A One-Class Peeling Method for Multivariate Outlier Detection with Applications in Phase I SPC

Waldyn G. Martinez, Maria L. Weese, L. Allison Jones-Farmer

Department of Information Systems & Analytics, Miami University

Abstract

In Phase I of Statistical Process Control (SPC), control charts are often used as outlier detection methods to assess process stability. Many of these methods require estimation of the covariance matrix, are computationally infeasible, or have not been studied when the dimension of the data, p, is large. We propose the one-class peeling (OCP) method, a flexible framework that combines statistical and machine learning methods to detect multiple outliers in multivariate data. The OCP method can be applied to Phase I of SPC, does not require covariance estimation and is well-suited to high-dimensional datasets with a high percentage of outliers. Our empirical evaluation suggests that the OCP method performs well in high dimensions and is computationally more efficient and robust than existing methodologies. We motivate and illustrate the use of the OCP method in a Phase I SPC application on a $N = 354$, $p = 1,917$ dimensional dataset containing Wikipedia search results for NFL players, teams, coaches, and managers. The example data set and R functions, *OCP.R* and *OCPLimit.R*, to compute the respective OCP distances and thresholds are available in the supplementary materials.

KEYWORDS: Control Chart, Gaussian Kernel, High Dimensional Data, Kernel Distance, One-Class Methods, Support Vector Data Description

1 Introduction

1.1 Background

Statistical Process Control (SPC) applications are often divided into two phases. In Phase I, the stability of a process is assessed, and a critical component of this assessment is often the application of an outlier detection method. Traditionally, control charts have been used in Phase I to identify outlying observations or subgroups that deviate from the normal operations of the process. These outlying observations or subgroups are signals to a potential out-of-control (OC) event. Overviews of Phase I methods in SPC are given in Jones-Farmer et al.[1] and Chakraborti et al.[2] . In Phase II, the process is monitored over time for departures from the Phase I conditions.

The detection of outliers is a critical component of the Phase I or any data analysis process. Nearly all "real" data may be contaminated by unusual observations; thus, preprocessing the data should precede any data analysis exercise. In some cases, such as fraud detection, network intrusion, or crime identification, outlier detection may be the end-goal of the analysis. In other cases, like in Phase I of SPC, outlier detection may only comprise an initial exploratory phase of an analysis.

One of the goals of a Phase I analysis is to establish an in-control (IC) baseline sample; however, a requirement for many Phase I and outlier detection methods to work is knowledge of the data distribution. Recently, several authors have proposed distribution-free Phase I methods to identify outliers, OC subgroups or changes in multivariate processes including Bell et al.[3], Cheng and Shiau[4], and Capizzi[5]. Each of these methods have been studied for multivariate processes with relatively low dimensions. For example, Bell et al.[3] and Capizzi[5] considered only dimensions up to $p = 10$, while Cheng and Shiau[4] considered cases of $p = 3$. In many modern processes such as real time health monitoring[6], additive manufacturing and image monitoring[7], much higher dimensional data is common.

The goal of this article is to develop the one-class peeling (OCP) method to detect outliers in high dimensions that works efficiently and competitively with wide data ($p > N$), does not require estimation of a covariance matrix, and can be applied in Phase I of SPC to establish an IC baseline sample. Like many methods, the OCP method performs best when the underlying data distribution is known; however, the method is somewhat robust to distribution misspecification. We also offer an approach for determining outliers that has suitable performance when the underlying data distribution is unknown. We apply the OCP method to detect outliers in a $p = 1,917$ dimensional dataset ($N = 354$) containing Wikipedia searches for the National Football League (NFL). In addition, we conduct a simulation study to show the benefits of using the OCP method versus existing methods for outlier detection in high dimensions.

1.2 Motivating Example

This article is motivated by the need for methods to determine an outlier-free baseline Phase I sample for high-dimensional multivariate processes that can be used to develop a Phase II monitoring scheme. We consider an updated version of the NFL Wikipedia search data that originally appeared in Weese et al.[8]. Monitoring of web searches and social media data for changes in volume, content, or sentiment is becoming increasingly important; however, much of this data is high dimensional, and few methodologies exist that address the retrospective outlier detection problem in this context.

The data is based on Wikipedia activity of NFL-related pages. A dictionary of NFL team names, coaches, managers, and active players was gathered as of 9/15/2014. The number of Wikipedia searches per hour for any Wikipedia page can be downloaded from `http://dumps.wikimedia.org/other/pagecounts-raw`, and we did so for all terms in our dictionary for the periods ranging between 9/01/2014 00:00 UTC (Coordinated Universal Time) and 9/14/2014 17:00 UTC. These dates were chosen because the 2014 season had several high- and low-profile controversies that could be retrospectively identified (see Table C.2). For example, Adrian Peterson, former running back of the Minnesota Vikings, was indicted on child abuse charges in September 2014. The data consists of $p = 1,917$ variables (pages) and $N = 354$ observations (hours).

This data was first analyzed in Weese et al.[8], who modified the K^2-chart of Sukchotrat et al.[9] for use in Phase I and applied it to the first week of the data in order to establish a baseline sample. The Phase II K^2-chart was then applied to the second week of the data, but many of the events known to occur early in this NFL season were not detected. The goal of the analysis in this article is to establish an outlier-free baseline sample in Phase I so that a Phase II monitoring scheme can be developed.

1.3 Background Literature

The literature on multivariate outlier detection methods is vast. Although many of these methods are developed for cross-sectional data, some can be adapted to data that occurs over time, and may be appropriate for use in determining a baseline sample in Phase I. Many multivariate, distance-based outlier detection methods follow two main steps: (1) robust estimation of center and scale of the data; and (2) evaluation of a measure of "outlyingness", often a distance measure. One challenge with this approach is that the existence of outliers can negatively affect the sample estimates of the center and scale making it difficult to correctly identify unusual observations. A common solution is to assume the multivariate normal model and use robust estimators of the center, $\boldsymbol{\mu}$, and spread, $\boldsymbol{\Sigma}$, and also a robust distance measure[10–17]. A well-known method is to use the Minimum Covariance Determinant (MCD) or the Minimum Volume Ellipsoid (MVE) estimators with robust squared distance measures to identify unusual observations[14,18]. These methods can be computationally intensive in high dimensions[16,19,20], and the statistical properties of the distance measures are asymptotic, which can limit their ability to detect outliers in finite samples[21,22]. The Finite Sample Reweighted Minimum Covariance Determinant (FSRMCD) method was developed by Cerioli[23] and is based on the reweighted version of the MCD[16,24]. The FSRMCD performs well in finite samples under conditions of multivariate normality. Like other methods based on squared statistical distance measures[19,25], the FSRMCD cannot be implemented on wide data.

Most of the distance-based methods require covariance estimation and cannot be directly applied when $p > N$ without the use of principal components. A regularization term with the MCD approach that can be used when $p > N$ is suggested in Fritsch et al.[26], but this method is only evaluated under multivariate normality. A Robust Minimum Diagonal Product (RMDP) estimator of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is suggested by Ro et al.[27]. In this method, the statistical distances used to measure outlyingness are computed from only the diagonal elements of the sample covariance matrix. This method can be applied when $p > N$ and works well when the distribution is close to multivariate normal and the covariance matrix is sparse, but has a high false positive rate when these conditions are not met. Another approach to analyze wide data is to reduce the dimensionality, often by means of principal component analysis (PCA)[28,29]. For example, Filzmoser et al.[20] developed a procedure known as PCOut that combines robust PCA with weighted distance measures to identify outliers.

Data depth was introduced by Tukey[30] as a way to measure the depth or centrality of a point within a sample. The data depth of $\mathbf{x} \in \mathbb{R}^p$ measures how deep $\mathbf{x}$ is with respect to a certain probability distribution F. There are a number of data depth functions that transform $\mathbf{x}$ from a $p-$dimensional vector to a univariate depth value. These depth values vary with respect to computational complexity and mathematical properties. For outlier detection, depth values are often ranked. Ranks of simplicial depth were applied by Liu[31] to Phase II SPC and studied by Stoumbos and Jones[32]. Rank-based Phase I charts for subgrouped data based on data depth were considered by Bell et al.[3]. While their method can be used with any measure of data depth, they specifically considered Mahalanobis depth and simplicial depth. The method in Bell et al.[3] was better able to correctly classify a higher percentage of location shifts when using Mahalanobis depth than when using simplicial depth. Although Mahalanobis depth is easy to compute, it requires computation of the covariance matrix and is not feasible for cases when $p > N$. The order of computation, $O(N^{p+1})$, for simplicial depth makes it impractical for applications in large samples with high dimensions. Although there have been successful attempts at decreasing the computational complexity of simplicial depth, this progress has been made in lower dimensions ($p = 2$)[33].

A number of authors have proposed cluster-based algorithms for outlier detection, and some of these methods have been applied to SPC. For example, Thissen et al.[34] considered the application of Gaussian mixture models to describe non-normally distributed data as a mixture of clusters. A multi-step algorithm for SPC to overcome the problem of masking, or multiple outliers that effect the parameter estimates in a retrospective analysis was developed by Jobe and Pokojovy[35] . This method, which requires estimation of the covariance matrix, combines a modified moving average to smooth the data, multivariate density estimation, and mixture-based clustering to identify the cluster with the largest number of observations. Simulation results suggest that Jobe and Pokojovy[35]'s method detects outliers with a higher probability than Hotelling's T^2 chart when there are multiple outliers and the shift size is large. The work of Jobe and Pokojovy[35] was extended by Jobe and Pokojovy[36] by including the FSRMCD estimator of Cerioli[23]. This cluster-based method, requires two sets of simulated limits to control for family-wise and per-comparison error rates, which are determined from multivariate standard normal data. Because these methods require estimation of the covariance matrix, they are not directly applicable when $p > N$.

An outlier detection method similar to those based on statistical distance is based on the concept of density estimation methods. Support Vector Data Description (SVDD) was introduced by Tax and Duin[37] as method of multivariate kernel density estimation and applied this to the outlier detection problem. SVDD, which is discussed in detail in section 2.1, finds a flexible minimum volume boundary with radius, R around a multivariate set of data. Observations whose distance from the center, $\boldsymbol{\mu}$, exceed the radius of the boundary, $\|\mathbf{x}_i - \boldsymbol{\mu}\|^2 > R^2$, are considered outlying. The SVDD method was adapted to SPC by Sun and Tsung[38], which developed the k-chart. It is shown in Weese et al.[39] that the k-chart does not perform well in distinguishing IC from OC observations in Phase I.

1.4 OCP Overview

Our proposed method, the OCP, borrows from the kernel-density approaches and the statistical distance approaches. The method begins with the iterative application of one-class algorithms such as the SVDD methodology proposed by Tax and Duin[37],[40] to estimate the center of the data. Scaled kernel distances of each observation from the estimated center are then used to identify potential outlying observations. The OCP method does not require estimation of the covariance matrix. In addition, OCP is data-driven and can be used with any kernel function that is appropriate for the data. The rest of the paper is organized as follows. In Section 2, we introduce the OCP method. In Section 3, we apply the OCP method to the NFL Wikipedia search dataset. We evaluate and compare the performance of the OCP method with the modified RMDP method of Ro et al.[27] using simulations in Section 4. In Section 5 we introduce an a robust approach for determining limits for the OCP when the distribution of the data is unknown. Finally, in Section 6 we discuss the flexibility and potential limitations of the proposed method and provide our concluding remarks.

2 OCP Method

The OCP method includes robust estimation of the center and the use of a kernel distance measure between each observation and the center. To detect outliers, we recommend using a threshold applied to the kernel distance measures. Observations that are the most dissimilar from the center data are flagged as potential OC events. Algorithm 1 gives the pseudocode for the OCP method.

2.1 Estimating the Center of the Multivariate Data

The first step in using the OCP method is to determine the center of the multivariate data. We determine the center of data using an iterative peeling approach with the boundaries derived from SVDD, similar in principle to that of convex hull peeling[41].

The mean estimated by convex hull peeling is an affine equivariant estimator, but it is not robust to outliers. The robustness of estimators is often measured in terms of the Finite Sample Replacement Breakdown Point (FSRBP). The FSRBP of a location estimator, $\mathbf{t}_N$, based on a data set, S, is the smallest fraction, m/N, of outliers that can take the estimate "over all bounds"[42]:

$$\epsilon^*(\mathbf{t}_N, S) = \min_{1 \leq m \leq N} \left\{ \frac{m}{N} : \sup_{S_{\text{O}}} ||\mathbf{t}_N(S) - \mathbf{t}_N(S_{\text{O}})|\,| = \infty \right\},$$

where the supremum is taken over all possible corrupted samples, $S_{\text{O}} = \{\mathbf{y}_1, \cdots, \mathbf{y}_m\}$, for $\mathbf{y}_i \in \mathbb{R}^p$, obtained by replacing m points from S with arbitrary values. It is shown in Donoho and Gasko[43] that convex hull peeled means can have a FSRBP of no more than $\frac{1}{p+1}$, which is not robust for data in very high dimensions. We introduce a more robust method for estimating the multivariate center using boundaries determined from SVDD.

In Tax and Duin[40] SVDD is introduced as a method for both outlier detection and classification. SVDD is a one-class method that finds a flexible minimum volume boundary around a multivariate data set via optimization. Similar to Support Vector Classification[44], the SVDD boundary is defined only by a few points based on their proximity to the center of the data. These points are referred to as the support vectors. One obtains a SVDD hypersphere with radius R for data vectors, $\mathbf{x}_i \in \mathbb{R}^p, i = 1, ..., N$, by minimizing:

$$\begin{aligned} &\underset{R,\xi_i}{\text{minimize}} \quad F(R, \xi_i) = R^2 + C\sum_i \xi_i \\ &\text{subject to} \quad \|\mathbf{x}_i - \boldsymbol{\mu}\|^2 \leq R^2 + \xi_i,\ \xi_i \geq 0\ \forall i. \end{aligned} \tag{1}$$

Here, $\boldsymbol{\mu}$ is the center of the hypersphere, ξ_i is an error term, and C is a parameter that controls the fraction of observations outside of the boundary. The parameter, $C \leq \frac{1}{Nq}$, where q is the desired fraction of observations rejected. For the OCP method, we set $q = 0.0001$, limiting the number of observations allowed to be outside the hypersphere at each peel to practically zero.

Using Lagrange multipliers, Tax and Duin[40] showed that incorporating the constraints from Equation (1) results in the need to maximize

$$L = \sum_i \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_j \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j) \tag{2}$$

with respect to α_i such that $0 \leq \alpha_i \leq C$. Maximizing Equation (2) results in a set of α_i such that when an observation, $\mathbf{x}_i$ satisfies $\|\mathbf{x}_i - \boldsymbol{\mu}\|^2 < R^2$ in Equation (1), the corresponding $\alpha_i = 0$, and these observations are inside of the boundary. When, $\|\mathbf{x}_i - \boldsymbol{\mu}\|^2 = R^2$, $\alpha_i > 0$ and these observations are the support vectors providing the description of the data and, thus, the one-class boundary. Observations where $\|\mathbf{x}_i - \boldsymbol{\mu}\|^2 > R^2$, are beyond the boundary and are also considered to be support vectors, but belong to the outlier class.

To construct a boundary with a flexible shape, the inner products in Equation (2) can be replaced by a kernel (similarity) function $\text{KS}_G(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$, where $\Phi(\mathbf{x})$ maps the vector $\mathbf{x}$ to a different feature space[45]. Commonly, SVDD is implemented using the Gaussian kernel function, defined as

$$\text{KS}_G(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{s^2}\right), \tag{3}$$

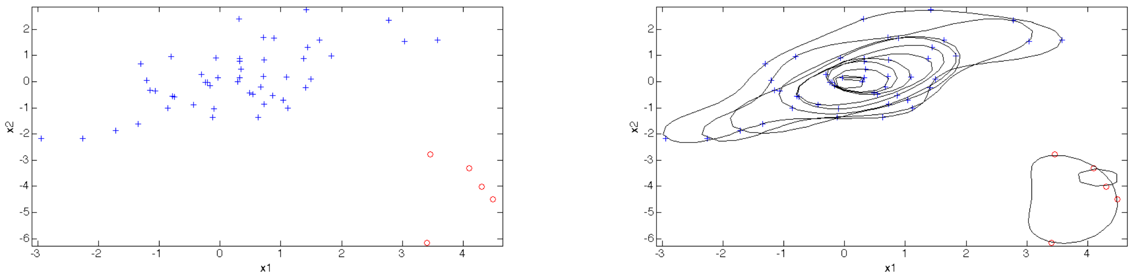

Figure 1: A toy example using a sample of $N = 50$, $p = 2$ correlated multivariate Normal data with 5 outlier points (red circles) and the corresponding SVDD boundaries.

where s is a scale parameter that must be specified in order to complete the optimization in Equation (2). Several authors have considered data-dependent methods for determining appropriate values of s for SVDD using the Gaussian kernel including Tax and Duin[40] and Ning and Tsung[46]. Weese et al.[39] shows that for a specified value of C, $s \approx p$ works well when the data are scaled to approximate unit variance. We present boundaries created with $s = p$ for the remainder of the paper.

To compute the proposed one-class peeled mean, $\hat{\boldsymbol{\mu}}_{\mathrm{OCP}}$, one constructs sequential one-class boundaries, peeling away the support vectors that create the boundaries at each iteration. After many successive peels, the remaining final observations are averaged to estimate the center of the data. Figure 1 shows the successive boundaries generated using the OCP method on toy data generated from $N_2\left(\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}\right)$ and 5 outlier points (shown as circles) from outlier distribution $N_2\left(\boldsymbol{\mu} + \boldsymbol{\delta} = \begin{bmatrix} 4 \\ -4 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}\right)$. Notice in the second graph in Figure 1 there are two separate boundaries around the outlying observations; thus, these observations were removed in the first two iterations. The remaining SVDD boundaries peeled the support vectors from the outside resulting in two remaining observations. We refer to the minimum number of observations remaining as n. The analyst can specify any number, $n < N$, for the size of the reduced dataset upon which to base the robust estimate. Our results suggest that peeling to a reduced set of $n = 2$ gives the highest FSRBP values in the cases we considered (see Appendix A).

Although it is not possible to analytically derive the FSRBP of the one-class peeling estimate of the mean, we perform simulations to study the empirical breakdown properties of this estimator for samples from a variety of multivariate normal, lognormal, and t-distributions. For a baseline comparison, the simulations also contain mean estimates from convex hull peeling. The results, given in Appendix A, suggest that empirically, the OCP estimator does not begin to breakdown until the sample contamination is around $30-35\%$ for the dimensions and sample sizes considered. Note, the convex hull peeling method breaks down empirically at less than 10% contamination.

2.2 Measuring Distance from the Estimated Center

The second step in the OCP method is to determine how close observations are to the estimated center of the data. To measure the distance between each observation, $\mathbf{x}_i$, and the center of the data, $\hat{\boldsymbol{\mu}}_{\text{OCP}}$, we propose the use of the Gaussian kernel distance given by

$$\text{KD}(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_{\text{OCP}}) = 1 - \text{KS}_G(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_{\text{OCP}}). \tag{4}$$

It is clear from Equation (3) that the Gaussian kernel similarity, KS_G, is a decreasing function of the Euclidean distance between two arbitrary points, $\mathbf{x}_i$ and $\mathbf{x}_j$, scaled by the width parameter s, and $\text{KS}_G(\mathbf{x}_i, \mathbf{x}_j) = 1$ for $i = j$.

The distance in Equation (4) is a half kernel distance such that $0 \leq \text{KD}(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_{\text{OCP}}) \leq 1$, with $\text{KD}(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_{\text{OCP}}) = 0$ if and only if $\mathbf{x}_i = \hat{\boldsymbol{\mu}}_{\text{OCP}}$. Thus, smaller values of $\text{KD}(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_{\text{OCP}})$ (closer to 0) indicate observations close to the estimated mean $\hat{\boldsymbol{\mu}}_{\text{OCP}}$, while larger values of $\text{KD}(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_{\text{OCP}})$ (closer to 1) indicate observations far away from $\hat{\boldsymbol{\mu}}_{\text{OCP}}$. While other distance measures could be used, we choose $\text{KD}(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_{\text{OCP}})$ because (1) it captures the underlying mechanism upon which the

SVDD boundaries were constructed, (2) does not require estimating $\mathbf{\Sigma}$, and (3) is derived from a positive-definite family of kernels with advantageous properties in learning non-linear spaces[47]. There are, however, some limitations the end-user needs to be aware of. Most importantly, the Gaussian Kernel similarity assumes that either the partial distributions of the variables are similar, or the data has been standardized so that variables with higher variances do not dominate the similarity metric.

2.3 Determining Outlyingness

To determine outlyingness, we recommend rescaling the kernel distances in Equation (4) using the following robust linear transformation

$$\mathrm{sRKD}_i = \frac{\mathrm{KD}(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_{\mathrm{OCP}}) - \mathrm{med}(\mathrm{KD})}{\mathrm{MAD}(\mathrm{KD})}. \tag{5}$$

Here, med(KD) is the median of the vector of N kernel distances, where N is the sample size, and MAD(KD) is the median absolute deviation of the N kernel distances. The MAD(KD) has the best possible breakdown (50% FSRBP) and is defined by Rousseeuw and Croux[48] for a sample $\{x_1, x_2, ..., x_n\} \in \mathbb{R}^p$ as

$$\mathrm{MAD} = \mathrm{med}_i |x_i - \mathrm{med}_j x_j|. \tag{6}$$

To detect potential OC observations, we recommend a threshold for the sRKD_i from Equation (5). Because the distribution of the sRKD_i varies due to the underlying data distribution, the threshold values, h, can be empirically determined to achieve an approximate Type I error rate using a bisection method (see Algorithm 2 in Appendix B). Finding the threshold, h using Algorithm 2 requires some knowledge of the underlying data distribution. If the underlying data distribution is unknown, an approximate robust value for h can be used as described in Section 5.

In the next section, we illustrate the use of the OCP method which includes a graph to help visually determine outliers. The visual implementation of this method strongly enhances its use in

Algorithm 1 The OCP Method

```
procedure OCP($S$, $n$, $h$)    ▷ $S = \{(\mathbf{x}_i \in \mathbb{R}^p), i = 1, ..., N\}$
    $r \leftarrow N$    ▷ assign $N$ to $r$
    $S_{\mathrm{r}} \leftarrow S$    ▷ assign $S$ to $S_r$
    while $r > n$ do    ▷ $n$ is min number of obs after final peel. $n = 2$ is recommended.
        $\mathbf{x}_{SV} \leftarrow \mathrm{SVDD}(S_{\mathrm{r}}, q = 0.0001, s = p)$    ▷ $\mathbf{x}_{SV}$ are the boundary support vectors
        $S_r \leftarrow S_r \backslash \mathbf{x}_{SV}$    ▷ update $S_r$ by dropping $\mathbf{x}_{SV}$
        $r \leftarrow \mathrm{rows}(S_{\mathrm{r}})$    ▷ update $r$ to number of rows of $S$
    end while
    $\hat{\boldsymbol{\mu}}_{\mathrm{OCP}} \leftarrow \mathrm{mean}(S_{\mathrm{r}})$
    $\mathrm{sRKD}_i \leftarrow \frac{\mathrm{KD}_i(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_{\mathrm{OCP}}) - \mathrm{med(KD)}}{\mathrm{MAD(KD)}}$
    $\mathrm{flag}_{OCP} \leftarrow I(\mathrm{sRKD}_i > h)$    ▷ $h$ is the threhold for desired Type I error rate
    return $\hat{\boldsymbol{\mu}}_{\mathrm{OCP}}$, $\mathrm{sRKD}_i$, $\mathrm{flag}_{OCP}$
end procedure
```

practice, and we recommend this approach. The OCP method is summarized as follows:

1. Determine a threshold value, h, using the bisection method (see Algorithm 2 in Appendix B). If the joint joint distribution of the $\mathbf{x}_i$ is unknown use a robust limit calculation (see Section 5).
2. Compute the robust estimate of the center of the data, $\hat{\boldsymbol{\mu}}_{\text{OCP}}$ using SVDD with the Gaussian kernel function to derive the boundaries and remove the support vectors at each peeling step.
3. Compute the kernel distance between each vector of observations and the robust estimate of the center of the data, $\hat{\boldsymbol{\mu}}_{\text{OCP}}$.
4. Robustly scale the distances are using Equation (5).
5. Flag observations with robustly-scaled kernel distances larger than h as potential outliers.

The R code implementation of the method is included in the supplemental material of this article as *OCP.R*. Algorithm 1 shows the pseudocode of the OCP method and describes the steps for flagging outliers in our simulation starting with a contaminated sample, S. The *OCP.R* function has as parameters the unscaled data, S, the minimum number of observations remaining, n, defaulting to two observations, the threshold value, h, that can be determined using a robust limit (see Section 5) as default or use the *OCPLimit.R* function (described in Algorithm 2 in Appendix B).

3 Analysis of the NFL Data

The NFL data consists of the number of Wikipedia hits on $p = 1,917$ pages representing NFL teams, players, coaches, managers, and active players between 09/01/2014 00:00 UTC and 09/15/2014 17:00 UTC. Due to a daily cyclical pattern in the data, we first preprocessed the data and used the residuals from an additive Holt-Winters model lagged by 24 hours[49]. Our goal is to determine a baseline dataset free of unusual events that could disrupt the normal Wikipedia traffic related to the NFL. This baseline sample can then be used to establish a method for monitoring for an increase/decrease in future traffic. Events that disrupt the normal Wikipedia traffic include:

- Games: interest in players, teams, coaches, and managers increase as the teams play normal and playoff games and consequently the Wikipedia pages have higher than usual traffic.
- Controversies: including player suspensions due to the use illegal substances, performance enhancing drugs, domestic disturbances, controversial postures and other high profile events.
- Breaking News: including updates to controversies and new developments.

In this example, we are comparing the OCP method to an existing outlier detection method for high dimensional data to determine how well it performs in identifying several disruptive events known to occur during the two-week study period. We used a structured methodology described in Table C.1 in Appendix C to identify time periods associated with games, controversies and breaking news. For example, game events were identified as one hour prior to kickoff of the first NFL game until one hour after the final down of the last game on NFL game days. All times were converted to UTC for consistency.

In total, 105 out of the 354 hours in the NFL data set are considered events of interest. We apply the proposed OCP method to the labeled data to determine if the flagged observations correspond to the predetermined events. To determine outlyingness using the OCP method, we use an empirically determined threshold described in Algorithm 2 in Appendix B. The data we are evaluating are prewhitened from the application of an additive Holt Winter's model. An additive model was chosen because some of the $p = 1917$ variables had some zero counts. The residuals from the Holt Winter's model show moderate departures from normality and are right-skewed. There is slight correlation among the $p = 1917$ residual vectors, the first quartile of the absolute value of the Pearson's correlation coefficients is $Q_1 = .017$ and the third quartile is $Q_3 = .073$. Although the distributions of the residuals are skewed in most cases, we evaluate outlyingness using thresholds generated from multivariate normal, lognormal, and $t-$distributions in order to compare the performance across a variety of distributional assumptions. We select the threshold values from simulated cases with an average correlation near zero ($N = 354$, $p = 1917$, $\rho = 0$) using the methodology described in Algorithm 2 in Appendix B. (See table of empirical threshold values in Appendix D). We emphasize that no exact statistical model exists for the data we are analyzing. The OCP method is intended to be an approximate, data-driven method that flags unusual events while retaining the majority of the IC data.

For comparative purposes, we include an analysis using the RMDP outlier detection method of Ro et al.[27] with a modified threshold value generated according to Algorithm 2 in the Appendix. The RMDP is a classical statistical distance method, which evaluates observations based on the distance of observations from the center of the data. The RMDP distances are based on a minimum diagonal product estimator of the covariance matrix using only an outlier free subset of data and the diagonal elements of the covariance matrix. An asymptotic cutoff value for the statistical distance is recommended by Ro et al.[27] to determine outlyingness; however, we used an empirically derived thresholds to cast the RMDP method in the best possible light. The distribution of the RMDP distances are asymptotically normal in p when the underlying data distribution is multivariate normal. We found that the threshold recommended by Ro et al.[27] resulted in Type I error probabilities that were much higher than desired in many cases, especially for non-normally distributed data as originally pointed out by Ro et al.[27]. Thus, we used an empirically derived thresholds for case $N = 354$, $p = 1917$, $\rho = 0$ that gave simulated Type I error rates of 5% under the multivariate normal, lognormal, and $t-$distributions (see table of empirical threshold values in Appendix D). Figure 2 illustrates how the asymptotically derived limits proposed by Ro et al.[27] can result in a large number of Type I errors for a 2-dimensional correlated t-distributed example with with 20% outlier contamination. Notice the larger number of "+" signs in the the plot on the

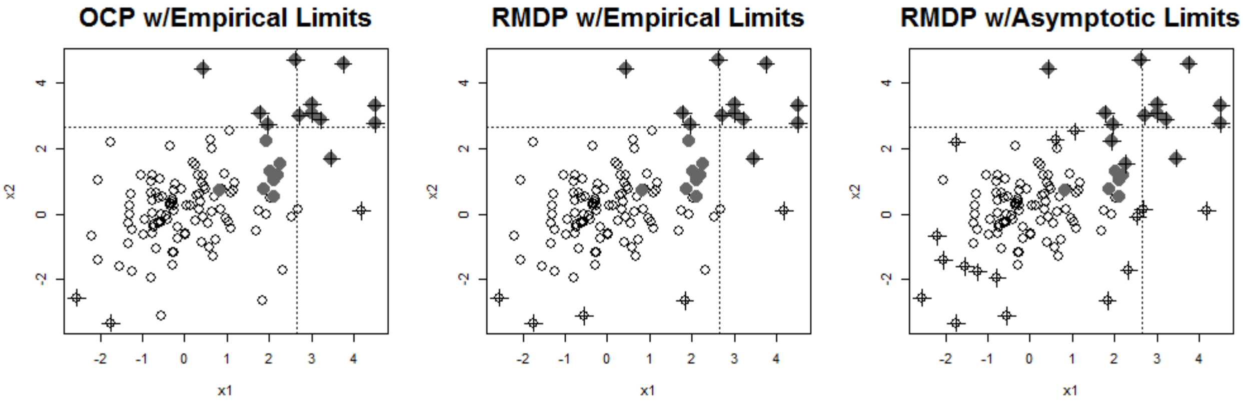

Figure 2: 2-dimensional correlated ($\rho = 0.25$) t-distribution ($N = 100$) with 20% outlier contamination. Solid circles indicate outliers, points marked with + indicate flagged points. Dashed lines indicate the mean of the outlier points.

right compared two the plots on the left.

Throughout the remainder of the paper we use empirically derived limits for the RMDP distances instead of the asymptotic limits, which we will still refer to as the RMDP method. As a result of this, the reader should keep in mind that the RMDP performance presented in this paper does not necessarily correspond to, and improves upon, the performance of the RMDP as proposed by Ro et al.[27], especially in non-normal, correlated distributions. We chose the RMDP method for comparison because it works for wide data , that is, when $p > N$, and was shown to have superior performance against several competing methods in Ro et al.[27].

Figure 3 gives a graphical display of the OCP and RMDP methods with the thresholds based on the multivariate t-distribution, which we feel most closely resembles the distribution of the residuals. When using the $\text{t}_{df=10}$ threshold values, the OCP method correctly classifies 88.70% of the observations and detects 65.71% of the events of interest. These results compare favorably against the RMDP method, which correctly classifies 84.18% of the the observations and detects only 50.48% of the events of interest. If the OCP method were used to identify outliers for this example, 281 observations would be retained for a baseline sample, of which 245 would be IC observations. By comparison, if the RMDP method were used to identify outliers in this case, 297 observations would be retained for the baseline sample with 245 as IC observations.

Table 1 gives a summary of the results of the analysis for the thresholds derived from the three different multivariate distributions. It is important to note that, in this example, the OCP has similar threshold values and performs similarly in terms of correct classification rate, regardless of the assumption of the underlying distribution. On the other hand, the performance of the RMDP method is quite sensitive to the assumption of the underlying data distribution, changing substantially between the three distributions. For example, the number of false positives for the OCP method is 8, 4, and 4 under the assumption of normal, lognormal, and $t-$distributed data. The number of false positives for the RMDP method is 145, 9, and 4 under the same respective cases. Inspection of the limits in Appendix D shows that for a given sample size, dimension, and average correlation, the OCP limits are relatively stable across distributions. For example, the limits range from $h = 2.448$ when $N = 354$, $p = 1917$, and $\rho = 0$ for normally distributed observations to $h = 4.535$ in the same case for observations from a multivariate t-distribution. For the same cases, the empirically derived RMDP limits range from $h = 1.463$ to $h = 52.500$. The stability of the of the OCP across distributions seems to hold when the distributions are symmetric or the correlations among the variables are lower ($\rho \leq .5$) but may not be true in skewed distributional cases with high correlation. Our simulation study in Section 4 will bring more clarity to the performance of the OCP method.

In terms of runtime, we measured the time in seconds to complete both the OCP and RMDP methods for the NFL example using an Intel Core i7-6700 clocked at 3.40Ghz not using parallel multicore processing. The OCP method uses 15.00 seconds of total computational time, while the RMDP method uses 93.17 seconds of total computational time. The RMDP method is generally

Table 1: Detection of events of interest in NFL Example

		FP	TP	TN	FN	Detection Rate	Correctly Classified
Normal	OCP	8	76	241	29	72.38%	89.55%
	RMDP	145	97	104	8	92.38%	56.78%
Lognormal	OCP	4	71	245	34	67.62%	89.27%
	RMDP	9	66	240	39	62.86%	86.44%
$t_{df=10}$	OCP	4	69	245	36	65.71%	88.70%
	RMDP	4	53	245	52	50.48%	84.18%

about 6 times slower than the OCP method on average for all the simulations considered in this paper. The reader should note that the runtime does not intend to measure the theoretical computational complexity of the OCP method, but to merely compare the time it takes to run each each method under the same circumstances with a fixed physical computational power.

4 Simulation Study

To study the performance of the OCP method in a variety of scenarios, an extensive simulation study was conducted.

4.1 Study Design

The simulation study compares the OCP method to the RMDP method with empirically derived thresholds. The OCP and RMDP limits were derived such that the Type I error is approximately 5% for each the distributions and cases considered. The thresholds for the OCP and the RMDP methods were determined using Algorithm 2 in Appendix B in order to achieve a Type I error rate of about 5%. The limits for both the OCP and RMDP methods are given in Appendix D.

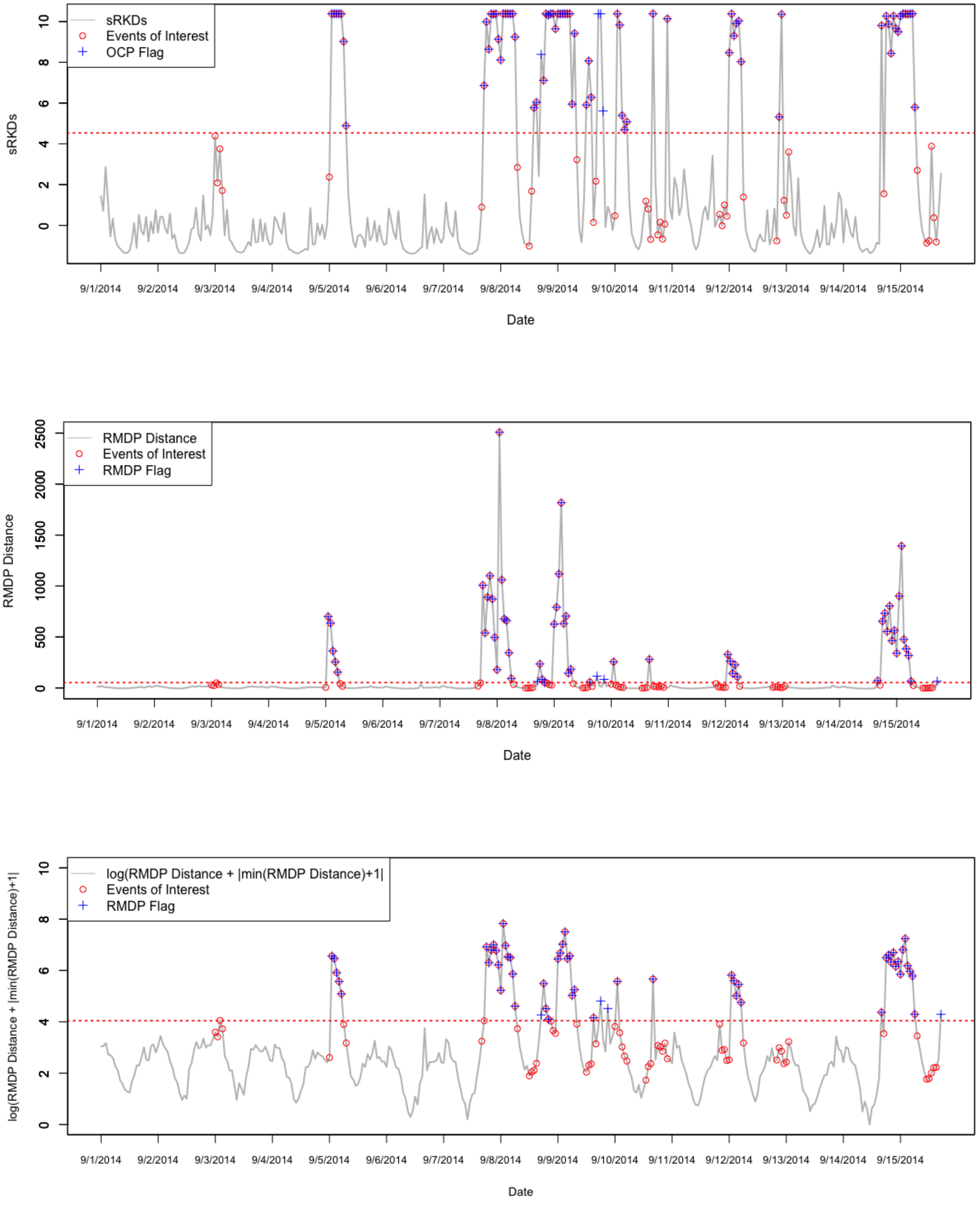

Figure 3: NFL Wikipedia hits analysis using the OCP and RMDP Methods. TOP: OCP sRKDs, MID: RMDP Distances, BOTTOM: log of shifted RMDP distances for better visualization.

The study considers observations generated from multivariate normal, lognormal, and t-distributions ($df = 10$) with sample size and dimensions of $(N = 50, p = 50), (N = 50, p = 100), (N = 100, p = 100), (N = 354, p = 1917), (N = 500, p = 2000)$. For each distribution, the correlation matrix is generated with off-diagonal elements of $\rho = 0, 0.1, 0.25, 0.75$. For each case considered, 1000 replications are summarized to obtain the performance measures.

Because of the differences in the distributions and to maintain similar shift sizes across vastly different dimensions, we shifted the OC observations probabilistically. For a given probability distribution, $F_X(\cdot)$, a p-dimensional IC sample, $S_I = \{\mathbf{x}_1, \cdots, \mathbf{x}_{N-m}\}$, of size $(N-m) \times p$, with mean, $\boldsymbol{\mu}_0$, and covariance $\boldsymbol{\Sigma}$ is generated. Then a p-dimensional OC sample, $S_O = \{\mathbf{y}_1, \cdots, \mathbf{y}_m\}$, of size $m \times p$ from $F_Y(\cdot)$, which has a shifted mean of $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \boldsymbol{\delta}$ is generated. The parameter, $\boldsymbol{\delta}$ is chosen such that the mean of the OC distribution $\boldsymbol{\mu}_0 + \boldsymbol{\delta}$ is unlikely to originate from $F_X(\cdot)$. For this study, we chose a small value for γ, such that $\gamma = .023 = 1 - \int_{-\infty}^{\mu_1+\delta_1} \ldots \int_{-\infty}^{\mu_p+\delta_p} f_X(x_1, \ldots, x_p) dx_1 \ldots dx_p$, where $f_X(\cdot)$ is the corresponding IC density function and $\boldsymbol{\mu}_0 = [\mu_1, ..., \mu_p]^\top$.The 2-dimensional plots in Figure 4 show the location of $\boldsymbol{\mu}_1$ for $\gamma = 0.023$ under the distributions considered. The location of the equicoordinate quantiles $(\boldsymbol{\mu}_0 + \boldsymbol{\delta})$ satisfying a $\gamma = 0.023$ under each distribution was obtained numerically using R package *mvtnorm* [50,51]. The multivariate shift can be compared to approximately a 2σ mean shift in a univariate normal model.

For each case, the shift is evenly distributed over p dimensions. The contaminated sample S is then defined as $S = S_I \cup S_O$ with contamination level $m/N = 0\%, 5\%, 20\%$. Table 2 gives a summary of the simulation conditions. For each possible combination of sample size, distribution, shift type, correlation, and % outliers, 1000 replications were performed.

In the IC scenarios, the primary performance measure is the average of the empirical Type I error percent. We compute

$$\hat{\alpha}_i = \frac{\mathrm{FP}_i}{N} \times 100,$$

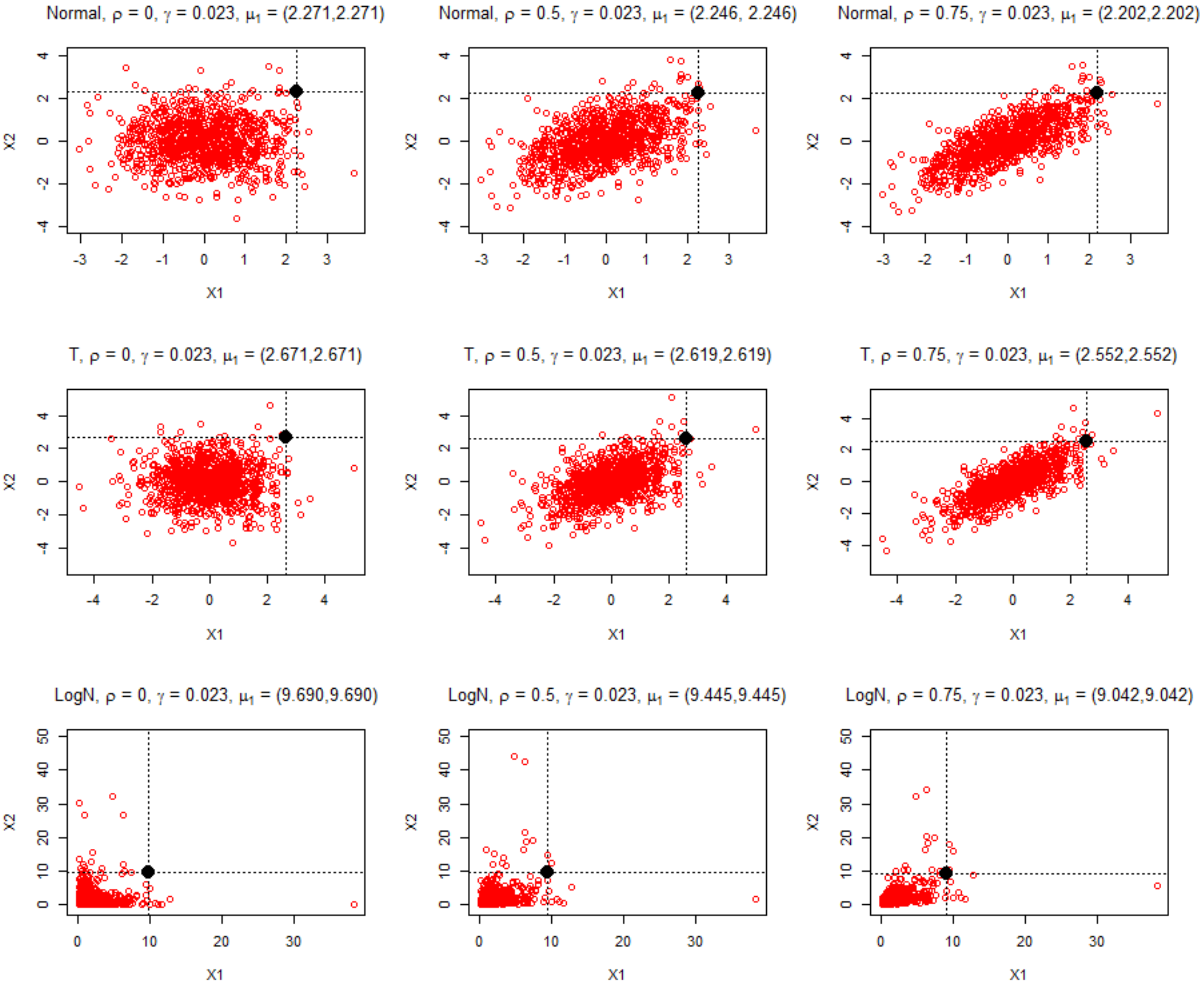

Figure 4: 2-dimensional examples for the normal, $t_{df=10}$ and lognormal distributions. Outlier mean $(\boldsymbol{\mu}_1)$ generated with $\gamma = 0.023$ indicated by solid point in crossmarks.

Table 2: Summary of OC simulation conditions.

Sample Size	Distribution	Correlation	%Outliers
$N = 50$, $p = 50$	Normal	$\rho = 0$	0
$N = 50$, $p = 100$	Lognormal	$\rho = 0.1$	5
$N = 100$, $p = 100$	$\mathrm{t}_{df=10}$	$\rho = 0.25$	20
$N = 354$, $p = 1917$		$\rho = 0.5$	
$N = 500$, $p = 2000$		$\rho = 0.75$	

where FP_i is the number of false positives in each of the $i = 1, ..., 1000$ replicates for the sample size (N) and dimension (p) cases. The $\hat{\alpha}_i$ values are averaged giving $\overline{\alpha} = \sum_{i=1}^{1000} \hat{\alpha}_i \Big/ 1000$.

In the OC scenarios, the primary performance measures are the average of the percent of the observations that are correctly classified ($\overline{\%\mathrm{CC}}$) as either IC or OC and the average of the empirical Detection Rate ($\overline{\mathrm{DR}}$) of OC observations. For each case, we compute

$$\%\mathrm{CC}_i = \frac{\mathrm{TP}_i + \mathrm{TN}_i}{N} \times 100, \text{ and } \mathrm{DR}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i} \times 100. \tag{7}$$

Here, TP_i is the number of true positives, TN_i is the number of true negatives, and FN_i is the number of false negatives for the i^{th} simulation case. These are averaged giving $\overline{\%\mathrm{CC}} = \sum_{i=1}^{1000} \%\mathrm{CC}_i \Big/ 1000$ and $\overline{\mathrm{DR}} = \sum_{i=1}^{1000} \mathrm{DR}_i \Big/ 1000$.

4.2 Simulation Results

Table 3 shows the IC performance of the OCP and RMDP methods. As previously mentioned the OCP and RMDP methods were designed to achieve a desired Type I error of 5%. Both methods achieve IC performance near the desired level when using the empirically derived threshold values. Figures 5 and 6 give a graphical summary of OC performance of the OCP and RMDP methods

Table 3: Empirical Type I error percentage for the OCP and the RMDP methods.

		$\rho = 0$		$\rho = 0.1$		$\rho = 0.25$		$\rho = 0.5$		$\rho = 0.75$	
		OCP	RMDP	OCP	RMDP	OCP	RMDP	OCP	RMDP	OCP	RMDP
Normal	N=50, p=50	5.444	4.836	5.018	5.106	4.564	5.040	4.844	4.980	5.152	5.062
	N=50, p=100	5.500	5.110	5.264	4.968	5.680	4.952	5.176	5.296	5.686	4.944
	N=100, p=100	5.501	5.625	5.455	4.358	5.410	5.175	4.609	4.927	5.274	5.050
	N=354, p=1917	5.169	5.127	4.979	5.021	4.703	5.260	4.927	4.803	4.822	4.991
	N=500, p=2000	5.143	4.763	4.831	5.317	5.061	4.888	4.623	4.663	4.798	5.476
Lognormal	N=50, p=50	4.718	5.270	5.160	5.182	4.708	4.902	4.896	4.784	5.204	4.766
	N=50, p=100	4.938	4.732	4.944	4.448	4.718	5.236	4.948	5.362	4.922	5.130
	N=100, p=100	4.612	5.424	5.193	5.151	5.088	4.995	4.934	4.020	4.874	4.816
	N=354, p=1917	4.728	4.768	4.844	4.980	4.868	5.250	4.846	5.092	5.162	4.902
	N=500, p=2000	5.151	4.912	4.662	5.132	4.805	5.145	4.796	4.874	4.907	5.208
$\text{t}_{df=10}$	N=50, p=50	4.934	5.004	4.782	4.826	5.130	4.944	5.118	4.758	5.138	4.832
	N=50, p=100	5.066	4.976	4.522	5.264	4.578	5.016	4.774	5.092	4.936	5.038
	N=100, p=100	5.083	5.241	4.984	4.632	5.047	4.959	5.142	5.018	4.841	5.067
	N=354, p=1917	5.209	4.989	5.061	5.532	4.700	4.834	5.003	5.194	5.145	4.758
	N=500, p=2000	5.114	4.454	4.964	5.294	4.665	4.784	4.979	4.926	5.132	5.100

with 5% and 20% contamination when the data follow multivariate lognormal, normal, and *t*-distributions with different levels of correlation.

Figure 5 shows the $\overline{\%\text{CC}}$ of the OCP and RMDP methods. Ideally, a method with a high value of $\overline{\%\text{CC}}$ will preserve the majority of the observations in Phase I, while correctly signaling those observations that should be flagged as potential outliers. In the case of 5% contamination, the OCP and RMDP methods performed similarly in terms of $\overline{\%\text{CC}}$ with the OCP method performing slightly better in most cases. In the case of 20% contamination, the OCP method has higher values

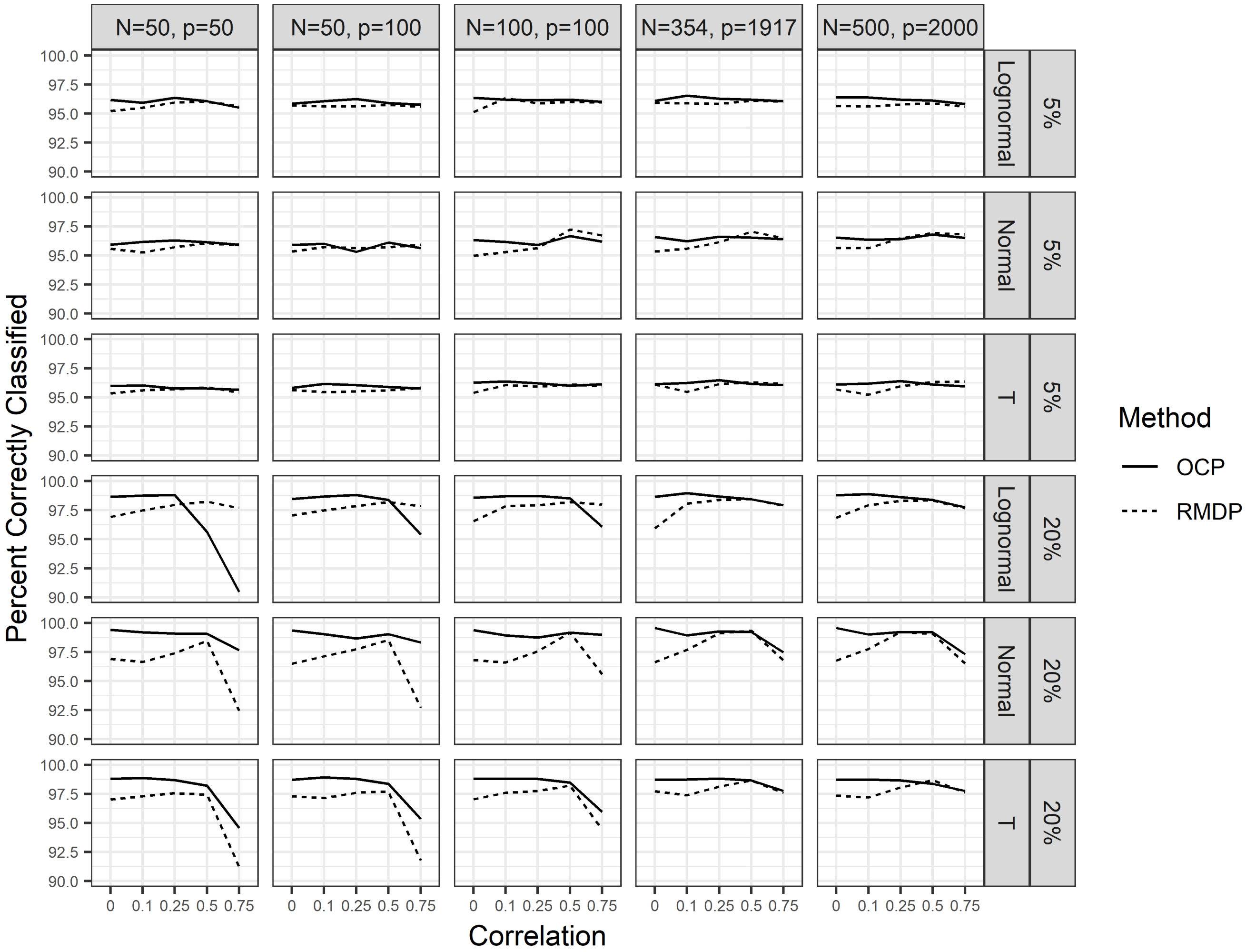

Figure 5: Summary of the Percent Correctly Classified for infected samples.

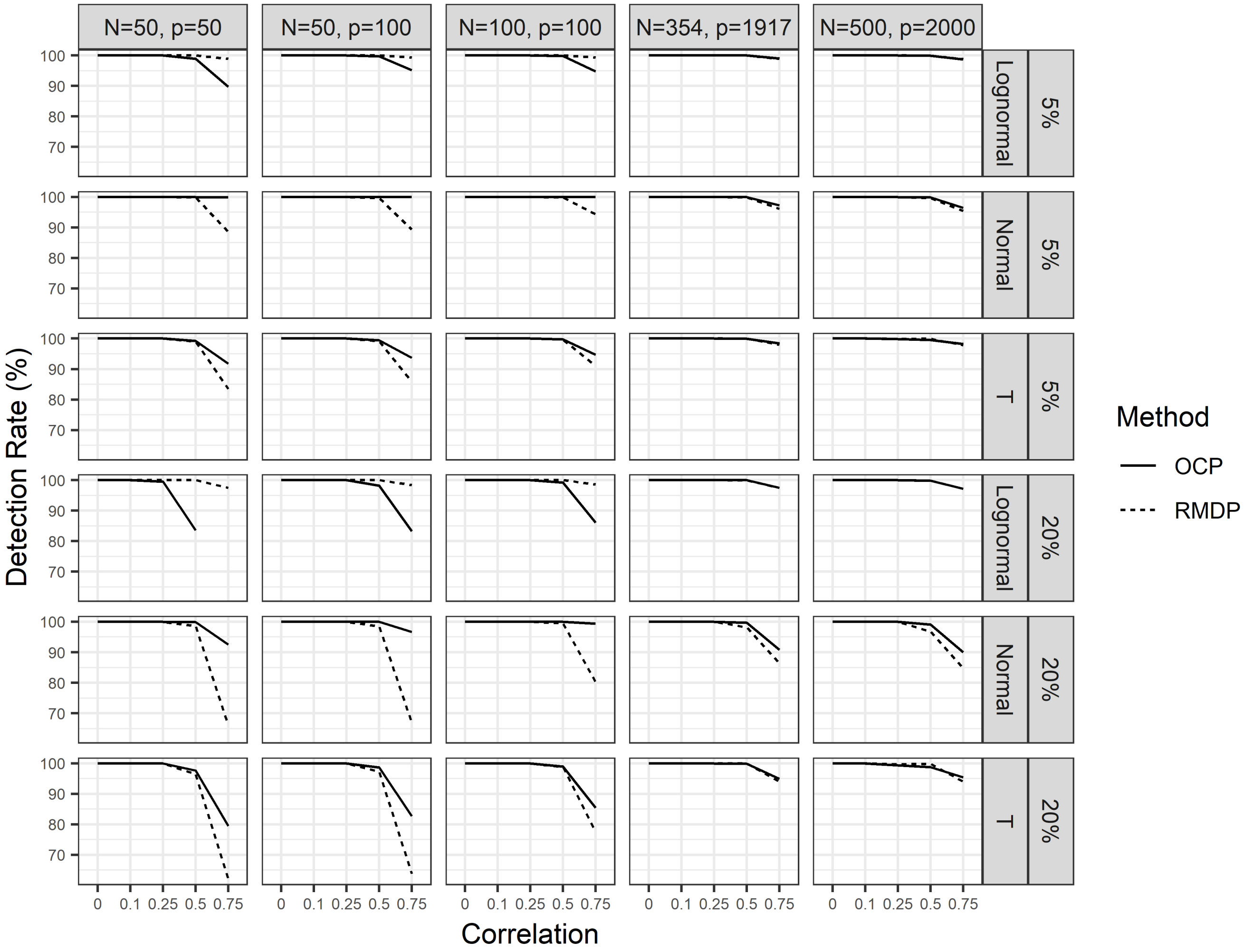

Figure 6: Summary of the Detection Rate (as a percentage) for infected samples.

of $\overline{\%\text{CC}}$ in most cases. An exception is when the data are skewed (lognormal), highly correlated, and the sample size is small. The performance of both methods generally deteriorates at higher correlation levels (i.e. $\rho = 0.75$). This is partly due to the fact that as the correlation increases, the mean shift controlling for sample size and dimensionality decreases to capture a given value of γ. See Figure 4 for visual a intuition in 2-dimensional examples. Figure 6 shows the $\overline{\text{DR}}$ of the OCP and RMDP methods. In all but the highly correlated ($\rho \geq .25$) lower dimensional ($p \leq 100$) cases when the data are lognormally distributed, the $\overline{\text{DR}}$ of the OCP method is higher than that of the RMDP.

In all of the cases considered, the OC observations were generated to be probabilistically unlikely. It is interesting to also consider how the OCP method compares to the RMDP method in detecting less extreme OC observations. To investigate this, we used the same methodology described in Section 4.1, but generated shifts with differing degrees of outlyingness. In all prior results, we used a value of $\gamma = .023$, which corresponds to the 97.7^{th} percentile, or about a 2σ shift in a univariate normal distribution. In Figure 7 we consider $0 < \gamma < 0.5$ for the case of $N = 100, p = 100$ and the multivariate lognormal, normal, and $t-$distributions. In Figure 7 we see that the $\overline{\text{DR}}$ for both the OCP and RMDP methods declines as the shift size gets smaller or γ gets larger. The $\overline{\%\text{CC}}$ behaves similarly. We also see that the OCP has equivalent or higher values of $\overline{\text{DR}}$ in all but the lognormal case with 20% contamination.

4.3 Results Summary

The proposed OCP method outperforms RMDP method in terms of correctly classifying observations from the samples and detection rate of OC observations in most cases considered. A central goal of a Phase I method is to correctly distinguish IC observations from OC observations. The correct classification of the IC and OC observations allows the practitioner to investigate the true

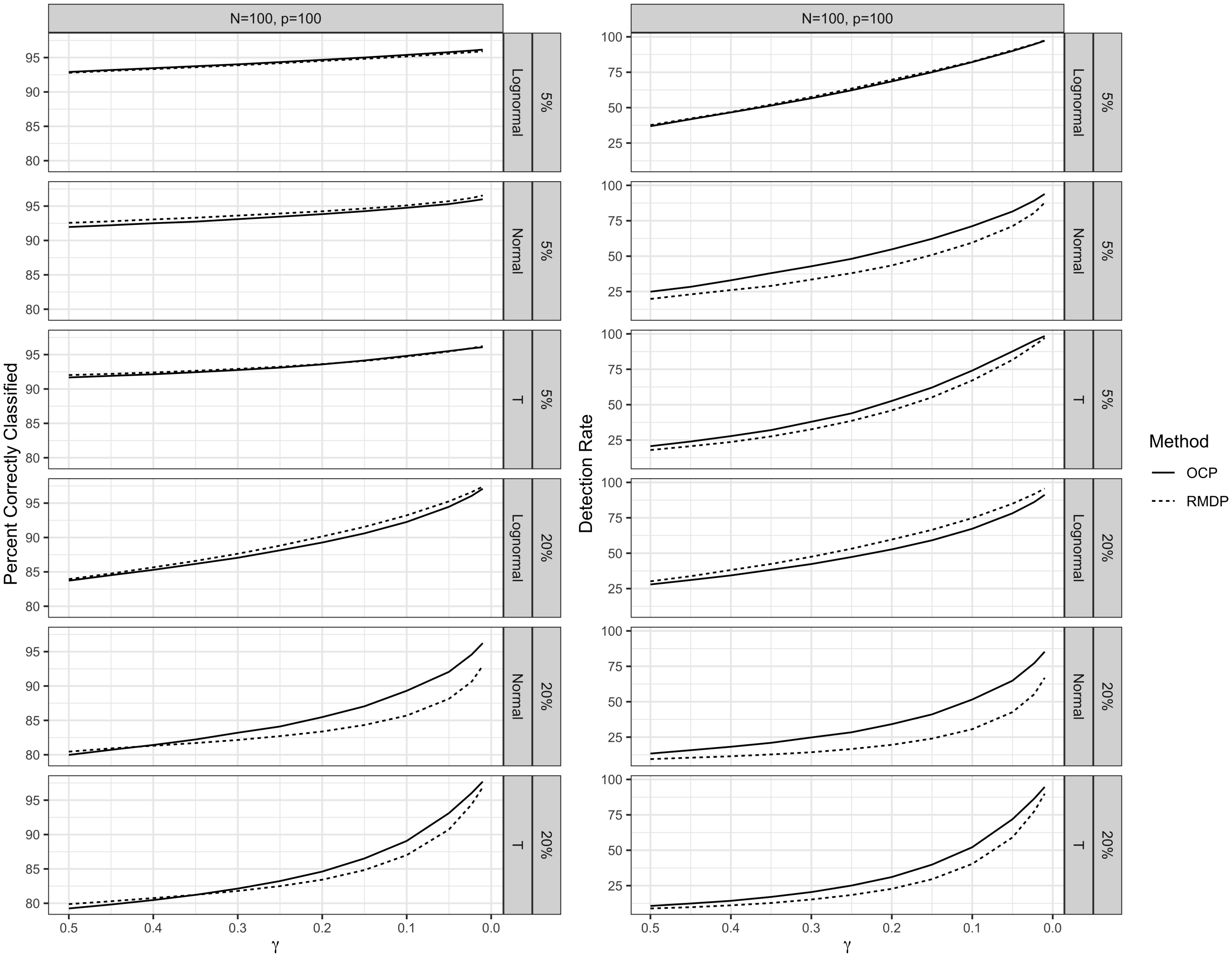

Figure 7: Summary of the of Correctly Classified and Detection Rate for varying shift sizes.

outlying observations, while retaining the largest sample possible as a baseline for use in Phase II.

The OCP method has higher $\overline{\%\text{CC}}$ and $\overline{\text{DR}}$ values than the RMDP method in most cases, especially in higher dimensions. While the RMDP method does a better job of detecting OC conditions when the data are skewed, highly correlated, and of lower dimension, as in the lognormal case with 20% contamination, the OCP method outperforms the RMDP method in all other cases considered, especially in the higher dimensional cases. It is also important to reiterate that the RMDP method has been modified here so that the method obtains better performance in non-normal, correlated settings.

5 Robust Limits

A requirement of many methods for retrospective evaluation of outlyingness is knowledge of the data distribution. It is often impossible to know the distribution of the data in practice, especially when working with data in higher dimensions. The OCP method (and the RMDP, as well as other methods) work best when the distribution is known. In the absence of distributional knowledge, practitioners can use the OCP threshold values given in Appendix D as approximate for symmetric, heavy-tailed, and skewed distributions. Alternately, we offer a simple, robust, data-driven method for determining the OCP threshold limits that works reasonably well in many situations.

We recommend using the upper fence of the traditional boxplot as an approximate robust threshold for the OCP method. We also test this approach with the RMDP method. In other words, this sample driven method calculates the threshold as follows:

$$h_{\text{BP}} = Q_3 + 1.5 \cdot IQR. \tag{8}$$

Here, Q_3 is the third quartile and IQR is the interquartile range of the distances from either the OCP or RMDP method.

Figure 8 shows the $\overline{\%\mathrm{CC}}$ and Figure 9 shows the $\overline{\mathrm{DR}}$ of the OCP and the RMDP method using the limts generated from Equation 8 for a variety of correlations and contamination levels. When the contamination rate is low (5%), both the OCP and RMDP method perform comparably in most cases, correctly classifying nearly 100% of the observations when using these robust thresholds. Both the OCP and the RMDP methods work relatively well and almost identically. However, it is important to note that in the case of very high contamination (say greater 25%), the robust thresholds are affected by the outliers, and the percent of correctly classified observations could be reduced, because the IQR would most likely be affected. Another limitation is that by using robust limits we can longer control for Type I error probability. Appendix E gives the empirical Type I error rates for the OCP and RMDP methods using robust limits in some example cases. These results are based on 1000 simulations.

6 Concluding Remarks

We have proposed a new OCP method to detect outlying observations in Phase I. Our proposed method works well in high dimensions and does not require covariance estimation. Because most traditional methods require estimation of the covariance matrix, they cannot be used with wide data. The OCP method allows a practitioner to perform a Phase I analysis with fewer samples than would be required for a traditional method.

The OCP method uses a kernel-density approach to estimate the center of the data, a distance approach to measure the distance of each observation from the center, and a simple threshold to detect OC observations. The method works well when there is a high percentage of contamination in the data set. When implemented using the Gaussian kernel, we recommend the bandwidth parameter, $s = p$, provided the data are scaled to unit variance. Optimizing the s for a given dataset is outside the scope of this paper, but could further improve the results of the OCP method,

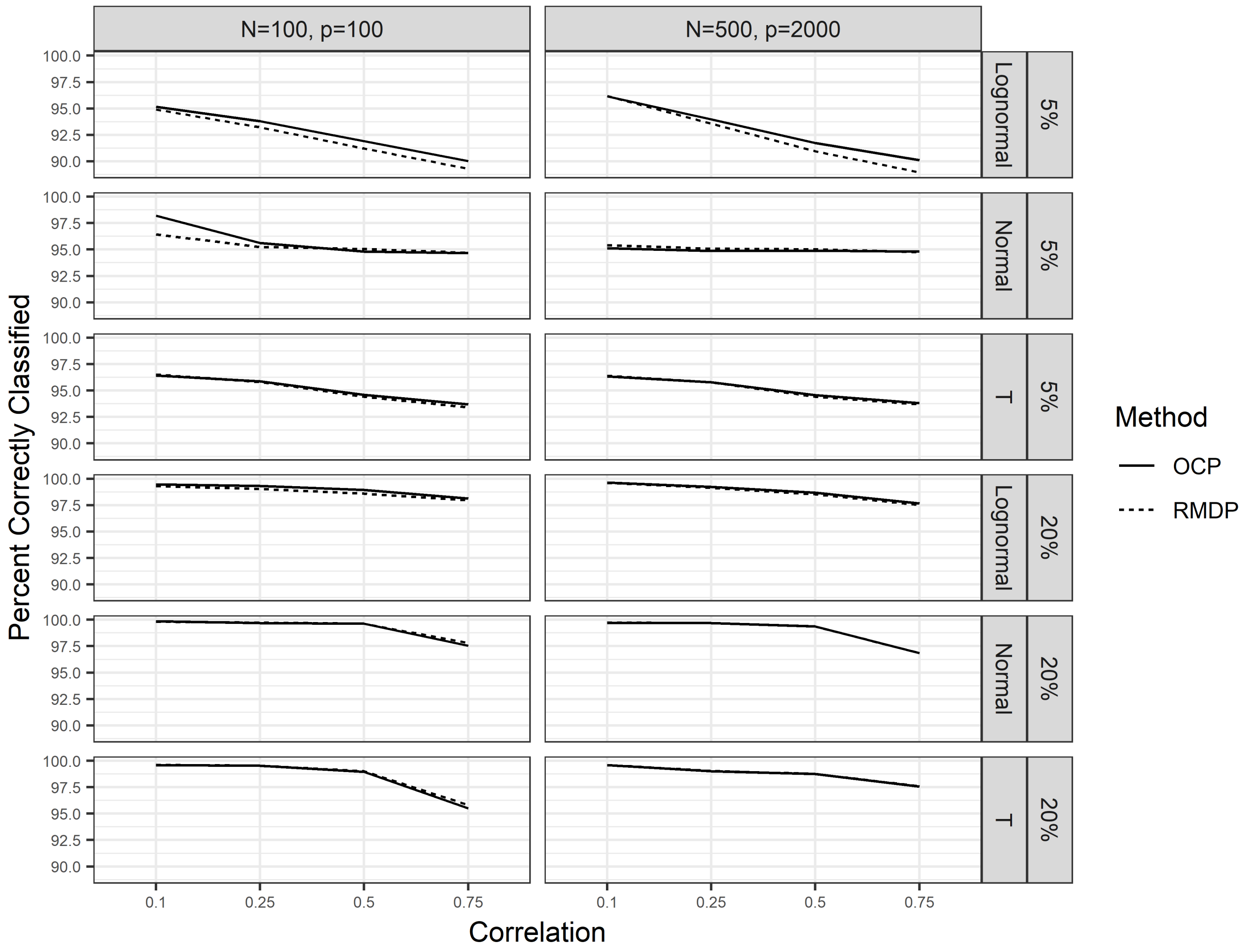

Figure 8: Summary of the Percent Correctly Classified using Robust Threshold Values.

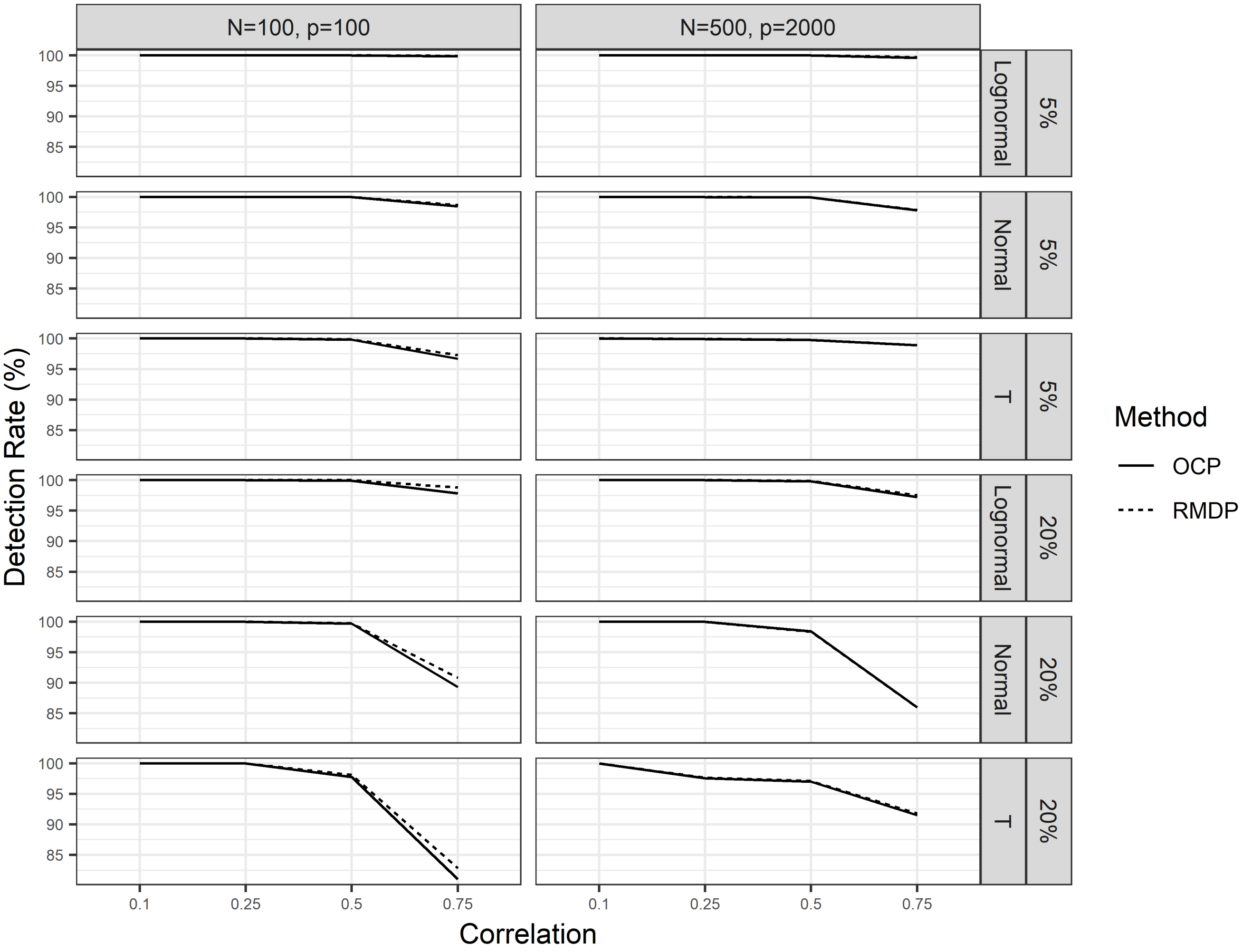

Figure 9: Summary of the Detection Rate using Robust Threshold Values.

especially for highly correlated cases. As with many multivariate methods, the SVDD boundaries can be dominated by a specific dimension if the scales of the variables are quite different. To avoid this, the practitioner might consider standardizing the data when using the OCP. In the supplemental materials we provide an R function, *OCP.R*, to compute the OCP distances and an R function, *OCPLimit.R* to find custom thresholds to determine outlyingness. In addition, we have suggested a robust method for determining a threshold for determining outlyingness when the distribution of the data is unknown.

Although there are no existing Phase I methods to serve as a benchmark for cases when $p > N$, we adapted the RMDP method of Ro et al.[27] for use in Phase I. Our results show that the OCP method results in the highest percentage of observations correctly classified as either IC or OC in most of the cases considered.

The OCP method is based on iterative peeling of boundary support vectors and requires quadratic optimization at each iteration; therefore, computational time increases with N, however empirical tests show that the OCP method has a faster computational time compared to the RMDP method. A practitioner using the OCP method should be aware that the computational time is higher for cases where $N >> p$. Because the OCP method is fairly robust to the choice of q, the desired fraction of observations rejected, when the sample size, N, is large, the practitioner can choose $q = 0.05$ to speed computation.

Natural extensions of this research include the investigation of the OCP method with kernels other than the Gaussian kernel to estimate the center of the data and determine the distance from the center. Other kernels that are more flexible and apply to a variety of data types, specifically kernels appropriate for categorical data and data of mixed types as well as kernels used for unstructured data such as images and text may prove as useful extensions of the OCP method to a new class of Phase I problems. In addition, it would be interesting to explore other methods for determining a threshold or outlyingness using the OCP method in addition to those described in our paper.

The OCP method is novel in that it departs from the traditional method of estimating the mean and covariance and measuring the distance from the center using a statistical distance. The traditional approach is restrictive to near normal models for continuous observations. Using the kernel methods as in the OCP provides a new paradigm and the opportunity for flexibility in terms of the size and types of data considered.

References

[1] L Allison Jones-Farmer, William H Woodall, Stefan H Steiner, and Charles W Champ. An overview of Phase I analysis for process improvement and monitoring. *Journal of Quality Technology*, 46(3):265–280, 2014.

[2] S Chakraborti, SW Human, and MA Graham. Phase I statistical process control charts: an overview and some results. *Quality Engineering*, 21(1):52–62, 2008.

[3] Richard C Bell, L Allison Jones-Farmer, and Nedret Billor. A distribution-free multivariate Phase I location control chart for subgrouped data from elliptical distributions. *Technometrics*, 56(4):528–538, 2014.

[4] Ching-Ren Cheng and Jyh-Jen Horng Shiau. A distribution-free multivariate control chart for Phase I applications. *Quality and Reliability Engineering International*, 31(1):97–111, 2015.

[5] Giovanna Capizzi. Recent advances in process monitoring: Nonparametric and variable-selection methods for Phase I and Phase II. *Quality Engineering*, 27(1):44–67, 2015.

[6] Demetris Trihinas, George Pallis, and Marios Dikaiakos. Low-cost adaptive monitoring techniques for the internet of things. *IEEE Transactions on Services Computing*, 2018.

[7] Bianca Maria Colosimo. Modeling and monitoring methods for spatial and image data. *Quality Engineering*, 30(1):94–111, 2018.

[8] Maria Weese, Waldyn Martinez, Fadel M Megahed, and L Allison Jones-Farmer. Statistical learning methods applied to process monitoring: An overview and perspective. *Journal of Quality Technology*, 48(1):4–24, 2016.

[9] Thuntee Sukchotrat, Seoung Bum Kim, and Fugee Tsung. One-class classification-based control charts for multivariate process monitoring. *IIE transactions*, 42(2):107–120, 2009.

[10] Peter J Huber et al. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

[11] Ricardo A Maronna and Víctor J Yohai. Robust estimation of multivariate location and scatter. *Wiley StatsRef: Statistics Reference Online*, 1976.

[12] Norm A Campbell. Robust procedures in multivariate analysis I: Robust covariance estimation. *Applied statistics*, pages 231–237, 1980.

[13] David L Donoho. Breakdown properties of multivariate location estimators. Technical report, Technical report, Harvard University, Boston. URL http://www-stat. stanford. edu/˜ donoho/Reports/Oldies/BPMLE. pdf, 1982.

[14] Peter J Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8:283–297, 1985.

[15] Ali S Hadi. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 761–771, 1992.

[16] Peter J Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

[17] Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*, volume 589. John wiley & sons, 2005.

[18] DG Simpson. Introduction to Rousseeuw (1984) least median of squares regression. In *Breakthroughs in Statistics*, pages 433–461. Springer, 1997.

[19] Nedret Billor, Ali S Hadi, and Paul F Velleman. BACON: blocked adaptive computationally efficient outlier nominators. *Computational Statistics & Data Analysis*, 34(3):279–298, 2000.

[20] Peter Filzmoser, Ricardo Maronna, and Mark Werner. Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3):1694–1711, 2008.

[21] Douglas M Hawkins. The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. *Computational Statistics & Data Analysis*, 17(2):197–210, 1994.

[22] Mia Hubert, Peter J Rousseeuw, and Stefan Van Aelst. High-breakdown robust multivariate methods. *Statistical Science*, pages 92–119, 2008.

[23] Andrea Cerioli. Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association*, 105(489):147–156, 2010.

[24] Douglas M Hawkins and David J Olive. Improved feasible solution algorithms for high breakdown estimation. *Computational Statistics & Data Analysis*, 30(1):1–11, 1999.

[25] Peter Filzmoser, Robert G Garrett, and Clemens Reimann. Multivariate outlier detection in exploration geochemistry. *Computers & geosciences*, 31(5):579–587, 2005.

[26] Virgile Fritsch, Gaël Varoquaux, Benjamin Thyreau, Jean-Baptiste Poline, and Bertrand Thirion. Detecting outlying subjects in high-dimensional neuroimaging datasets with regularized minimum covariance determinant. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011*, pages 264–271, 2011.

[27] Kwangil Ro, Changliang Zou, Zhaojun Wang, Guosheng Yin, et al. Outlier detection for high-dimensional data. *Biometrika*, 102(3):589–599, 2015.

[28] W Wu, DL Massart, and S De Jong. The kernel PCA algorithms for wide data. part I: theory and algorithms. *Chemometrics and Intelligent Laboratory Systems*, 36(2):165–172, 1997.

[29] W Wu, DL Massart, and S De Jong. Kernel-PCA algorithms for wide data part II: Fast cross-validation and application in classification of NIR data. *Chemometrics and Intelligent Laboratory Systems*, 37(2):271–280, 1997.

[30] J Tukey. Mathematics and picturing data.(rd james, ed.) 523–531. In *Canadian Math., Congress*, 1975.

[31] Regina Y Liu. Control charts for multivariate processes. *Journal of the American Statistical Association*, 90(432):1380–1387, 1995.

[32] Zachary G Stoumbos and L Allison Jones. On the properties and design of individuals control charts based on simplicial depth. *Nonlinear Studies*, 7(2), 2000.

[33] Peter J Rousseeuw and Ida Ruts. Algorithm as 307: Bivariate location depth. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 45(4):516–526, 1996.

[34] U Thissen, H Swierenga, AP De Weijer, R Wehrens, WJ Melssen, and LMC Buydens. Multivariate statistical process control using mixture modelling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 19(1):23–31, 2005.

[35] J Marcus Jobe and Michael Pokojovy. A multistep, cluster-based multivariate chart for retrospective monitoring of individuals. *Journal of Quality Technology*, 41(4):323–339, 2009.

[36] J Marcus Jobe and Michael Pokojovy. A cluster-based outlier detection scheme for multivariate data. *Journal of the American Statistical Association*, 110(512):1543–1551, 2015.

[37] David MJ Tax and Robert PW Duin. Support vector domain description. *Pattern recognition letters*, 20(11):1191–1199, 1999.

[38] Ruixiang Sun and Fugee Tsung. A kernel-distance-based multivariate control chart using support vector methods. *International Journal of Production Research*, 41(13):2975–2989, 2003.

[39] Maria L Weese, Waldyn G Martinez, and L Allison Jones-Farmer. On the selection of the bandwidth parameter for the k-chart. *Quality and Reliability Engineering International, doi: 10.1002/qre.2123*, 2017.

[40] David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54 (1):45–66, 2004.

[41] Vic Barnett. The ordering of multivariate data. *Journal of the Royal Statistical Society. Series A (General)*, pages 318–355, 1976.

[42] David L Donoho and Peter J Huber. The notion of breakdown point. *A festschrift for Erich L. Lehmann*, 157184, 1983.

[43] David L Donoho and Miriam Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, pages 1803–1827, 1992.

[44] Vladimir Naumovich Vapnik and Vlamimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[45] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[46] Xianghui Ning and Fugee Tsung. Improved design of kernel distance–based charts using support vector methods. *IIE transactions*, 45(4):464–476, 2013.

[47] Matthew P Wand and William R Schucany. Gaussian-based kernels. *Canadian Journal of Statistics*, 18(3):197–204, 1990.

[48] Peter J Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424):1273–1283, 1993.

[49] Galit Shmueli and Stephen E Fienberg. Current and potential statistical methods for monitoring multiple data streams for biosurveillance. In *Statistical Methods in Counterterrorism*, pages 109–140. Springer, 2006.

[50] Alan Genz and Frank Bretz. *Computation of multivariate normal and t probabilities*, volume 195. Springer Science & Business Media, 2009.

[51] Alan Genz, Frank Bretz, T Miwa, X Mi, F Leisch, F Scheipl, and T Hothorn. Mvtnorm: multivariate normal and t distributions. r package version 0.9-2. *URL http://CRAN. R-project. org/package= mvtnorm*, 53, 2009.

Appendices

A Breakdown Properties of the One-Class Peeled Mean

Donoho and Gasko[43] showed that convex hull peeled means can have a FSRBP of no more than $\frac{1}{p+1}$, so the FSRBP does not depend on the number of observations N, but instead depends on the dimensionality of the problem, with a deteriorating FSRBP with increasing dimension p. For the OCP estimator $\hat{\boldsymbol{\mu}}_{\text{OCP}}$ it is not possible to derive the FSRBP analytically. FSRBP properties are based on the premise that S is in general position, that is, $N \geq p$. Therefore, we perform simulations to study and estimate the breakdown properties of $\hat{\boldsymbol{\mu}}_{\text{OCP}}$ for samples from multivariate normal, lognormal and t- distributions. Obtaining estimates of the FSRBP properties through simulation is not new in the literature and has been used in the cases where analytical solutions are not available[19]. Our simulations protocol to estimate the FSRBP is described as follows: for a given probability distribution, $F_X(\cdot)$, we generate a p-dimensional inlier sample, S_I, of size $N-m$, with mean, $\boldsymbol{\mu} = \mathbf{0}$, and covariance $\mathbf{I}$ for uncorrelated distributions. For correlated distributions the correlation matrix is generated randomly such that it is positive semi-definite with correlations uniformly ranging from $[-1, 1]$. For each case considered, 500 replications are summarized to obtain the breakdown percentages. We generate the p-dimensional outlier sample, S_O, of size m from $F_Y(\cdot)$, with a shifted mean $\boldsymbol{\mu} + \boldsymbol{\delta}$, $\delta_i = 20\sigma_{ii}$, for $i = 1, \ldots, p$ for the multivariate normal and t- distributions, and $\delta_i = e^{20\sigma_{ii}}$ for the multivariate lognormal distribution. The value of δ_i was chosen such that the means of the inlier and outlier distributions are as far apart as numerically measurable. We consider breakdown when $.05 > 1 - \int_{-\infty}^{\hat{\mu}_1} \ldots \int_{-\infty}^{\hat{\mu}_p} f_X(x_1, ..., x_p) dx_1 ... dx_p$, where $\hat{\boldsymbol{\mu}}_{\text{OCP}}^{\top} = [\hat{\mu}_1, ..., \hat{\mu}_p]$ is the one-class peeling estimate of the mean of $F_X(\cdot)$. The same process is considered to estimate breakdown for the Convex Hull Peeling estimator, $\hat{\boldsymbol{\mu}}_{\text{CHP}}$.

Figure A.1 gives the percentage of cases where breakdown occurred for several dimensions and

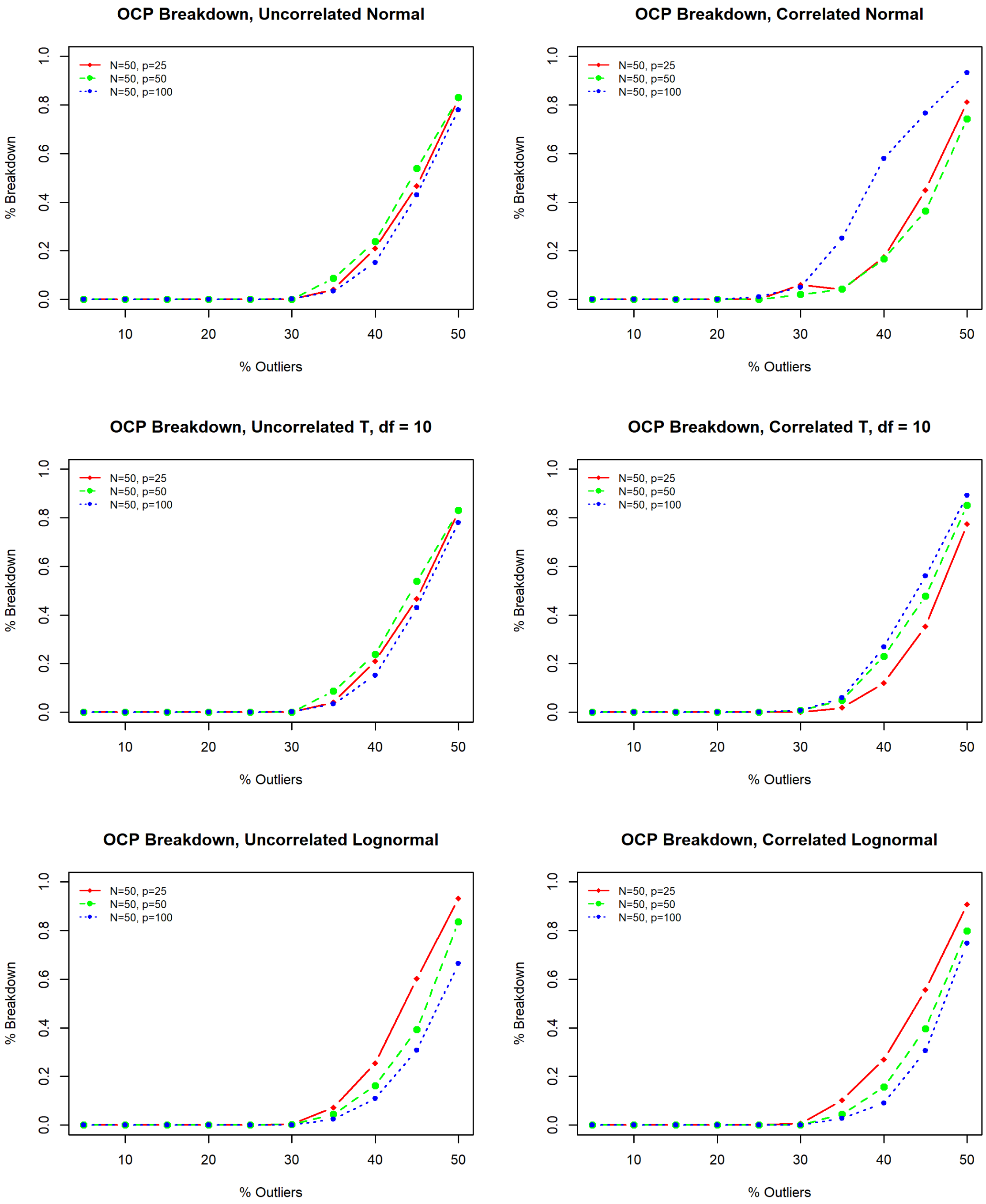

Figure A.1: Percentage of breakdown cases in 500 simulations versus percentage of outliers for $\hat{\boldsymbol{\mu}}_{\mathrm{OCP}}$ for sample size $N = 50$ and $p = 25, 50, 100$.

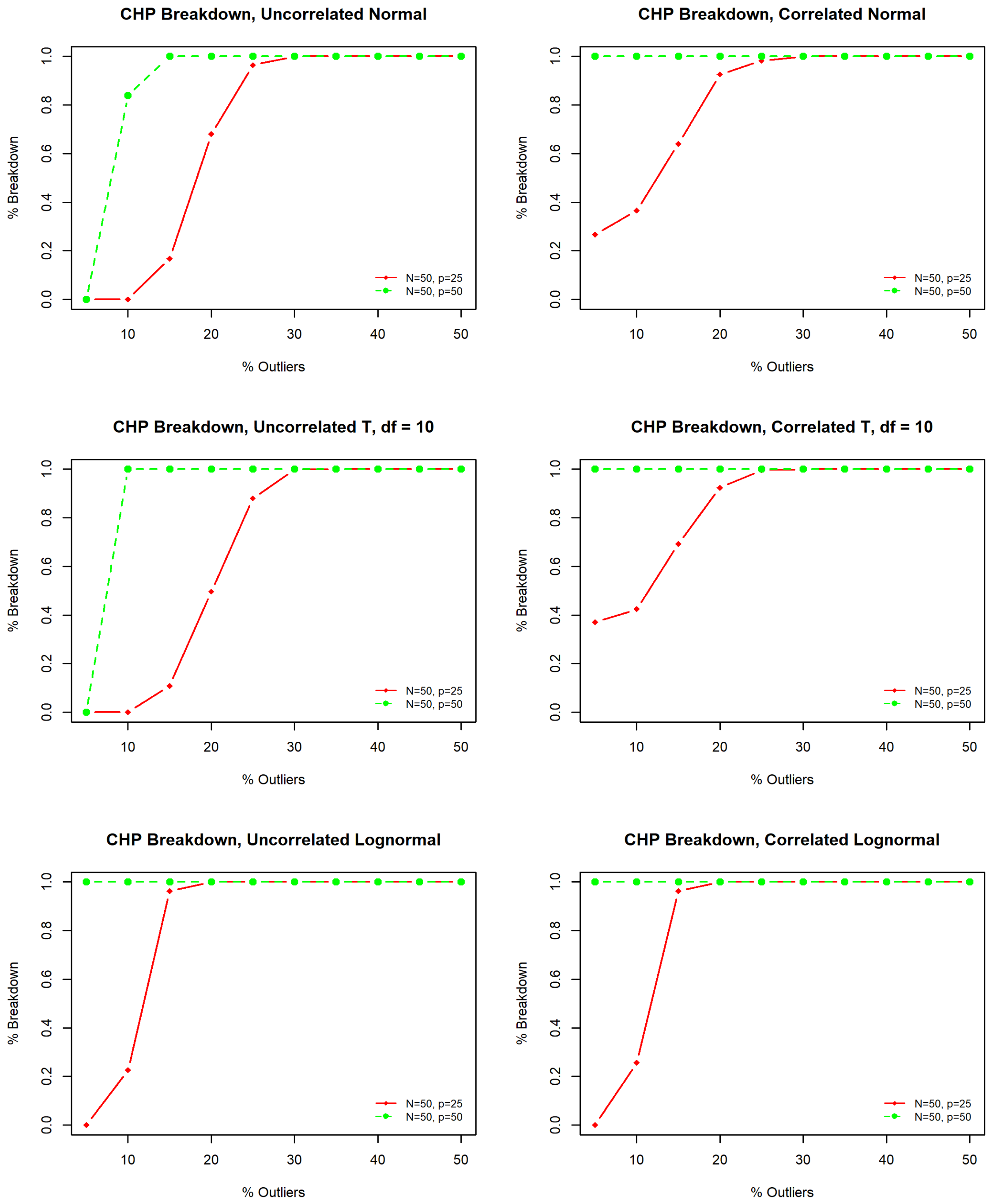

Figure A.2: Percentage of breakdown cases in 500 simulations versus percentage of outliers for $\hat{\boldsymbol{\mu}}_{\text{CHP}}$ for $N = 50$ and $p = 25, 50$.

distributions when $N = 50$ for the OCP estimator $\hat{\boldsymbol{\mu}}_{\text{OCP}}$, while Figure A.2 illustrates the breakdown performance $\hat{\boldsymbol{\mu}}_{\text{CHP}}$ for cases where $N \geq p$. The results suggest that the one-class peeling estimator, $\hat{\boldsymbol{\mu}}_{\text{OCP}}$, begins to break down at around 30-35% contamination for the cases considered, while $\hat{\boldsymbol{\mu}}_{\text{CHP}}$ breaks down more rapidly and the estimated FSRBP depends on the dimensionality p. Similar results were obtained with other sample sizes and dimensions not shown here. Figure A.3 illustrates a typical behavior of how $\hat{\boldsymbol{\mu}}_{\text{OCP}}$ breaks down if the OCP method peels to different values of n. The results suggest that the OCP method obtains the highest FSRBP values when $n = 2$. Simulations with other sample sizes and dimensions not shown here also validate this conclusion.

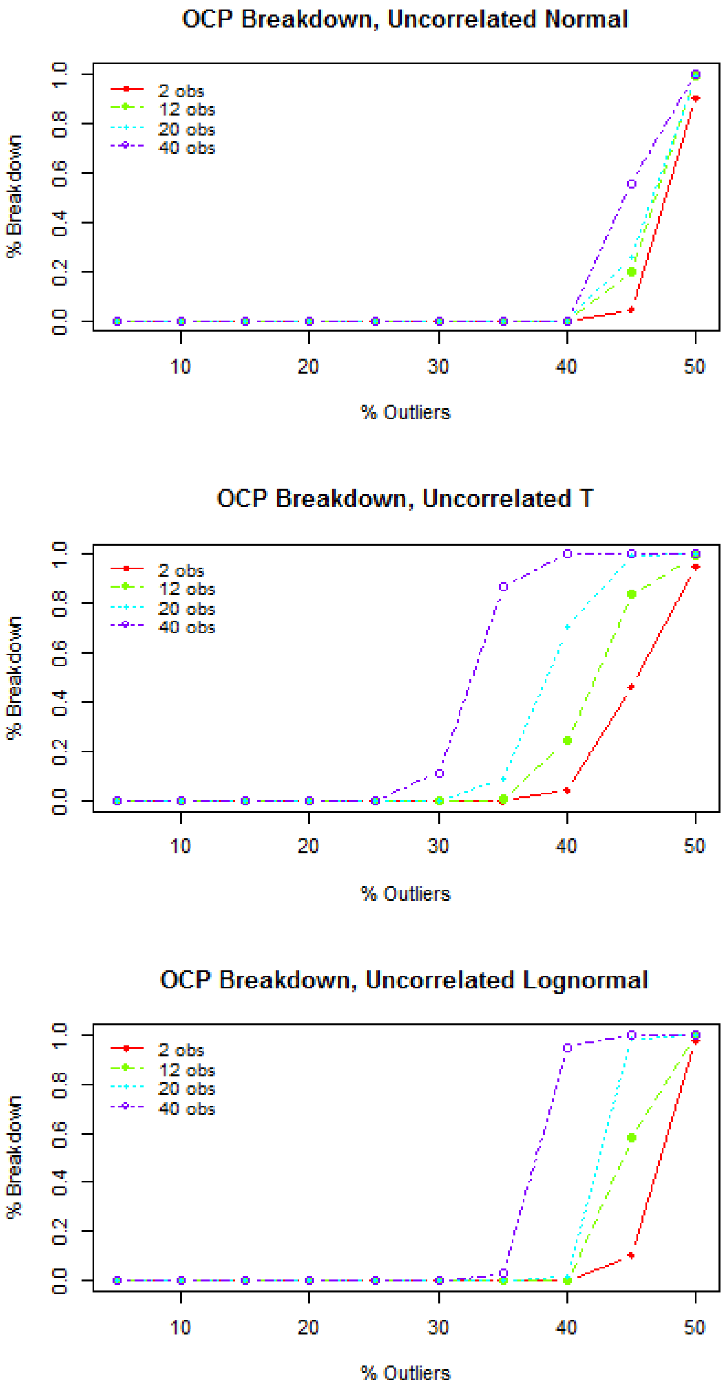

Figure A.3: Percentage of breakdown cases in 500 simulations versus percentage of outliers for $\hat{\boldsymbol{\mu}}_{\mathrm{OCP}}$ for various values of n

B Bisection Algorithm

The starting values for the bisection method are determined using an inverse prediction from a linear model for Type I error rate, $\widetilde{\alpha}$, as function of factors: N, p, h, average correlation (ρ), and distribution (Normal, Lognormal or $\mathrm{t}_{df=10}$). Empirical Type I errors are calculated from the average of 1000 simulations for each of 30 factor combinations generated using an I-optimal design. The factors in the design cover values of $50 \leq N \leq 1000$, $50 \leq p \leq 2000$, subject to $0.5 \leq p/n \leq 2$, $2.5 \leq h \leq 5.5$ and $0.1 \leq \rho \leq 0.8$ for normal, $\mathrm{t}_{df=10}$, and lognormal distributions. The starting value for h in Algorithm 2 is determined by supplying the desired Type I error rate, N, p, distribution (Normal, heavy tailed, skewed), correlation (as an estimate of the average correlation in the data) using the linear model to generate an inverse estimate along with a 99% confidence interval. The algorithm was tested using common error levels $\widetilde{\alpha} \in \{0.01, 0.05, 0.10\}$. Numerical issues could arise for $\widetilde{\alpha} < 0.01$. The code for generating a custom threshold value can be found in the supplementary materials.

Algorithm 2 Bisection Method

```
procedure BISECTION(h$_l, h_u, h, \widetilde{\alpha} = 0.05, tol = 3e^{-3}$) ▷ $h, h_u, h_u$ = estimate, lower, upper C.I.
    $S_j \leftarrow \{(\mathbf{x}_i \in \mathbb{R}^p), i = 1, ..., N\} \sim$F$(\mu, \Sigma)$    ▷ generate j = 500 samples from distribution $F$
    $\omega_{ji} \leftarrow$ METHOD$(S_j)$    ▷ The $\omega_{ji}$ are the distances. METHOD is either OCP or RMDP
    $\bar{\alpha} \leftarrow \sum_j \frac{\sum_i \mathbb{1}_{\{\omega_{ji} > h\}}/N}{500}$
    $\epsilon \leftarrow |\bar{\alpha} - \widetilde{\alpha}|$
    while $\epsilon > tol$ do
        if $\bar{\alpha} < 0.05$ then
            $h_u \leftarrow h$
            $h_l \leftarrow h_l$
            $h \leftarrow (h_u + h_l)/2$
        else
            $h_u \leftarrow h_u$
            $h_l \leftarrow h$
            $h \leftarrow (h_u + h_l)/2$
        end if
        $S_j \leftarrow \{(\mathbf{x}_i \in \mathbb{R}^p), i = 1, ..., N\} \sim$F$(\mu, \Sigma)$
        $\omega_{ji} \leftarrow$ METHOD$(S_j)$
        $\bar{\alpha} \leftarrow \sum_j \frac{\sum_i \mathbb{1}_{\{\omega_{ji} > h\}}/N}{500}$
        $\epsilon \leftarrow |\bar{\alpha} - \widetilde{\alpha}|$
    end while
    return $h$
end procedure
```

C NFL Example Event Identification Tables

Table C.1: Description of event identification for NFL example.

Step	Action
1	Searched “NFL” on Google news in a custom search for a custom date range by day (based on PDT converted to UTC) during the two week period.
2	Considered only news stories that appeared on the first page of the search results (approximately 10 search returns).
3	If a news story was new to any period consider it to be breaking news. If it was classified as breaking news then navigate to the source of the news which first reported the story to find the exact time that the story was released and converted it to UTC.
4	If the story drops during the overnight hours 1am and 8am start our period at 8am EDT the morning following (convert times to to UTC).
5	If the story drops between 6am and 1am EDT consider it a potential signal starting within a 4 hour lag of the time of the story hitting (times converted to UTC).

Table C.2: Description of NFL Example Events

Event Type	Timeline	Description of Event
Games	9/5/2014 0:00 UTC to 7:00 UTC	Thursday Night Kickoff Game
	9/7/2014 16:00 UTC to 9/8/2014 7:00 UTC	Sunday Games
	9/8/2014 22:00 UTC to 9/9/2014 8:00 UTC	Monday Games
	9/11/2014 23:00 UTC to 9/12/2014 6:00 UTC	Thursday Night Games
	9/14/2014 16:00 UTC to 9/15/2014 7:00 UTC	Sunday Games
Controversies	9/3/2014 0:00 UTC to 3:00 UTC	Wes Welker Suspension for Amphetamines
	9/8/2014 12:00 UTC to 15:00 UTC	TMZ posts video of Ray Rice Domestic Abuse
	9/8/2014 18:00 UTC to 21:00 UTC	Baltimore suspends Ray Rice
	9/10/2014 1:00 UTC to 4:00 UTC	Lesean Mccoy News on 20 cent Tip
	9/10/2014 18:00 UTC to 22:00 UTC	NFL Commisioner Under Fire for Ray Rice Decisions
	9/11/2014 20:00 UTC to 23:00 UTC	Law Enforcement contradicts NFL commisioner
	9/12/2014 20:00 UTC to 23:00 UTC	Child Abuse Case Photos (Adrian Peterson) Becomes Public
	9/15/2014 12:00 UTC to 15:00 UTC	News on Adrian Peterson Abusing Another Son
Breaking News	9/9/2014 12:00 UTC to 15:00 UTC	Ray's Rice wife issues statement of support
	9/10/2014 0:00 UTC to 3:00 UTC	Marquise Goodwin Documentary
	9/10/2014 13:00 to 16:00 UTC	News on New York Giants Performance Woes
	9/12/2014 22:00 UTC to 9/13/2014 1:00 UTC	Adrian Peterson Indicted on Child Abuse

D Table of Empirical Threshold Values

Table D.1: Empirical threshold values for the OCP and the RMDP methods.

		$\rho = 0$		$\rho = 0.1$		$\rho = 0.25$		$\rho = 0.5$		$\rho = 0.75$	
		OCP	RMDP	OCP	RMDP	OCP	RMDP	OCP	RMDP	OCP	RMDP
Normal	N=50, p=50	2.574	2.500	2.970	2.847	4.564	4.814	6.765	8.398	8.236	7.488
	N=50, p=100	2.492	2.500	3.115	3.345	4.604	6.074	7.289	9.331	8.288	6.873
	N=100, p=100	2.541	1.937	3.148	2.627	4.940	4.893	7.714	8.284	8.238	4.868
	N=354, p=1917	2.448	1.463	5.984	5.627	8.045	7.438	8.327	4.313	8.389	1.213
	N=500, p=2000	2.442	1.547	6.127	5.720	7.873	7.627	8.613	4.101	8.456	1.204
Lognormal	N=50, p=50	5.908	40.625	6.312	63.303	9.230	104.055	17.718	148.203	43.105	103.358
	N=50, p=100	5.189	40.625	5.979	69.933	9.248	110.345	18.681	150.734	48.289	93.743
	N=100, p=100	5.595	30.313	5.985	64.651	8.566	100.450	16.441	129.646	42.000	77.858
	N=354, p=1917	3.794	23.688	5.052	86.390	7.977	115.000	15.215	80.000	30.000	25.500
	N=500, p=2000	3.674	24.125	5.172	85.417	8.121	117.500	15.591	85.000	33.000	28.250
$t_{df=10}$	N=50, p=50	4.221	9.625	4.356	9.083	4.676	8.264	6.251	8.332	8.444	7.286
	N=50, p=100	4.240	12.750	4.569	10.701	5.021	9.097	6.479	8.195	8.922	6.517
	N=100, p=100	4.471	12.438	4.590	10.949	5.036	8.156	6.470	6.752	8.871	4.641
	N=354, p=1917	4.535	52.500	4.649	14.471	5.322	6.367	6.744	2.929	8.619	1.300
	N=500, p=2000	4.560	50.000	4.679	14.562	5.360	6.500	6.793	3.000	8.701	1.313

E Robust Limits IC Performance

Table E.1: Empirical Type I error percentages for the robust control limits.

		$\rho = 0.1$		$\rho = 0.25$		$\rho = 0.5$		$\rho = 0.75$	
		OCP	RMDP	OCP	RMDP	OCP	RMDP	OCP	RMDP
Normal	N=100, p=100	2.681	5.056	6.254	7.305	7.669	7.255	7.68	7.582
	N=500, p=2000	7.195	7.066	7.695	7.433	7.657	7.514	7.666	7.738
Lognormal	N=100, p=100	6.939	7.091	8.659	9.110	11.249	11.907	13.570	14.223
	N=500, p=2000	5.605	5.508	8.375	8.849	11.077	12.067	13.137	14.630
$\mathrm{t}_{df=10}$	N=100, p=100	4.999	4.868	5.814	6.072	7.519	7.775	8.520	8.789
	N=500, p=2000	5.245	5.189	5.988	6.083	7.693	7.856	8.768	8.830