.5	99.2
	4%	50	100	*	13.7	98.5	99.4	7.95	34.1	99.8
		100	100	*	29.8	98.7	100	12.9	30.3	99.6
		200	100	*	41.0	98.9	100	48.0	30.6	99.6
		400	100	7.65	47.1	98.2	100	100	7.10	99.6
	6%	50	100	*	8.0	96.6	99.9	9.25	33.1	99.3
		100	72.0	*	10.6	96.2	100	6.55	27.2	99.1
		200	36.4	*	14.2	94.7	100	6.00	22.7	98.8
		400	22.7	*	16.3	94.3	100	6.15	5.60	99.0

Note: * indicates that the test is not unbiased.

the focus is the outlier detection for linear profiles. In practice, we recommend to use a two-step FDOT procedure to alleviate the masking effect. At the first step, we choose $d = 1$ and a relativelt large significance level (such as $\alpha = 0.1$ in our simulation). Through the SFOD procedure, we exclude some "candidate outliers" from the sample and then use the filtered sample to estimate the mean curve and the covariance function. At the second step, we conduct the FDOT procedure using the estimators of the mean curve and covariance function based on the "cleaner" sample obtained from the first step. Figure 2 shows the size ($\rho = 0$) and power comparison between the two-step FDOT and one-step FDOT for Case II when $K = 30$. The nominal significance level α was still chosen as 0.05. The simulated results in Figure 2 show that the two-step FDOT procedure alleviates the masking effect to certain degree. It is also worth noting that due to the extra variation introduced by the additional steps, the two-step procedure tends to have larger size than the one-step procedure. However, as shown by the simulation results in Section 3.1, the one-step FDOT procedure is usually somewhat conservative. This results in the sizes of the two-step procedure are generally not far away from the nominal size.

3.4   The performance of SFOD procedure

We compare the performance of our SFOD procedure with the depth-based functional outlier detection procedure introduced by Febrero et al. (2008) (denoted as DFOD). In this comparison, we set $K = 200$ and the significance level α was chosen as 0.1. To evaluate the statistical performance of the SFOD procedure, we consider two accuracy measures, $r_1 = \frac{NTID}{NID}$ and $r_2 = \frac{NTID}{NT}$, where NTID is the number of true outliers identified, NID is the number of samples identified as outlier, and NT is the number of true outliers in the sample. These two indexes provide certain indication of the precision of the detection results. Large r_1 and r_2 indicate superior detection.

Table 3 shows the simulated values of these two indexes for Cases (I)-(III) when the ratio of the true outliers is chosen to be 2%, 4% and 6%. The results show that the SFOD procedure generally performs better than DFOD. From the measure r_1, we find that the SFOD procedure identifies true outliers more successfully as the sample size increases. As shown in Table 3, when the sample size is 400, most of the curves in the set $\mathcal{O}_N$ (identified by the SFOD procedure) are true outliers. Moreover, from the results of the measure r_2,
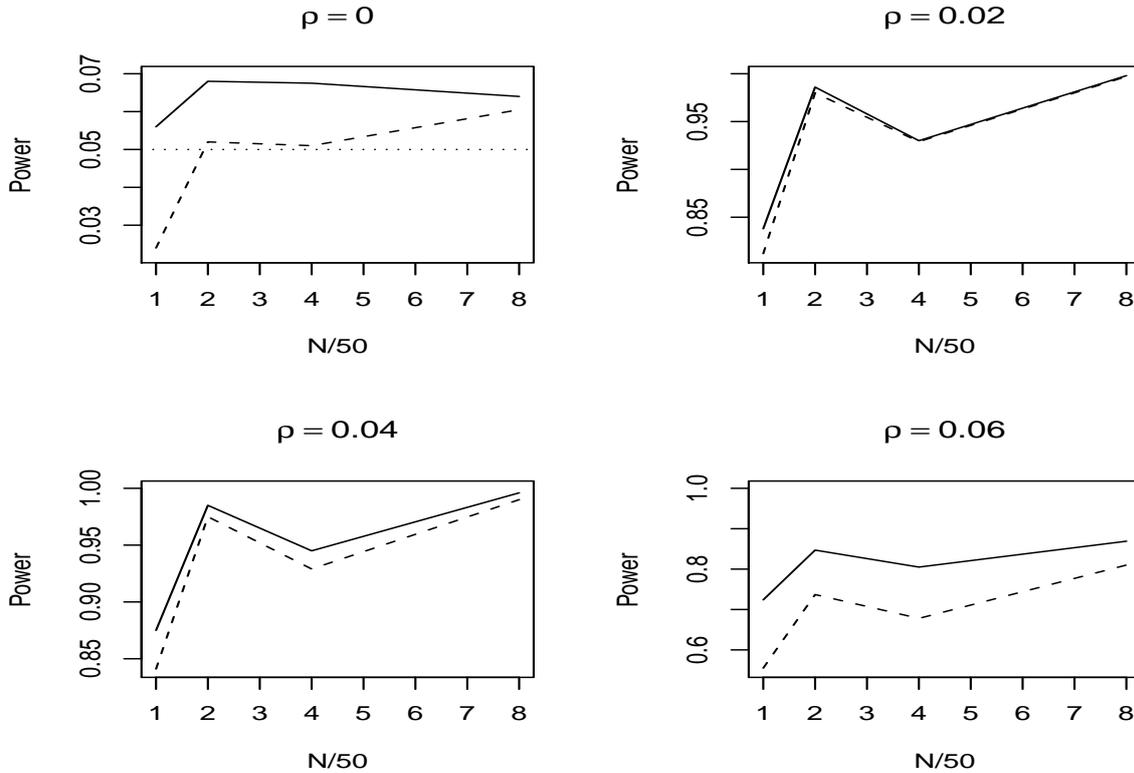
17

Figure 2: The size and power comparison between the two-step FDOT (solid line) and one-step FDOT (dashed line). The dotted line in the first plot represents the nominal significance level $\alpha = 0.05$.

we can see that the SFOD procedure could identify most of the true outliers except for a handful of cases that the ratio of the outliers is large while the shift of them is small.

4 A Real-data application in industrial manufacturing: profile monitoring

The proposed methodology is demonstrated in an aluminium electrolytic capacitor (AEC) manufacturing process in this section. More detailed discussion about the AEC example may be found in Qiu et al. (2010). Regarding quality of AECs, the most important characteristic is dissipation factor (DF), which can be automatically measured by an electronic device. However, it is known that DF measurements would change significantly with environmental temperature, and there is a specific requirement about the adaptability of AECs to the

Table 3: The Performance Comparison of SFOD and DFOD When $K = 200$

		Case (I)				Case (II)				Case (III)			
		$r_1(\%)$		$r_2(\%)$		$r_1(\%)$		$r_2(\%)$		$r_1(\%)$		$r_2(\%)$	
ρ	N	SFOD	DFOD	SFOD	DFOD	SFOD	DFOD	SFOD	DFOD	SFOD	DFOD	SFOD	DFOD
	50	92.1	44.1	97.4	46.8	78.0	76.0	83.5	81.5	93.5	87.3	99.5	92.5
2%	100	96.2	42.1	97.1	28.3	89.8	82.9	79.5	69.3	96.1	92.2	100	84.9
	200	97.7	37.2	94.0	13.2	96.4	88.5	72.3	51.4	98.0	95.1	100	71.1
	400	98.8	40.4	90.5	7.06	98.4	92.2	65.8	35.0	98.7	96.7	100	51.4
	50	96.2	48.2	98.5	35.8	84.7	86.4	79.2	79.7	95.5	96.2	99.3	100
4%	100	98.0	43.9	96.7	17.2	93.4	89.9	75.5	64.3	97.4	97.6	99.8	99.9
	200	99.0	39.7	94.7	7.17	97.4	92.5	68.5	42.1	98.7	98.6	100	98.0
	400	99.4	41.9	91.5	3.65	99.0	94.6	60.6	25.5	99.3	99.1	100	77.9
	50	97.2	41.8	97.3	23.7	71.2	88.1	61.7	76.0	96.7	96.2	99.9	100
6%	100	98.2	40.2	96.4	10.3	74.8	90.9	51.7	57.5	98.4	98.2	100	99.9
	200	99.2	36.8	94.3	4.27	83.7	93.0	41.6	33.6	99.1	99.0	100	97.6
	400	99.3	37.6	90.7	2.14	86.7	94.9	23.0	17.3	99.5	99.1	100	68.1

temperature. In order to monitor the adaptability, engineers put a sampled AEC in a container. Then, the container's temperature is controlled, and the temperature is supposed to stably increase from -26^oF to 78^oF. In this process, measurements of dissipation factor (DF) and the actual temperature inside the container are taken at 53 equally spaced time points. The actual temperature inside the container is reported by a temperature sensor. Figure 3 shows some AEC curves which represent the functional relationship between DF and temperature.

Most existing profile monitoring methods (e.g., Zou et al. 2008; 2009) require a fundamental assumption that observations within a profile are independent of each other, which is apparently invalid in applications. To properly describe within-profile correlation, Qiu et al. (2010) proposed a nonparametric mixed-effects model (cf., Wu and Zhang 2002), which allows a flexible variance-covariance structure. Alternatively, we may consider the AEC profiles as functional observations with some random errors and thus can be adequately described by the FDA model (1).

The entire AEC dataset contains 144 curves. As discussed in Section 2.2, the discrete sampled curves were converted to functional observations by using 15 Fourier basis functions.
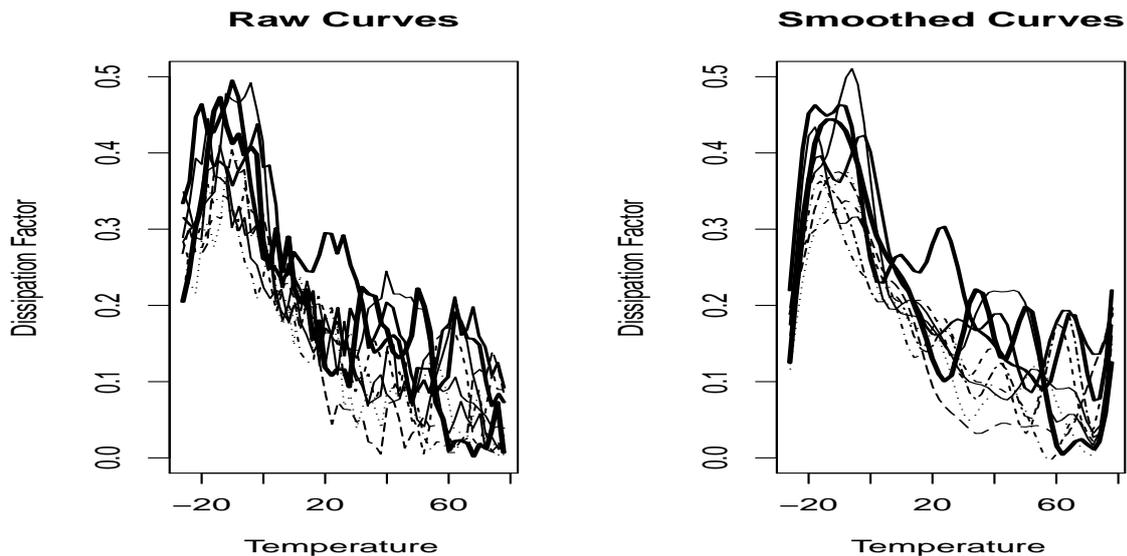
Figure 3: The first six (thin lines), the $115^{th}, 116^{th}, 117^{th}$, and the 123rd (thick lines) AEC functional curves.

Then we used the cumulative percentage variance approach which finds $d = 2$ explaining 85% of the variance. We then computed the statistic $S_{N,d}$, the candidate outliers subset $\mathcal{O}_N = \{i_0 : \arg\max_{1 \leqslant i \leqslant N} \sum_{1 \leqslant k \leqslant d} \frac{\hat{\eta}_{ik}^2}{\hat{\lambda}_k}\}$ and the p-value by the asymptotic distribution in Theorem 1. Then, we removed the sample in the subset $\mathcal{O}_N$ and repeated the same computation steps several times. Table 4 gives the results of the first ten steps of the stepwise detection procedure. The first p-value 0.0069 strongly indicates that there are outliers in the data. If we specify the significance level to be 0.05, we can identify the indices of the outlying observations to be 115, 116, 117 and 123.

Table 4: Detection results of the AEC example

Rank	p-value	Index(SFOD method)	Depth	Index(Depth method)
1	0.0069	115	14.836	117
2	0.0072	123	14.863	115
3	0.0265	117	15.579	121
4	0.0327	116	16.402	120
5	0.0881	121	16.886	118
6	0.0954	120	17.852	116
7	0.2295	118	18.107	118
8	0.3587	110	18.455	56
9	0.2066	23	18.813	123
10	0.2756	30	19.155	139

20

For the purpose of comparison, we also detected the outlying AEC curves by the h-modal functional depth introduced in Febrero et al. (2008). Outlying AEC curves are expected to be far away from the center of the data and therefore correspond to curves of significantly low depth. The depth values corresponding to the curves with the ten smallest h-model functional depths are also shown in Table 4. We found that two methods acquire similar results. The indices of the candidate outlying curves are both between 115 and 123. Figure 4 shows the first three non-outlying curves, the mean curve and the 115th, 116th, 117th and 123th curves after smoothing. The deviation of the outlying curves and the non-outlying curves is quite clear.
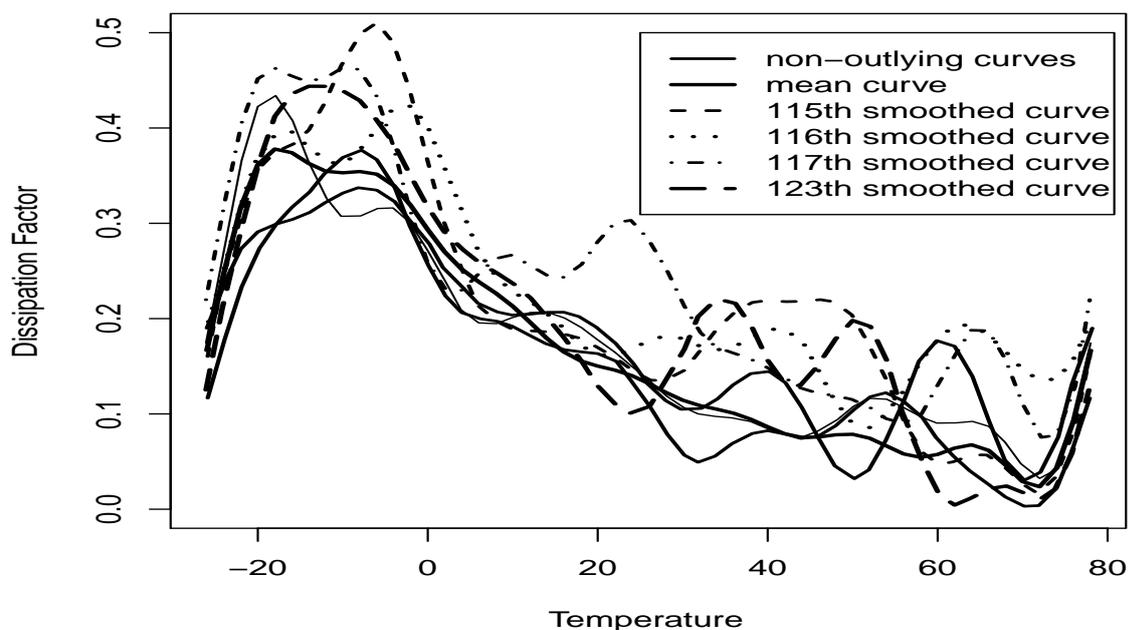
Figure 4: The first three non-outlying curves, the mean curve and the 115th, 116th, 117th and 123th curves after smoothing

By this detection result we have the following two findings: firstly, the first ninety-six AEC curves in this example can be regarded as in-control functions. This justifies the use of these curves as the training (reference) dataset to fit the model and design the control chart in Qiu et al. (2010). Secondly, by the monitoring result in Qiu et al. (2010), the outliers we found are contained in the set of OC curves (actually 112-120). Some other OC curves are not identified by our approach, such as 112-114. This is not surprising to

us because the control chart in Qiu et al. (2010) is designed to be efficient in detecting the step-change. Their charts borrow strength across multiple curves while our approach is individually testing each curve. The detection result for the AEC curves indicates that our proposed procedure performs well in applications.

5   Concluding remarks

Data sets with multiple outliers or clusters of outliers are subject to masking and swamping effects (Barnett and Lewis 1994; Pena and Prieto 2001). Similar to classical outlier detection methods, our procedure assumes the data contain only one outlier in each retrospective step and thus the power may decrease if the percentage of outlying function curves in the data is high as shown in the simulation study. In some situations, our method does not always succeed in detecting most of outliers, simply because it is affected by the functional observations that it is supposed to identify. Although the proposed two-step procedure seems to work well, a systematic method which is able to handle this issue to certain extent is needed. Moreover, an ongoing effort of the authors is to develop a scheme integrating a "data-driven" adaptive smoothing parameter selection method to improve the performance of FDOT in situations involving masking and swamping.

In addition, our simulation results (not reported here) show that the considered approach for choosing d, cumulative percentage variance, may not produce a most powerful test. This finding is not surprising because such a criteria is tailored for estimation and similar conclusions have been made in some other testing problems. For instance, in the nonparametric regression testing problem, the power and size of a typical test would depend on the bandwidth used in regression function estimation, and it is recognized that optimal bandwidth for nonparametric curve estimation may not be optimal for testing (cf., Hart 1997). See also Berkes et al. (2009) for a related discussion. An ongoing effort of the authors is to develop a more proper adaptive selection of d to make the test nearly optimal.

Appendix: The technical conditions used in Theorem 2

In order to establish the consistency of the proposed test under the alternative hypothesis, we also need the following additional conditions:

(C3) Under H_1, denote the number of outliers as m_N, we assume that

$$\frac{\sum_{i \in \mathcal{A}_N} \|\mu_i(t)\|^2}{m_N} = O(1), \quad \text{as} \quad N \to \infty.$$

(C4) (The condition on the number of outliers) Under H_1, the number of outliers satisfies

$$m_N \to \infty, \quad \frac{m_N \sqrt{\log(m_N)}}{N} \to 0, \quad \text{as} \quad N \to \infty.$$

(C5) (The condition on the means of outlying curves) Define

$$\delta_{ik} = \frac{1}{\sqrt{\lambda_k}} \int_a^b (\mu_i(t) - \mu_0(t)) \upsilon_k(t) dt.$$

Under H_1, we assume that

$$\liminf_{N \to \infty} \frac{\max_{i \in \mathcal{A}_N} \sum_{k=1}^d \delta_{ik}^2}{\log N} > 8 + 4\sqrt{3}.$$

Remark 3 Condition (C3) is fairly common technical assumption for the study of the consistency of the test. Similar to the outlier test for the multivariate samples (Hadi 1992), our proposed test is not consistent when assuming that there is only one outlier in the sample. In order to construct the consistency of our proposed test, Conditions (C4) and (C5) are required. Condition (C4) is reasonable because the number of outliers $|\mathcal{A}_N| = m_N$ always grows with the number of sample N in many cases. Furthermore, the number of outliers m_N cannot be too large in order to help us distinguish the outliers from "normal" data. Condition (C5) is a purely technical condition which guarantees that the mean of outlying functions dominates the chance variability caused by the random error.

References

Barnett, V., and Lewis, T. (1994), *Outliers in Statistical Data*, John Wiley, New York.

Benko, M., Härdle, W. and Kneip, A. (2009), "Common Functional Principal Components," *The Annals of Statistics*, 37, 1–34.

Berkes, I., Gabrys, R., Horváth, L. and Kokoszka, P. (2009), "Detecting Changes in the Mean of Functional Observations," *Journal of Royal Statistical Society: Series B*, 71, 927–946.

Bosq, D. (2000), *Linear Processes in Function Spaces*, Springer, New York.

Cai, T., Zhang, C., and Zhou, H. (2010), "Optimal Rates of Convergence for Covariance Matrix Estimation,", *The Annals of Statistics*, 38, 2118–2144.

Cuevas, A., Febrero, M. and Fraiman, R. (2004), "An ANOVA Test for Functional Data," *Computational Statistics & Data Analysis*, 47, 111–122.

Fawcett T., and Provost F. (1997), "Adaptive fraud detection," *Data-mining and Knowledge Discovery*, 1, 291–316.

Febrero, M., Galeano, P., and González-Manteiga, W. (2008), "Outlier Detection in Functional Data by Depth Measures, with Application to Identify Abnormal NOx Levels," *Environmetrics*, 19, 331–345.

Hadi, A. S. (1992), "Identifying Multiple Outliers in Multivariate Data," *Journal of the Royal Statistical Society: Series B*, 54, 761–771.

Hall, P., Hosseini-Nasab, M. (2006), "On Properties of Functional Principal Components Analysis," *Journal of the Royal Statistical Society: Series B*, 68, 109–126.

Hall, P, and Van Keilegom, I. (2007), "Two-Sample Tests in Functional Data Analysis Starting From Discrete Data," *Statistica Sinica*, 17, 1511–1531.

Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*, Springer, New York.

Hawkins, D. (1980), *Identification of Outliers*, Chapman and Hall, New York.

Hyndman, R. J., and Ullah, Md. S. (2007), "Robust Forecasting of Mortality and Fertility Rates : A Functional Data Approach," *Computational Statistics & Data Analysis*, 51, 4942–4956.

James, G. M., Hastie, T. J., and Sugar, C. A. (2000), " Principal Component Models for Sparse Functional Data," *Biometrika*, 87, 587–602.

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983), *Extremes and Related Properties of Random Sequences and Processes*, Springer, New York.

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., and Cohn, K. L. (1999), "Robust Principal Component Analysis for Functional Data (with discussion)," *Test*, 8, 1–73.

Pena, D., and Prieto, F. J. (2001), "Multivariate Outlier Detection and Robust Covariance Matrix Estimation", *Technometrics*, 43, 286–310.

Qiu, P., Zou, C., and Wang, Z. (2010), "Nonparametric Profile Monitoring By Mixed Effect Modeling (with discussions)," *Technometrics*, in press.

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis*, Springer, New York.

Rousseeuw P., and Leory A., (1987), *Robust Regression and Outlier Detection*, John Wiely& Sons, New York.

Rousseeuw P., van Zomeren B., (1990) "Unmasking multivariate outliers and leverage points," Journal of the American Statistical Association, 85, 633-639.

Spitzner, D. J., Marron, J. S., and Essick, G. K. (2003), "Mixed-Model Functional ANOVA for Studying Human Tactile Perception," *Journal of the American Statistical Association*, 98, 263–272.

Woodall, W. H., Spitzner, D. J., Montgomery, D. C., and Gupta, S. (2004), "Using Control Charts to Monitor Process and Product Quality Profiles," *Journal of Quality Technology*, 36, 309–320.

Wu, H., and Zhang, J. (2002), "Local Polynomial Mixed-Effects Models for Longitudinal Data," *Journal of the American Statistical Association*, 97, 883–897.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005a), "Functional Linear Regression Analysis for Longitudinal Data," *The Annals of Statistics*, 33, 2873–2903.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005b), "Functional Data Analysis for Sparse Longitudinal Data," *Journal of the American Statistical Association*, 100, 577-590.

Yeh, A. B., Huwang, L., and Li, Y. M. (2009), "Profile Monitoring for Binary Response," *IIE Transactions*, 41, 931–941.

Zou, C., Qiu, P., and Hawkins, D. M. (2009), "Nonparametric Control Chart for Monitoring Profile Using the Change Point Formulation," *Statistica Sinica*, 19, 1337–1357.

Zou, C., Tsung, F., and Wang, Z. (2007), "Monitoring General Linear Profiles Using Multivariate EWMA Schemes," *Technometrics*, 49, 395–408.

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

Supplemental file

In this supplemental file, we provide the proofs of Theorems 1 and 2 in the paper.

Proofs of theorems

In order to prove the theorems, we first state a necessary lemma. Denote

$$\phi_{ik} = \int_a^b \{X_i(t) - \bar{X}_N(t)\}\upsilon_k(t)dt, \quad i = 1, \dots N, \quad k = 1, \dots d,$$

$$\bar{Y}_N(t) = \frac{1}{N}\sum_{1\le i\le N} Y_i(t), \quad \bar{\xi}_k = \frac{1}{N}\sum_{1\le i\le N}\xi_{ik}, \quad k = 1,\dots d.$$

Lemma 1 *Suppose that $Z_1, \dots, Z_N$ are i.i.d. chi-square random variables with d degrees of freedom, then for each $x \in \mathbb{R}$, we have*

$$P(\max_{1\le i\le N} Z_i \le 2x + 2\log N + (d-2)\log\log N - 2\log\Gamma(d/2)) \to e^{-e^{-x}} \quad as \quad N \to \infty.$$

Proof. Denote $U_N = 2\log N + 2x + (d-2)\log\log N - 2\log\Gamma(d/2)$ and $F(x)$ as the cumulative distribution function of the χ_d^2 random variable. We have

$$N(1 - F(U_N)) = N\int_{U_N}^\infty \frac{e^{-u/2}u^{d/2-1}}{2^{d/2}\Gamma(d/2)}du = \frac{e^{-x}}{2^{d/2}}\int_0^\infty \left(\frac{t + U_N}{\log N}\right)^{d/2-1} e^{-t/2}dt$$

$$\xrightarrow{N\to\infty} \frac{e^{-x}}{2^{d/2}}\int_0^\infty 2^{d/2-1}e^{-t/2}dt = e^{-x}.$$

Hence the assertion follows immediately from Theorem 1.5.1 of Leadbetter et al. (1983). $\square$

Proof of Theorem 1

The main idea of the proof is similar to Berkes et al. (2009). Recall that $Y_i(t) = \sum_{1\le k<\infty}\lambda_k^{1/2}\xi_{ik}\upsilon_k(t)$ and $\{\xi_{ik}, i = 1, \dots N, k = 1, 2, \dots\}$ are i.i.d. standard normal random variables. As an immediate corollary to Lemma 1, we obtain for each k

$$\max_{1\le i\le N}\xi_{ik}^2 = O_p(\log N).$$

By using the inequality

$$\left|\max_{1\le i\le N}|a_i| - \max_{1\le i\le N}|b_i|\right| \le \max_{1\le i\le N}|a_i - b_i|,$$

we have

$$\left| \max_{1 \leq i \leq N} \sum_{k=1}^{d} \frac{\hat{\eta}_{ik}^2}{\hat{\lambda}_k} - \max_{1 \leq i \leq N} \sum_{k=1}^{d} \frac{\hat{\eta}_{ik}^2}{\lambda_k} \right| \leq \max_{1 \leq i \leq N} \left| \sum_{k=1}^{d} \hat{\eta}_{ik}^2 \left(\frac{1}{\hat{\lambda}_k} - \frac{1}{\lambda_k} \right) \right|$$

$$\leq \sum_{k=1}^{d} \left| \frac{1}{\hat{\lambda}_k} - \frac{1}{\lambda_k} \right| \max_{1 \leq i \leq N} \hat{\eta}_{ik}^2. \tag{A.1}$$

Also, observe that

$$\max_{1 \leq i \leq N} \hat{\eta}_{ik}^2 = \max_{1 \leq i \leq N} \left(\int_a^b (Y_i(t) - \bar{Y}_N(t)) \hat{v}_k(t) dt \right)^2 \leq \max_{1 \leq i \leq N} \| Y_i(t) - \bar{Y}_N(t) \|^2$$

$$\leq 2 \max_{1 \leq i \leq N} \| Y_i(t) \|^2 + 2 \| \bar{Y}_N(t) \|^2 \leq 2 \sum_{k=1}^{\infty} \lambda_k \max_{1 \leq i \leq N} \xi_{ik}^2 + 2 \| \bar{Y}_N(t) \|^2. \tag{A.2}$$

Since $\| \bar{Y}_N(t) \|^2 = o_p(1)$, $\max_{1 \leq i \leq N} \xi_{ik}^2 = O_p(\log N)$ and $\sum_{k=1}^{\infty} \lambda_k = E(\| Y_i(t) \|^2) < \infty$, we obtain that $\max_{1 \leq i \leq N} \hat{\eta}_{ik}^2 = O_p(\log N)$. Combining expressions (A.1), (A.2) and the equation (4), we have

$$\left| \max_{1 \leq i \leq N} \sum_{k=1}^{d} \frac{\hat{\eta}_{ik}^2}{\hat{\lambda}_k} - \max_{1 \leq i \leq N} \sum_{k=1}^{d} \frac{\hat{\eta}_{ik}^2}{\lambda_k} \right| = o_p(1). \tag{A.3}$$

Next, we have

$$\left| \max_{1 \leq i \leq N} \sum_{k=1}^{d} \frac{\hat{\eta}_{ik}^2}{\lambda_k} - \max_{1 \leq i \leq N} \sum_{k=1}^{d} \frac{\phi_{ik}^2}{\lambda_k} \right| \leq \max_{1 \leq i \leq N} \left| \sum_{k=1}^{d} \frac{\hat{\eta}_{ik}^2 - \phi_{ik}^2}{\lambda_k} \right|$$

$$\leq \sum_{k=1}^{d} \max_{1 \leq i \leq N} |\hat{\eta}_{ik} - \hat{c}_k \phi_{ik}| \cdot \frac{\max_{1 \leq i \leq N} |\hat{\eta}_{ik}| + \max_{1 \leq i \leq N} |\phi_{ik}|}{\lambda_k}$$

$$\leq \sum_{k=1}^{d} \| \hat{v}_k(t) - \hat{c}_k v_k(t) \| \cdot 2 \left(\max_{1 \leq i \leq N} \| Y_i(t) - \bar{Y}_N(t) \| \right)^2.$$

Combining the results of Bosq (2000)

$$\limsup_{N \to \infty} \left[N(E(\| \hat{c}_k v_k(t) - \hat{v}_k(t) \|^2)) \right] < \infty,$$

$$\limsup_{N \to \infty} \left[N(E(|\lambda_k - \hat{\lambda}_k|^2)) \right] < \infty,$$

and the fact that $\max_{1 \leq i \leq N} \| Y_i(t) - \bar{Y}_N(t) \|^2 = O_p(\log N)$, we have

$$\left| \max_{1 \leq i \leq N} \sum_{k=1}^{d} \frac{\hat{\eta}_{ik}^2}{\lambda_k} - \max_{1 \leq i \leq N} \sum_{k=1}^{d} \frac{\phi_{ik}^2}{\lambda_k} \right| = o_p(1). \tag{A.4}$$

2

Finally, we have

$$\left| \max_{1 \le i \le N} \sum_{k=1}^{d} \frac{\phi_{ik}^2}{\lambda_k} - \max_{1 \le i \le N} \sum_{k=1}^{d} \xi_{ik}^2 \right| = \left| \max_{1 \le i \le N} \sum_{k=1}^{d} (\xi_{ik} - \bar{\xi}_k)^2 - \max_{1 \le i \le N} \sum_{k=1}^{d} \xi_{ik}^2 \right|$$

$$\le \sum_{k=1}^{d} \bar{\xi}_k^2 + 2 \sum_{k=1}^{d} \max_{1 \le i \le N} |\xi_{ik}| \cdot |\bar{\xi}_k|$$

Since $\sqrt{N} \bar{\xi}_k \xrightarrow{d} N(0, 1)$ and $\max_{1 \le i \le N} |\xi_{ik}| = O_p(\sqrt{\log N})$, we obtain that

$$\left| \max_{1 \le i \le N} \sum_{k=1}^{d} \frac{\phi_{ik}^2}{\lambda_k} - \max_{1 \le i \le N} \sum_{k=1}^{d} \xi_{ik}^2 \right| = o_p(1). \tag{A.5}$$

Equations (A.3), (A.4), (A.5) and Lemma 1 yield the result in Theorem 1. $\qquad \square$

Proof of Theorem 2

Denote $\mu_N^*(t) = \frac{1}{m_N} \sum_{i \in \mathcal{A}_N} \mu_i(t)$, $\psi_N^*(t) = \mu_N^*(t) - \mu_0(t)$ and $\psi_i(t) = \mu_i(t) - \mu_0(t)$ for each $i \in \mathcal{A}_N$. Firstly, under H_1, we have

$$\hat{c}(t, s) = \frac{1}{N} \sum_{i=1}^{N} (X_i(t) - \bar{X}_N(t))(X_i(s) - \bar{X}_N(s))$$

$$= \frac{1}{N} \sum_{i \in \mathcal{A}_N} (Y_i(t) - \bar{Y}_N(t) + \psi_i(t) - \frac{m_N}{N} \psi_N^*(t))(Y_i(s) - \bar{Y}_N(s) + \psi_i(s) - \frac{m_N}{N} \psi_N^*(s))$$

$$+ \frac{1}{N} \sum_{i \notin \mathcal{A}_N} (Y_i(t) - \bar{Y}_N(t) - \frac{m_N}{N} \psi_N^*(t))(Y_i(s) - \bar{Y}_N(s) - \frac{m_N}{N} \psi_N^*(s))$$

By the central limit theorem for the Banach space (e.g., Hall and Hosseini-Nasab 2006) and Conditions (C3) and (C4), we obtain $\hat{c}(t, s) \xrightarrow{P} c(t, s)$. Therefore, under H_1, we also have

$$\limsup_{N \to \infty} [N\{E(\|\hat{c}_k v_k(t) - \hat{v}_k(t)\|^2)\}] < \infty,$$

$$\limsup_{N \to \infty} [N\{E(|\lambda_k - \hat{\lambda}_k|^2)\}] < \infty.$$

Define

$$\tilde{\phi}_{ik} = \begin{cases} \int_a^b (Y_i(t) - \bar{Y}_N(t)) v_k(t) dt + \lambda_k^{1/2} \delta_{ik}, & i \in \mathcal{A}_N, k = 1, \ldots, d, \\ \int_a^b (Y_i(t) - \bar{Y}_N(t)) v_k(t) dt, & i \notin \mathcal{A}_N, k = 1, \ldots, d. \end{cases}$$

Similar to the proof of Theorem 1, we can obtain that

$$\max_{1 \le i \le N} \left| \sum_{k=1}^{d} \frac{\hat{\eta}_{ik}^2}{\hat{\lambda}_k} - \sum_{k=1}^{d} \frac{\tilde{\phi}_{ik}^2}{\lambda_k} \right| = o_p(1). \tag{A.6}$$

3

Note that

$$\sum_{k=1}^{d} \frac{\tilde{\phi}_{ik}^2}{\lambda_k} = \begin{cases} \sum_{k=1}^{d}(\xi_{ik} - \bar{\xi}_k + \delta_{ik})^2, & i \in \mathcal{A}_N, k = 1, \ldots, d, \\ \sum_{k=1}^{d}(\xi_{ik} - \bar{\xi}_k)^2, & i \notin \mathcal{A}_N, k = 1, \ldots, d, \end{cases}$$

and

$$\max_{1 \leq i \leq N} \sum_{k=1}^{d} \frac{\tilde{\phi}_{ik}^2}{\lambda_k} \geq \max_{i \in \mathcal{A}_N} \sum_{k=1}^{d}(\xi_{ik} - \bar{\xi}_k + \delta_{ik})^2$$

$$\geq \max_{i \in \mathcal{A}_N} \sum_{k=1}^{d} \delta_{ik}^2 - \max_{1 \leq i \leq N} \sum_{k=1}^{d}(\xi_{ik} - \bar{\xi}_k)^2 - 2\left| \max_{1 \leq i \leq N} \sum_{k=1}^{d} \delta_{ik}^2 \cdot \max_{1 \leq i \leq N} \sum_{k=1}^{d}(\xi_{ik} - \bar{\xi}_k)^2 \right|^{1/2}.$$

By the Condition (C5) and the fact that $\max_{1 \leq i \leq N} \sum_{k=1}^{d}(\xi_{ik} - \bar{\xi}_k)^2 = O_p(2 \log N)$, we conclude that for each $\alpha \in (0, 1)$,

$$P_{H_1}\left\{ \max_{1 \leq i \leq N} \sum_{k=1}^{d} \frac{\tilde{\phi}_{ik}^2}{\lambda_k} > u_{N,d}(\alpha) \right\} \to 1.$$
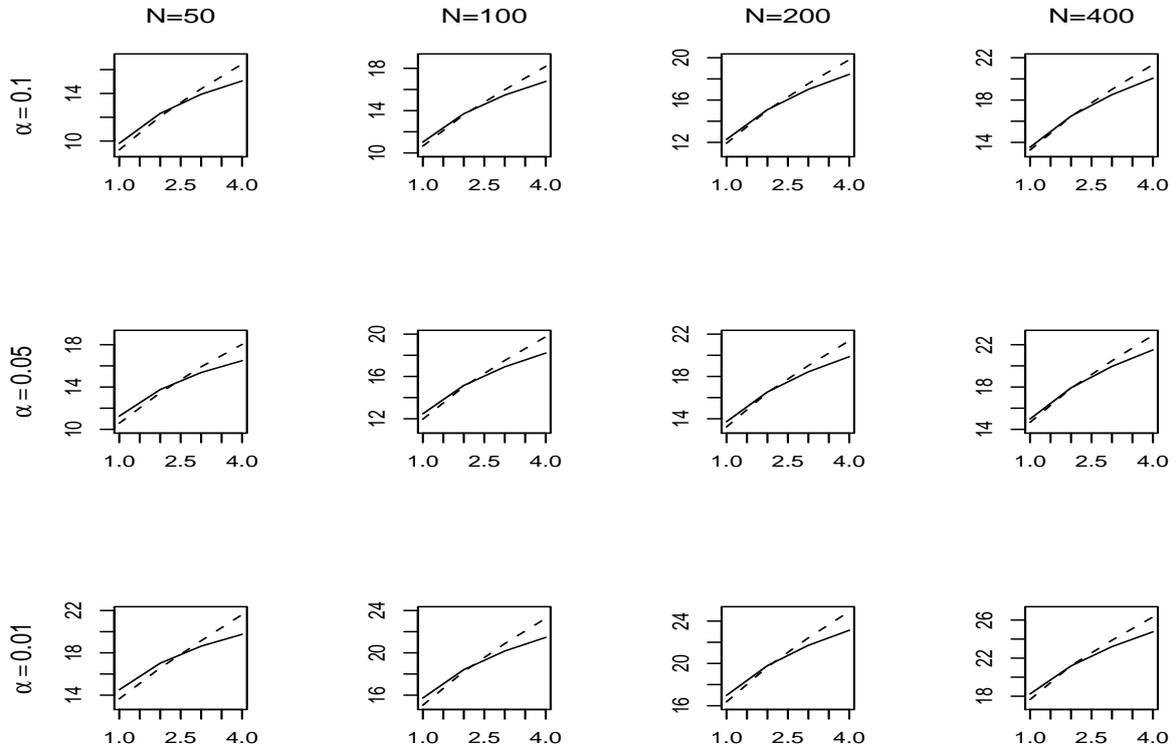
Theorem 2 now follows from (A.6). □

4

Figure A.1: Critical values for various N, d and α. The label for the x-axis is d and the label for the y-axis is critical value. The solid line is $U_{N,d}$ and the dashed line is $G_{N,d}$.