A Multivariate Robust Control Chart for Individual Observations

SHOJA'EDDIN CHENOURI and STEFAN H. STEINER

University of Waterloo, Waterloo, ON N2L 3G1, Canada

ASOKAN MULAYATH VARIYATH

Memorial University of Newfoundland, St.John's, NL A1C 5S7, Canada

To monitor a multivariate process, a classical Hotelling's T^2 control chart is often used. However, it is well known that such control charts are very sensitive to the presence of outlying observations in the historical Phase I data used to set the control limit. In this paper, we propose a robust Hotelling's T^2-type control chart for individual observations based on highly robust and efficient estimators of the mean vector and covariance matrix known as reweighted minimum covariance determinant (RMCD) estimators. We illustrate how to set the control limit for the proposed control chart, study its performance using simulations, and illustrate implementation in a real-world example.

Key Words: Hotelling's T^2; Multivariate Robust Estimation of Location and Scatter; Multivariate Statistical Process Control; Outliers; Reweighted Minimum Covariance Determinant Estimators.

MONITORING a process/product over time using a control chart allows quick detection (and correction) of unusual conditions. Control charts are implemented in two phases. In Phase I, some historical process data, assumed to come from an in-control process, are used to set the control limit(s). In Phase II, the process is monitored on an ongoing basis using the control limit(s) from Phase I. In Phase II, observations falling outside the control limit(s) or unusual patterns of observations signal that the process has shifted from the in-control process settings. Such signals trigger a search for an assignable cause and, if the cause is found, lead to corrective action to prevent its recurrence.

For many products or processes, overall quality is defined simultaneously by a number of quality characteristics. To monitor a multivariate process in a way that takes into account the correlations among the variates, we may use a Hotelling's T^2 control chart (Hotelling (1947), Tracy et al. (1992)). To implement the Hotelling's T^2 control chart in Phase I with n observations, for each individual observation j we calculate

$$T^2(j) = (\mathbf{x}_j - \overline{\mathbf{x}})' \mathbf{C}^{-1} (\mathbf{x}_j - \overline{\mathbf{x}}), \qquad (1)$$

where $\mathbf{x}_j = (x_{j1}, \ldots, x_{jp})'$, $j = 1, \ldots, n$, are the n $p-$variate Phase I observations with sample mean $\overline{\mathbf{x}} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i$ and covariance matrix $\mathbf{C} = (n-1)^{-1} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$. For each individual observation, we compare $T^2(j)$ with a control limit usually derived by assuming the $\mathbf{x}_j$'s are independent multivariate normal, i.e., $\text{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Note that, in this paper, we do not consider the situation where observations are correlated over time. Under these normality and independence assumptions, the control limit for Phase I data is based on a Beta distribution and an F distribution for a Phase II observation that is independent of the Phase I data (see Tracy et al. (1992)). A large value of $T^2(j)$ indicates that the process has shifted in some way.

Dr. Chenouri is an Assistant Professor in the Department of Statistics and Actuarial Science. His email address is schenouri@uwaterloo.ca.

Dr. Variyath is an Assistant Professor in the Department of Mathematics and Statistics. His email address is variyath@mun.ca.

Dr. Steiner is an Associate Professor in the Department of Statistics and Actuarial Science. He is a Senior Member of ASQ. His email address is shsteiner@uwaterloo.ca.

To motivate and illustrate this work, consider the final assembly of automobiles. A critical characteristic of each automobile, which can influence customer perception of quality, is alignment. One measure of alignment depends on four angles, namely front-right, and front-left wheel camber and caster. In this example, large volumes of data are collected because final 100% inspection includes measurement of the four camber and caster angles. Automobiles with any of the characteristics out of specification are reworked before shipment. To monitor the alignment process, we can use a multivariate control chart, such as a Hotelling's T^2 chart. However, the process produces the occasional outlier or flyer (that are reworked before shipment). As a result, it would be useful to have a control-chart setup procedure that is robust to outliers.

As in the alignment example, the assumption that the Phase I data comes from an in-control process is not always valid. Unusual observations in Phase I can lead to inflated control limits and reduced power to detect process changes in Phase II. For this reason, part of Phase I consists of the retrospective application of the control chart with the determined control limit(s) to the Phase I process data. Any Phase I observations outside the control limit(s) are investigated. If found to be due to an identified assignable cause that can be removed, the observation is eliminated and the control limit(s) recalculated. This iterative re-estimation procedure in Phase I can eliminate the effect of a small number of very extreme observations but will fail to detect more moderate outliers. For this reason, in Phase I, we propose to use robust estimators of the mean vector and covariance matrix in order to determine an appropriate control limit for Phase II data.

As shown in Equation (1), Hotelling's T^2 uses the classical sample mean and sample covariance matrix to estimate the population mean vector and covariance matrix. However, the sample mean vector and covariance matrix estimators are very sensitive to outliers in the Phase I data. Thus, Hotelling's T^2 suffers from a masking effect, where multiple outliers in the Phase I data yield T^2 values that are not large or unusual (Rousseeuw and van Zomeren (1990)). Sullivan and Woodall (1996, 1998) showed that, in certain situations, the T^2 statistic with the sample covariance matrix estimator is not effective in detecting shifts in the process mean vector. They proposed several different estimators of the covariance matrix and concluded that an estimator based on successive differences is more effective in detecting the process shift when certain conditions apply.

Vargas (2003) introduced robust control charts for identifying outliers in Phase I multivariate individual observations based on two robust estimates of mean vector and covariance matrix, namely, the minimum covariance determinant (MCD) and the minimum volume ellipsoid (MVE). The exact distribution of T^2 with robust estimators based on MVE and MCD is not available, so control limits for Phase I data need to be obtained empirically. Vargas (2003) and Jensen et al. (2007) estimated the control limits for the robust T^2 charts for Phase I data based on simulations. Jensen et al. (2007) tabulated these estimates for sample sizes of $n = 10, \ldots 100$, dimensions $p = 2, \ldots 10$, and an overall confidence level of $1 - \alpha = 0.95$ for any out-of-control points in Phase I. The performance of these robust control charts was assessed in terms of the probability of a signal (i.e., detecting an outlier) in Phase I only.

Our approach to the problem of monitoring the multivariate observations differs in two ways from Vargas (2003) and Jensen et al. (2007). We propose using robust estimators of the mean vector and the covariance matrix based on the reweighted MCD (Rousseeuw and Van Zomeren (1990), Lopuhaä and Rousseeuw (1991), Willems et al. (2002)). Reweighted MCD estimators inherit the nice properties of initial MCD estimators, such as affine equivariance, robustness, and asymptotic normality, while achieving a higher efficiency. Reweighted MCD estimates are not unduly influenced by the outliers and thus there is no need to identify outliers in the Phase I data, which is reflected in our simulations results. Second, we propose robust control charts for Phase II data based on the reweighted MCD estimates of the mean vector and covariance matrix from Phase I. Our simulation studies show that the robust control chart based on the reweighted MCD estimates has better performance than other existing control charts under certain conditions.

Organization of the remaining part of the paper is as follows. In the next section, we discuss the existing robust estimation methods. Following that, we formally introduce a robust control chart based on the reweighted MCD. The estimation of distribution quantiles using Monte Carlo methods needed to set the control limit is then presented. Following that, we compare the performance of reweighted MCD and other previously proposed robust control charts using simulations. Then the reweighted MCD control chart

is applied to the alignment example. A discussion and final conclusions are given in the last section.

Some Background on Robust Estimation

Consider the problem of estimating the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ based on a random sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from a p-variate normal $\mathrm{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. It is desirable that the estimators are independent of the choice of coordinate system. More formally, the estimators $\mathbf{t}_n$ and $\mathbf{C}_n$ of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively, are called affine equivariant if, for any nonsingular $p \times p$ matrix $\mathbf{A}$ and vector $\mathbf{b} \in \mathbb{R}^p$,

$$\begin{aligned} \mathbf{t}_n(\mathbf{A}\mathbf{x}_1 + \mathbf{b}, \ldots, \mathbf{A}\mathbf{x}_n + \mathbf{b}) &= \mathbf{A}\mathbf{t}_n(\mathbf{x}_1, \ldots, \mathbf{x}_n) + \mathbf{b} \\ \mathbf{C}_n(\mathbf{A}\mathbf{x}_1 + \mathbf{b}, \ldots, \mathbf{A}\mathbf{x}_n + \mathbf{b}) &= \mathbf{A}\mathbf{C}_n(\mathbf{x}_1, \ldots, \mathbf{x}_n)\mathbf{A}'. \end{aligned} \tag{2}$$

The finite sample breakdown point, introduced by Donoho and Huber (1983), is a very popular global measure of robustness. Intuitively, it is the smallest amount of contamination necessary to upset an estimator entirely. Formally, let $\mathbb{X}^{(o)} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a random sample of n observations and $\mathbf{t}_n(\mathbb{X}^{(o)})$ the corresponding estimator. Imagine replacing m arbitrary points in $\mathbb{X}^{(o)}$ by arbitrary values. Let the new data be represented by $\mathbb{X}^{(m)}$. The finite sample breakdown point of the estimator $\mathbf{t}_n$ for sample $\mathbb{X}^{(o)}$ is

$$\begin{aligned} &\epsilon_n^*(\mathbf{t}_n, \mathbb{X}^{(o)}) \\ &\quad = \min\left\{\frac{m}{n}; \sup_{\mathbb{X}^{(m)}} \|\mathbf{t}_n(\mathbb{X}^{(m)}) - \mathbf{t}_n(\mathbb{X}^{(o)})\| = \infty\right\}, \end{aligned} \tag{3}$$

where $\|\cdot\|$ is the Euclidean norm.

If $\epsilon_n^*(\mathbf{t}_n, \mathbb{X}^{(o)})$ is independent of the initial sample $\mathbb{X}^{(o)}$, we say the estimator $\mathbf{t}_n$ has the universal finite sample breakdown point $\epsilon_n^*(\mathbf{t}_n)$. In this case, we can calculate the limit $\epsilon^* = \lim_{n\to\infty} \epsilon_n^*(\mathbf{t}_n)$, which is often called the breakdown point or, sometimes, the asymptotic breakdown point. A higher breakdown point implies a more robust estimator. For example, for univariate data,

- The mean $\overline{x}$ has $\epsilon_n^*(\overline{x}) = 1/n$ and hence breakdown point $\epsilon^* = 0$.
- For odd sample sizes n, the median $\widetilde{x} = x_{((n+1)/2)}$ has $\epsilon_n^*(\widetilde{x}) = (n+1)/2n$ and hence breakdown point $\epsilon^* = 1/2$.

Relaxing the affine equivariance condition of estimators to invariance under the orthogonal transformation makes it easy to find an estimator with the highest possible breakdown point 1/2. But, if we are interested in finding an affine equivariant estimator and, at the same time, a robust one, things get complicated. The combination of affine equivariance and high breakdown is rare. Davies (1987) showed that the largest attainable finite sample breakdown point of any affine equivariant estimator of the location and scatter matrix is $\lfloor (n-p+1)/2 \rfloor / n$.

Classical estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, i.e., the sample-mean vector and covariance matrix, are affine equivariant but their finite sample breakdown point is $1/n$. This means that only one outlier can corrupt the estimators. Several multivariate robust estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ have been proposed in the literature. Examples include the M-estimators (Maronna (1976)), the Stahel–Donoho estimators (Stahel (1981), Donoho (1982)), the S-estimators (Rousseeuw and Yohai (1984), Davies (1987), Lopuhaä, (1989)), the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) estimators (Rousseeuw (1985)). The M-estimators are computationally cheap, but their breakdown point, under some general conditions, cannot exceed $1/(p+1)$ (Maronna (1976), Huber (1981)). This upper bound is disappointingly low in high dimension.

The Stahel–Donoho estimators are projection based, are reasonably efficient, and have finite sample breakdown point $\lfloor (n-2p+2)/2 \rfloor / n$ (Donoho (1982)). They are the first affine equivariant estimators with the highest possible breakdown point, $\epsilon^* = 1/2$. One major trouble with Stahel–Donoho estimators is that they are computationally expensive.

The S-estimators can attain a finite sample breakdown point of $\lfloor (n-p+1)/2 \rfloor / n$, with $\epsilon^* = 1/2$, under suitable conditions (Lopuhaä and Rousseeuw (1991)). However, these estimators are also very expensive to compute.

The MVE and MCD are two affine equivariant estimators introduced by Rousseeuw (1985) that have finite sample and asymptotic breakdown points $\lfloor (n-p+1)/2 \rfloor / n$ and 1/2, respectively. The MVE location estimator has a slow, $n^{-1/3}$, rate of convergence and a nonnormal asymptotic distribution (Davies (1992)). In addition, there is no existing fast algorithm to compute the MVE estimators. The MCD location and scatter estimators have a $n^{-1/2}$ rate of convergence, and the former has an asymptotic normal distribution (Butler et al. (1993)). A fast

algorithm is available in standard software packages to compute MCD estimators in high dimensions.

In this paper, we consider a modified and more efficient version of the MCD estimators of location and scatter. To begin, we formally define the MCD estimators. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a random sample taken from an absolutely continuous distribution F in $\mathbb{R}^p$. The MCD estimators of location and scatter of the distribution are determined by the subset of size $h = \lfloor n\gamma \rfloor$ (where $0.5 \leq \gamma \leq 1$), the covariance matrix of which has the smallest possible determinant. The MCD location estimate $\overline{\mathbf{x}}_{\mathrm{MCD}}$ is defined as the average of this subset of h points, and the MCD scatter estimate is given by $\mathbf{S}_{\mathrm{MCD}} = a^n_{\gamma,p}\mathbf{C}_{\mathrm{MCD}}$, where $\mathbf{C}_{\mathrm{MCD}}$ is the covariance matrix of the subset; the constant $a^n_{\gamma,p}$ is $c_{\gamma,p} \times b^n_{\gamma,p}$, where $c_{\gamma,p}$ is a consistency factor (see Croux and Haesbroeck (1999)); and $b^n_{\gamma,p}$ is a finite sample correction factor (see Pison et al. (2002)). Here $1-\gamma$ represents the (asymptotic) breakdown point of the MCD estimators, i.e., $\epsilon^* = 1 - \gamma$.

The MCD estimator has its highest possible finite-sample breakdown point when $h = \lfloor (n + p + 1)/2 \rfloor$ (see Rousseeuw and Leroy (1987)). Computing the exact MCD estimators $(\overline{\mathbf{x}}_{\mathrm{MCD}}, \mathbf{S}_{\mathrm{MCD}})$ is very expensive or even impossible for large sample sizes in high dimensions (see Woodruff and Rocke (1994)). However, various algorithms have been suggested for approximating the MCD. A fast algorithm was proposed independently by Hawkins and Olive (1999) and Rousseeuw and Van Driessen (1999). For small datasets, the algorithm of Rousseeuw and Van Driessen (1999), known as FAST-MCD, typically finds the exact MCD, whereas, for larger datasets, it is an approximation. The FAST-MCD is implemented in standard statistical softwares such as SPLUS, R, SAS, and Matlab.

For the multivariate normal distribution

$$\mathrm{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

the MCD estimator $(\overline{\mathbf{x}}_{\mathrm{MCD}}, \mathbf{S}_{\mathrm{MCD}})$ with

$$c_{\gamma,p} = \gamma / P(\chi^2_{(p+2)} \leq q_\gamma)$$

is consistent for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\chi^2_{(r)}$ represents a chi-square random variable with r degrees of freedom and q_γ is the γth quantile of $\chi^2_{(r)}$.

In addition to being highly robust against outliers, if robust multivariate estimators are going to be of use in statistical inference, they should offer reasonable efficiency under the multivariate normal distribution. There is usually a tradeoff between efficiency and robustness, but if one is interested in having both efficiency and robustness, the best proposal seems to be two-stage or reweighted estimators (Rousseeuw and van Zomeren (1990), Woodruff and Rocke (1994)). In this paper, we propose using reweighted MCD estimators as commonly defined in the literature (Willems et al. (2002)). This is because the reweighted MCDs are affine equivariant estimators with a high breakdown point, an $n^{-1/2}$ rate of convergence, high efficiency, and there exists a fast and good approximate algorithm for computational purposes. The reweighted MCD estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the weighted-mean vector,

$$\overline{\mathbf{x}}_{\mathrm{RMCD}} = \left(\sum_{i=1}^{n} w_i \mathbf{x}_i \right) \Big/ \left(\sum_{i=1}^{n} w_i \right), \tag{4}$$

and covariance matrix,

$$\mathbf{S}_{\mathrm{RMCD}} = c_{\eta,p} d^{n,p}_{\gamma,\eta} \frac{\sum_{i=1}^{n} w_i (\mathbf{x}_i - \overline{\mathbf{x}}_{\mathrm{RMCD}})(\mathbf{x}_i - \overline{\mathbf{x}}_{\mathrm{RMCD}})'}{\sum_{i=1}^{n} w_i}, \tag{5}$$

where the weights are based on the robust distances

$$D(\mathbf{x}_i) = \sqrt{(\mathbf{x}_i - \overline{\mathbf{x}}_{\mathrm{MCD}})' \mathbf{S}^{-1}_{\mathrm{MCD}} (\mathbf{x}_i - \overline{\mathbf{x}}_{\mathrm{MCD}})}. \tag{6}$$

Observations with $D(\mathbf{x}_i)$ below the cutoff value q_η, where q_η is the ηth quantile of the chi-square distribution with p degrees of freedom, are assigned weight 1, while all other observations are given weight 0, i.e.,

$$w_i = \begin{cases} 1 & \text{if } D(\mathbf{x}_i) \leq q_\eta \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

We use the value $\eta = 0.975$, which was advocated and used by Rousseeuw and Van Driessen (1999). Using $c_{\eta,p} = \eta / P(\chi^2_{(p+2)} \leq q_\eta)$ makes $\mathbf{S}_{\mathrm{RMCD}}$ consistent under the multivariate normal distribution. The factor $d^{n,p}_{\gamma,\eta}$ is a finite sample correction given by Pison et al. (2002).

The reweighted MCD estimators preserve the finite sample breakdown point of the MCD estimators (Lopuhaä and Rousseeuw (1991)). Because the MCD estimators are affine equivariant and the robust distance $D(\mathbf{x}_i)$ is invariant under an affine transformation of $\mathbf{x}_i$, the reweighted MCD estimators are affine equivariant. In addition, the reweighted MCD estimators have bounded influence functions and are asymptotically normal, just like the MCDs. In summary, the reweighted MCD estimators inherit the nice properties of the initial MCD estimators, such as

affine equivariance, robustness, and asymptotic normality while achieving a higher efficiency. The choice of $\gamma = 0.5$ yields the maximum asymptotic breakdown point for the MCD and reweighted MCD estimators, i.e., $\epsilon^* = 1 - \gamma = 0.5$.

A Robust Control Chart

Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a p-variate random sample of size n from $\text{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ that is considered the Phase I data in what follows. It is well known (see Wilks (1963), p. 263) that, for a Phase II observation $\mathbf{x}_f \notin \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, we have

$$T^2(f) \sim \left[\frac{p(n+1)(n-1)}{n(n-p)}\right] F(p, n-p), \quad (8)$$

where $T^2(f)$ is as defined in Equation (1) and $F(r_1, r_2)$ is F distribution with r_1 and r_2 degrees of freedom. To robustify the T^2 control chart based on Phase I data, we propose to replace $\overline{\mathbf{x}}$ and $\mathbf{S}$, the classical estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, by the reweighted MCD estimators. Suppose $\overline{\mathbf{x}}_{\text{RMCD}}$ and $\mathbf{S}_{\text{RMCD}}$ represent the reweighted MCD mean vector and covariance matrix estimators, respectively. We define a robust Hotelling's T^2 for $\mathbf{x}_f$ based on these RMCD estimates by

$$T^2_{\text{RMCD}}(f) = (\mathbf{x}_f - \overline{\mathbf{x}}_{\text{RMCD}})' \mathbf{S}^{-1}_{\text{RMCD}} (\mathbf{x}_f - \overline{\mathbf{x}}_{\text{RMCD}}), \quad (9)$$

where $f = n+1, n+2, \ldots$. The finite sample distributions of the MCD and reweighted MCD estimators and thus $T^2_{\text{MCD}}(f)$ and $T^2_{\text{RMCD}}(f)$ are unknown. Asymptotic properties of these estimators have been investigated in Butler et al. (1993), Croux, and Haesbroeck (1999), and Lopuhaä (1999). To find the asymptotic distribution of $T^2_{\text{RMCD}}(f)$, we first note that $\overline{\mathbf{x}}_{\text{RMCD}}$ and $\mathbf{S}_{\text{RMCD}}$ are consistent estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. Furthermore, $\mathbf{x}_f \sim \text{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, thus applying the Slutsky theorem (see Serfling (1980)), as $n \to \infty$

$$\begin{aligned}(\mathbf{x}_f - \overline{\mathbf{x}}_{\text{RMCD}})' \mathbf{S}^{-1}_{\text{RMCD}} (\mathbf{x}_f - \overline{\mathbf{x}}_{\text{RMCD}}) \\ \xrightarrow{\mathcal{D}} (\mathbf{x}_f - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_f - \boldsymbol{\mu}) \sim \chi^2_{(p)}, \end{aligned} \quad (10)$$

i.e., $T^2_{\text{RMCD}}(f)$ has an asymptotic $\chi^2_{(p)}$ distribution. However, this asymptotic distribution only works for large sample sizes. In the next section, we apply Monte Carlo simulations to estimate quantiles of the distribution of $T^2_{\text{RMCD}}(f)$ for several combinations of sample size and dimension. For each dimension, we further fit a smooth curve between the sample size and quantiles of $T^2_{\text{RMCD}}(f)$. These fits can be used to estimate appropriate quantiles of $T^2_{\text{RMCD}}(f)$ for small Phase I sample sizes ($n \leq 200$).

Construction Procedures

Estimation of Control Limits

In order to estimate the 99%, and 99.9% quantiles of $T^2_{\text{RMCD}}(f)$ for a given Phase I sample size n, dimension p, and breakdown point $1 - \gamma$, we generate $K = 10{,}000$ samples of size n from a standard multivariate normal distribution $\text{MVN}_p(\mathbf{0}, \mathbf{I}_p)$. For each data set of size n, we compute the reweighted MCD mean vector and covariance matrix estimates, $\overline{\mathbf{x}}_{\text{RMCD}}(k)$, and $\mathbf{S}_{\text{RMCD}}(k)$, $k = 1, \ldots K$. In addition, for each data set, we randomly generate a new observation $\mathbf{x}_{f,k}$ from $\text{MVN}_p(\mathbf{0}, \mathbf{I}_p)$ (treated as a Phase II observation) and calculate the corresponding $T^2_{\text{RMCD}}(k, f)$ value as given by Equation (9). The empirical distribution function of $T^2_{\text{RMCD}}(f)$ is based on the simulated values

$$T^2_{\text{RMCD}}(1, f), T^2_{\text{RMCD}}(2, f), \ldots, T^2_{\text{RMCD}}(K, f). \quad (11)$$

By inverting the empirical distribution function of $T^2_{\text{RMCD}}(f)$, we obtain Monte Carlo estimates of the 99%, and 99.9% quantiles. We construct the empirical distribution of $T^2_{\text{RMCD}}(f)$ for any combination of $p = 2, \ldots, 10$ and $n = 20, 21, \ldots, 50, 55, 60, \ldots, 200$.

Figure 1 shows scatter plots of the empirical 99%, and 99.9% quantiles of $T^2_{\text{RMCD}}(f)$ versus the sample size n for dimensions $p = 2, 6$, and 10. These scatter plots for different dimensions suggest that we could model the quantiles using a family of regression curves of the form $f(n) = b_1 + b_2/n^{b_3}$. Because the asymptotic distribution of $T^2(f)$ is $\chi^2_{(p)}$, it is sensible to use the following two parameter family of curves instead:

$$f_{p,1-\alpha,\gamma}(n) = \chi^2_{(p,1-\alpha)} + \frac{a_{1,p,1-\alpha,\gamma}}{n^{a_{2,p,1-\alpha,\gamma}}}, \quad (12)$$

where $\chi^2_{(p,1-\alpha)}$ is the $1 - \alpha$ quantile of the χ^2 distribution with p degrees of freedom and $a_{1,p,1-\alpha,\gamma}$ and $a_{2,p,1-\alpha,\gamma}$ are constants. Fitting this curve to the data will help us predict the desired quantiles of $T^2_{\text{RMCD}}(f)$ for any Phase I sample size n. Note that, as n increases, $f_{p,1-\alpha,\gamma}(n)$ approaches $\chi^2_{(p,1-\alpha)}$. Table 1 gives the least-square estimates of the parameters $a_{1,p,1-\alpha,\gamma}$ and $a_{2,p,1-\alpha,\gamma}$, for dimensions $p = 2, 3, \ldots 10$ and the 99%, 99.9% quantiles. Using Table 1 and Equation (12), we can compute the 99% and 99.9% quantiles of $T^2_{\text{RMCD}}(f)$ for $p = 2, \ldots, 10$ and any Phase I sample size n. The regression curves given by Equation (12) fit well to all the cases in Table 1, yielding R^2 values of at least 88%. For dimensions $p \geq 11$, a similar pattern is expected, although for a given situation, a practitioner may sim-

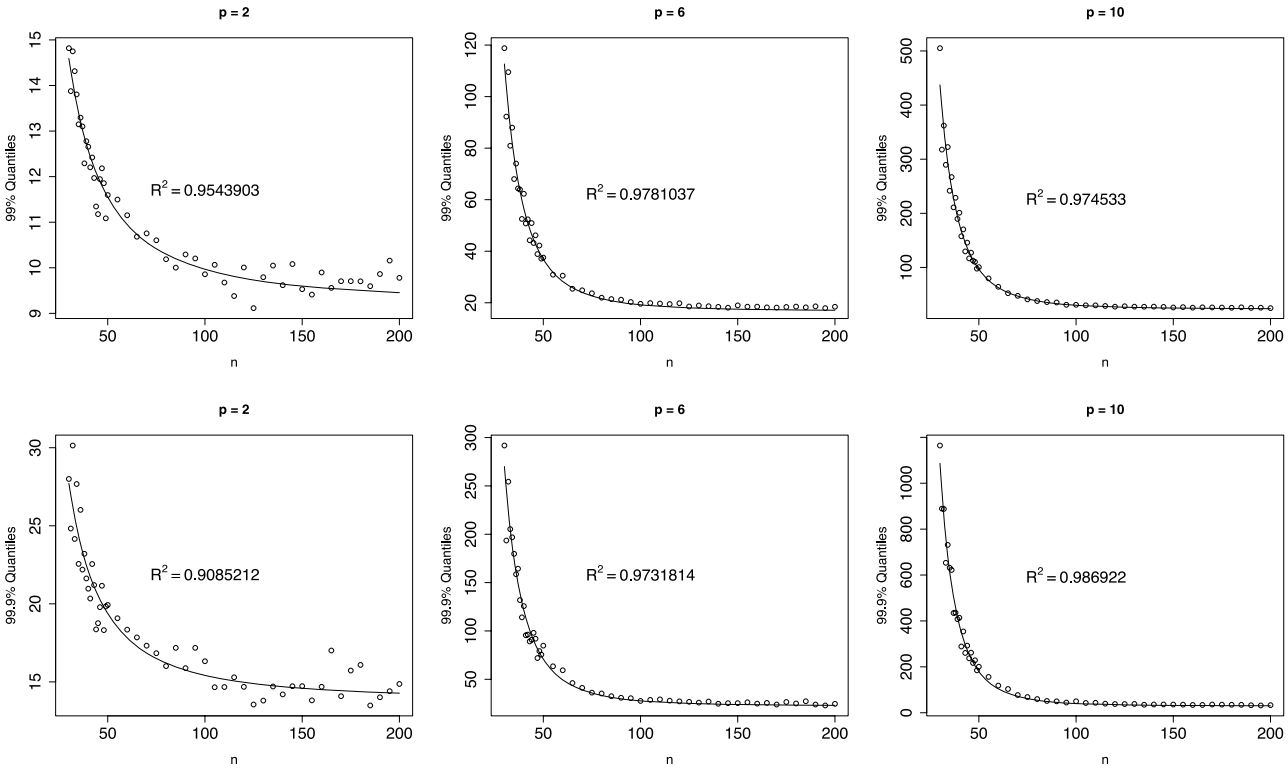

FIGURE 1. Simulated Quantiles of $T^2_{\text{RMCD}}(f)$ and the Fitted Curves for $p = 2, 6, 10$, $\gamma = 0.5$ and $\alpha = 0.01$ (Upper Panel) and $\alpha = 0.001$ (Lower Panel).

ulate the control limit directly. Our simulations used the function CovMcd() in the rrcov package written by Valentin Todorov (2007) in R software, which is available from the R project website http://r-projects.org/. It is worthwhile to note that CovMcd() also provides the initial MCD estimators using the function slot() with the arguments raw.center and raw.cov.

TABLE 1. The Least-Squares Estimates of the Regression Parameters $a_{1,p,1-\alpha,\gamma}$, $a_{2,p,1-\alpha,\gamma}$ for Dimensions $p = 2, \ldots, 10$, Confidence Levels $1 - \alpha = 0.99, 0.999$, and Breakdown Points $1 - \gamma = 0.5, 0.25$

	$\gamma = 0.5$				$\gamma = 0.75$			
	99% quantile		99.9% quantile		99% quantile		99.9% quantile	
p	$\widehat{a}_{1,p,0.99,0.5}$	$\widehat{a}_{2,p,0.99,0.5}$	$\widehat{a}_{1,p,0.999,0.5}$	$\widehat{a}_{2,p,0.999,0.5}$	$\widehat{a}_{1,p,0.99,0.75}$	$\widehat{a}_{2,p,0.99,0.75}$	$\widehat{a}_{1,p,0.999,0.75}$	$\widehat{a}_{2,p,0.999,0.75}$
2	1387.415	1.632	6225.543	1.795	208.836	1.251	1476.590	1.568
3	13533.973	2.018	71901.268	2.204	830.500	1.474	3530.978	1.647
4	110115.9	2.420	1897062	2.917	1709.908	1.563	23453.370	2.050
5	401744.3	2.618	2261387	2.838	7625.221	1.868	22914.710	1.950
6	3168654	3.060	12987610	3.195	13075.115	1.925	55097.744	2.103
7	2733044	2.904	10857430	3.019	43535.449	2.166	219090.500	2.407
8	5828231	3.009	12730200	2.976	64711.622	2.197	145095.600	2.223
9	9063979	3.048	27445690	3.114	80949.116	2.184	195972.600	2.231
10	41396480	3.385	471116200	3.824	91663.370	2.154	227923.500	2.209

Implementation Procedures

A step-by-step approach for constructing a T^2_{RMCD} control chart is given as follows:

Phase I

1. Decide on the sample size n, number of variables p, and confidence level $1-\alpha$.
2. Collect the Phase I data $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ at well-defined periodic intervals.
3. Using the Phase I data, compute the reweighted MCD estimates $\overline{\mathbf{x}}_{\text{RMCD}}$ and $\mathbf{S}_{\text{RMCD}}$ with breakdown point $1-\gamma = 0.5$ or 0.25.
4. For the desired α and p values, choose the least-square estimates $\widehat{a}_{1,p,1-\alpha,\gamma}$ and $\widehat{a}_{2,p,1-\alpha,\gamma}$ from Table 1, and then compute the control limit using Equation (12).

Phase II

5. Compute T^2_{RMCD} for each of the new observation as per Equation (9) and plot it on a control chart with the limit derived in Phase I (step 4).
6. Interpret the chart and look for out-of-control points or patterns. Diagnose the process if needed.

Performance of Robust T^2 Control Charts

In order to assess the performance of T^2_{RMCD} control charts, we conduct a number of simulation studies that consider different Phase I data structures and the amount of shift in the process mean vector in the Phase II data. The performance of the control chart is judged based on the probability of detecting changes in the process behavior based on the Phase II data. A shift in the process mean vector is measured by the noncentrality parameter (ncp) δ^2 as

$$\delta^2 = (\boldsymbol{\mu} - \boldsymbol{\mu}_A)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_A), \qquad (13)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_A$ represent in-control and out-of-control mean vectors, respectively. In this paper, we assume that there are no changes in covariance structure. Without loss of generality, because of affine equivariance, we generate in-control (no outlier) Phase I data from the standard multivariate normal distribution $\text{MVN}_p(\mathbf{0}, \mathbf{I}_p)$. In Phase I, $100\pi\%$ of the data are generated from $\text{MVN}_p(\boldsymbol{\mu}_I, \mathbf{I}_p)$ and $100(1-\pi)\%$ from $\text{MVN}_p(\mathbf{0}, \mathbf{I}_p)$, where $\delta_I^2 = \|\boldsymbol{\mu}_I\|^2$ and $\pi = 0, 0.10, 0.20$. Phase II data are generated from $\text{MVN}_p(\boldsymbol{\mu}_{II}, \mathbf{I}_p)$ where $\delta_{II}^2 = \|\boldsymbol{\mu}_{II}\|^2$. We considered the following different cases in our performance studies. We consider each combination of the above Phase I and II scenarios for Phase I sample sizes $n = 50, 150$, dimensions $p = 2, 6, 10$, and breakdown points $1-\gamma = 0.5, 0.25$, with the control limit set for a level of confidence of $1-\alpha = 0.99$. We present only the result for $\alpha = 0.01$ here and note that similar conclusions hold for other values of α. The performance of the control charts is judged by the probability of signal that is estimated as the proportion of $T^2_{\text{RMCD}}(f)$ values that fall above the control limit based on 10,000 simulations. In each simulation, we generate a sample of size n and compute the reweighted MCD estimates. Using these estimates, we compute the $T^2_{\text{RMCD}}(f)$ from Equation (9) for each observation in the Phase II data. The computed $T^2_{\text{RMCD}}(f)$ values were then compared with the approximate control limit to estimate the probability of signal. This is done for breakdown points of 50% and 25%, and the respective probabilities of signal are denoted by Re-MCD50 and Re-MCD75 in Figures 2–6.

For comparison purposes, we also estimate the probability of signal, on the same data sets, for other methods, such as the standard T^2 chart, the robust T^2 chart based on raw MCD and MVE estimators discussed in Vargas (2003) and Jensen et al. (2007), to identify outliers in Phase I data. We extend the raw MCD and MVE approaches to Phase II by using the robust estimators on the Phase I data to eliminate outliers. Then we construct the standard T^2

TABLE 2. Different Data Cases in the Performance Study

Case	Phase I	Phase II
1	No outliers ($\pi = 0$)	Process shifted with $\delta_{II}^2 = 0, 5, 10, 15, 20, 25, 30$
2	10% ($\pi = 0.10$) of the data from $\delta_I^2 = 5$	Process shifted with $\delta_{II}^2 = 0, 5, 10, 15, 20, 25, 30$
3	10% ($\pi = 0.10$) of the data from $\delta_I^2 = 30$	Process shifted with $\delta_{II}^2 = 0, 5, 10, 15, 20, 25, 30$
4	20% ($\pi = 0.20$) of the data from $\delta_I^2 = 5$	Process shifted with $\delta_{II}^2 = 0, 5, 10, 15, 20, 25, 30$
5	20% ($\pi = 0.20$) of the data from $\delta_I^2 = 30$	Process shifted with $\delta_{II}^2 = 0, 5, 10, 15, 20, 25, 30$

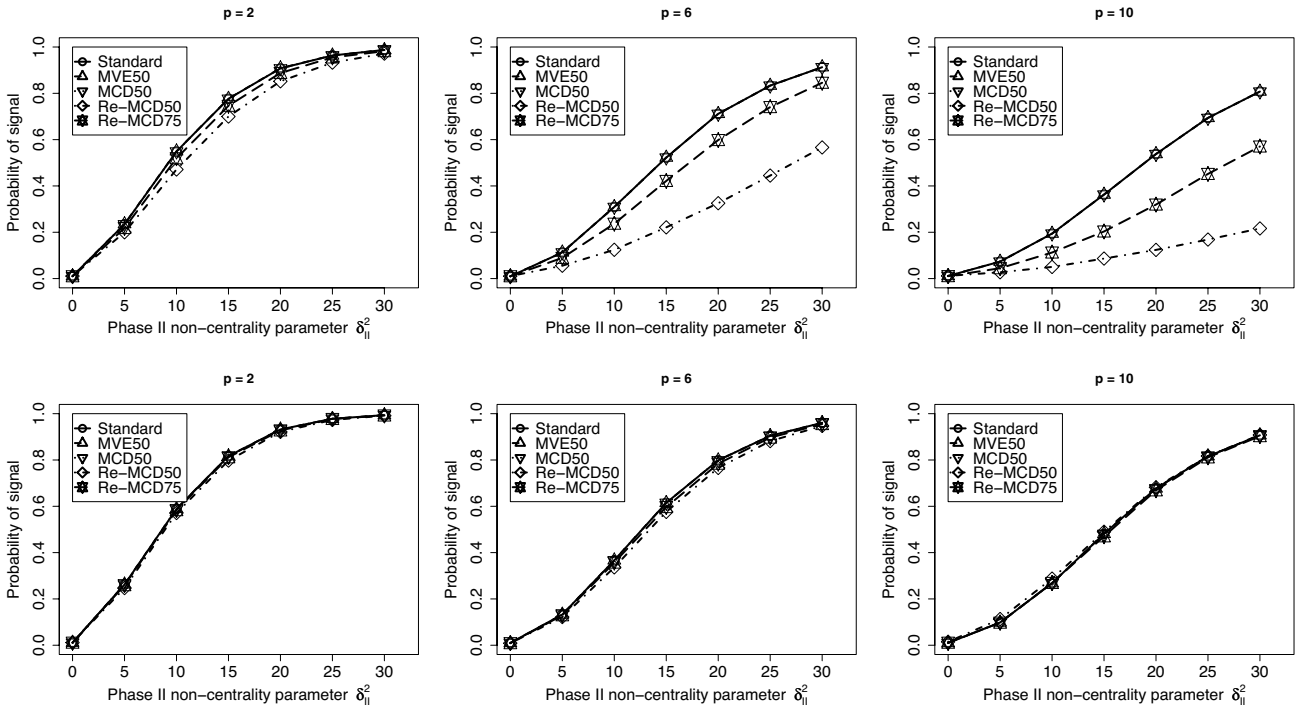

FIGURE 2. Probability of Signal When the Phase I Data Sets of Size $n = 50$ (Upper Panel) and $n = 150$ (Lower Panel) Are Outlier Free (See Case 1 in Table 2).

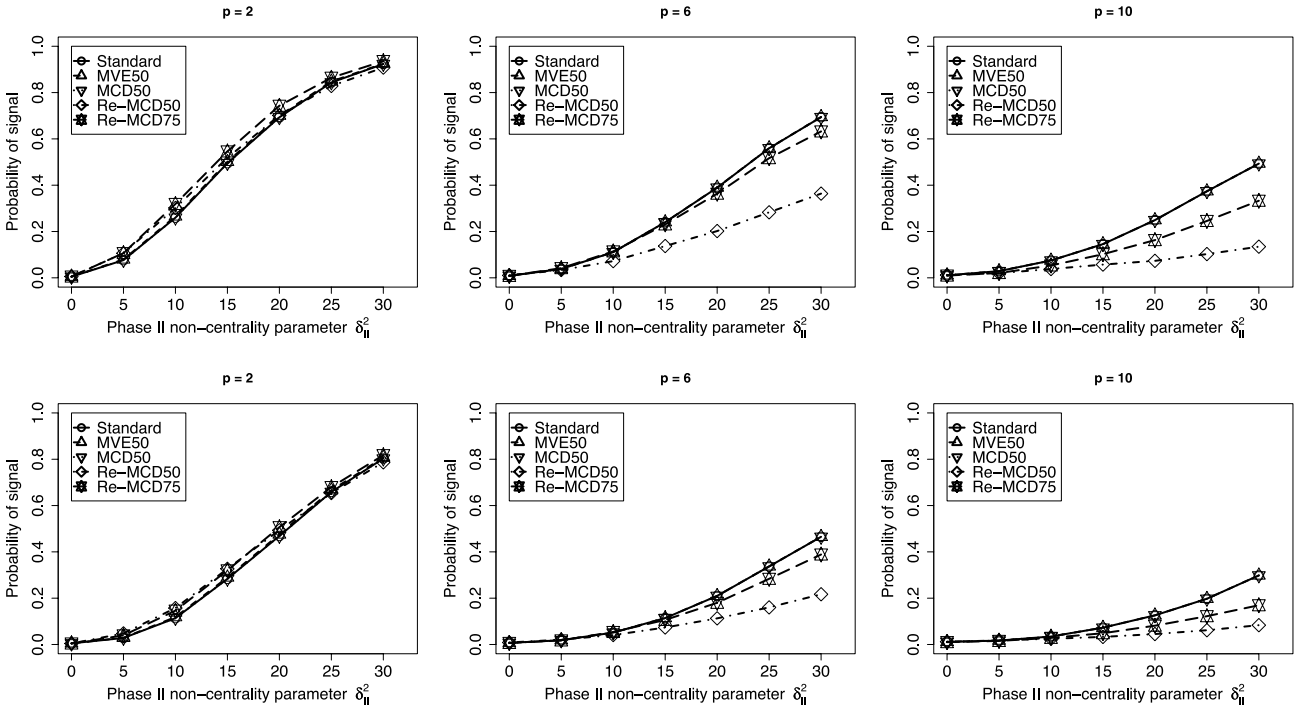

FIGURE 3. Probability of Signal When the Phase I Data Set Has 10% (Upper Panel) and 20% (Lower Panel) Outliers with $\delta_I^2 = 5$ and Sample Size $n = 50$ (see Cases 2 and 4 in Table 2).

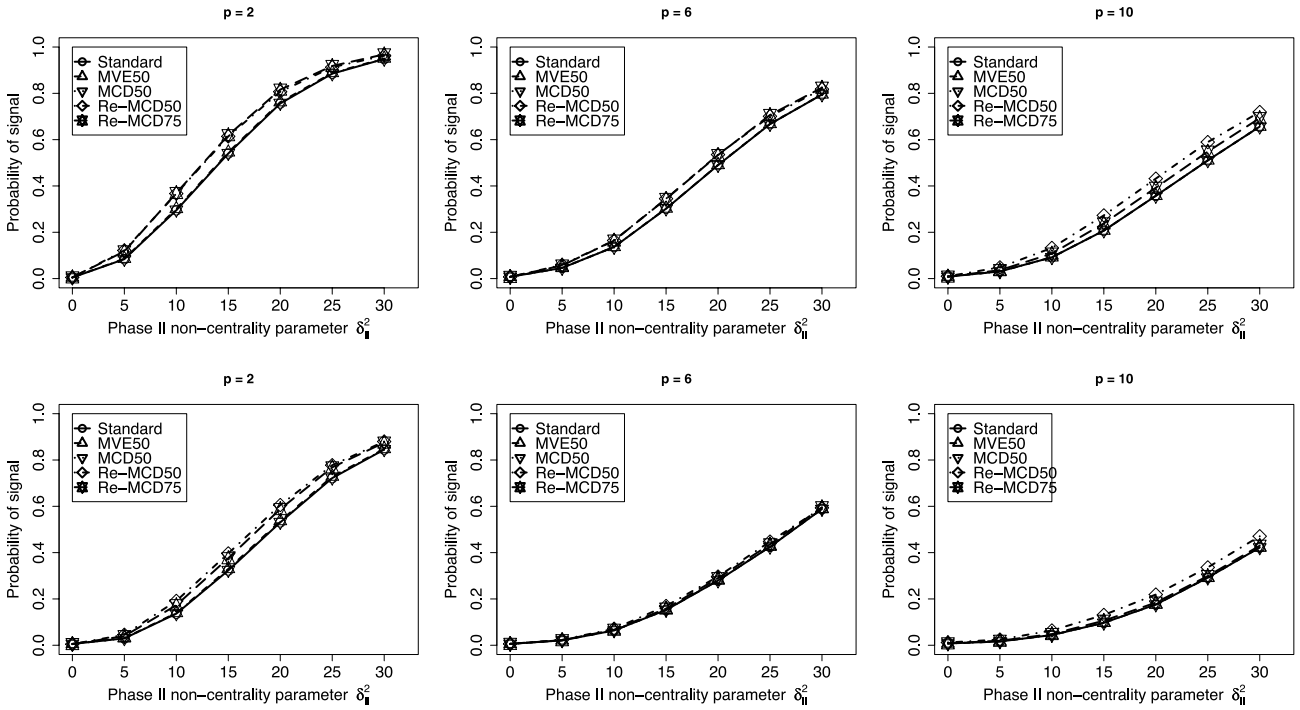

FIGURE 4. Probability of Signal When the Phase I Data Set Has 10% (Upper Panel) and 20% (Lower Panel) Outliers with $\delta_I^2 = 5$ and Sample Size $n = 150$ (See Cases 2 and 4 in Table 2).

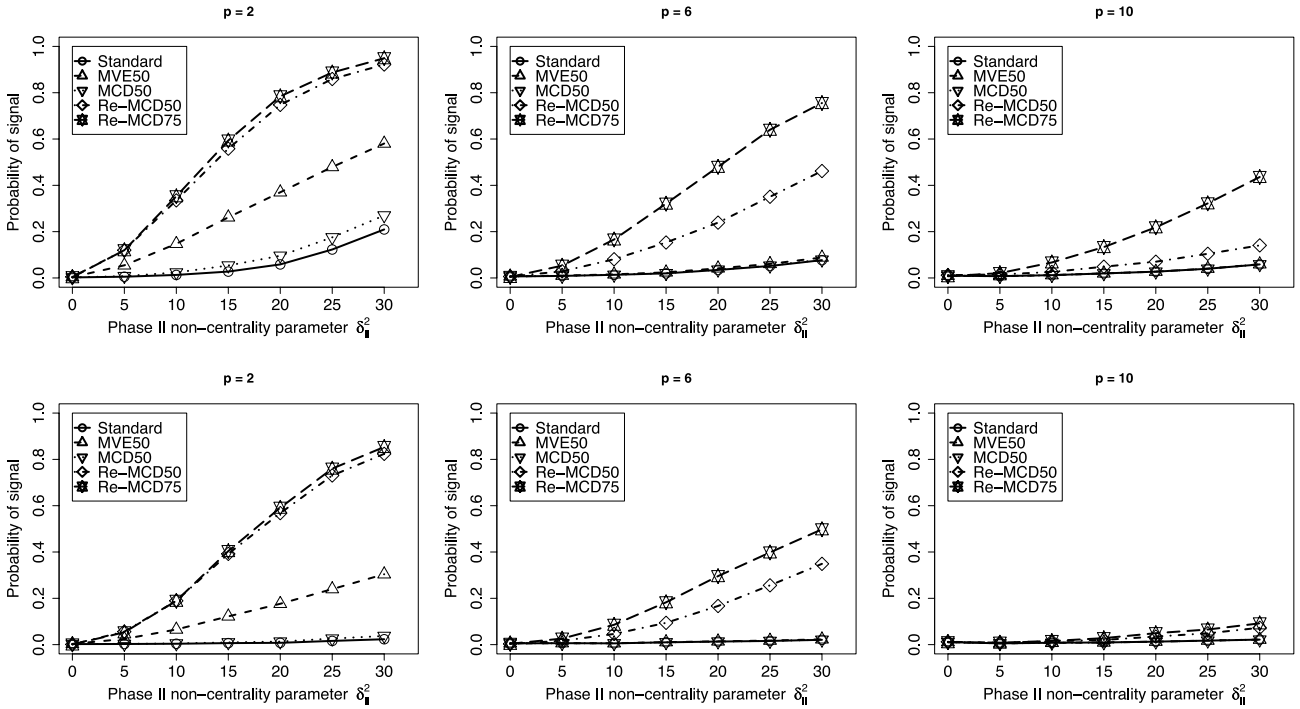

FIGURE 5. Probability of Signal When the Phase I Data Set Has 10% (Upper Panel) and 20% (Lower Panel) Outliers with $\delta_I^2 = 30$ and Sample Size $n = 50$ (See Cases 3 and 5 in Table 2).

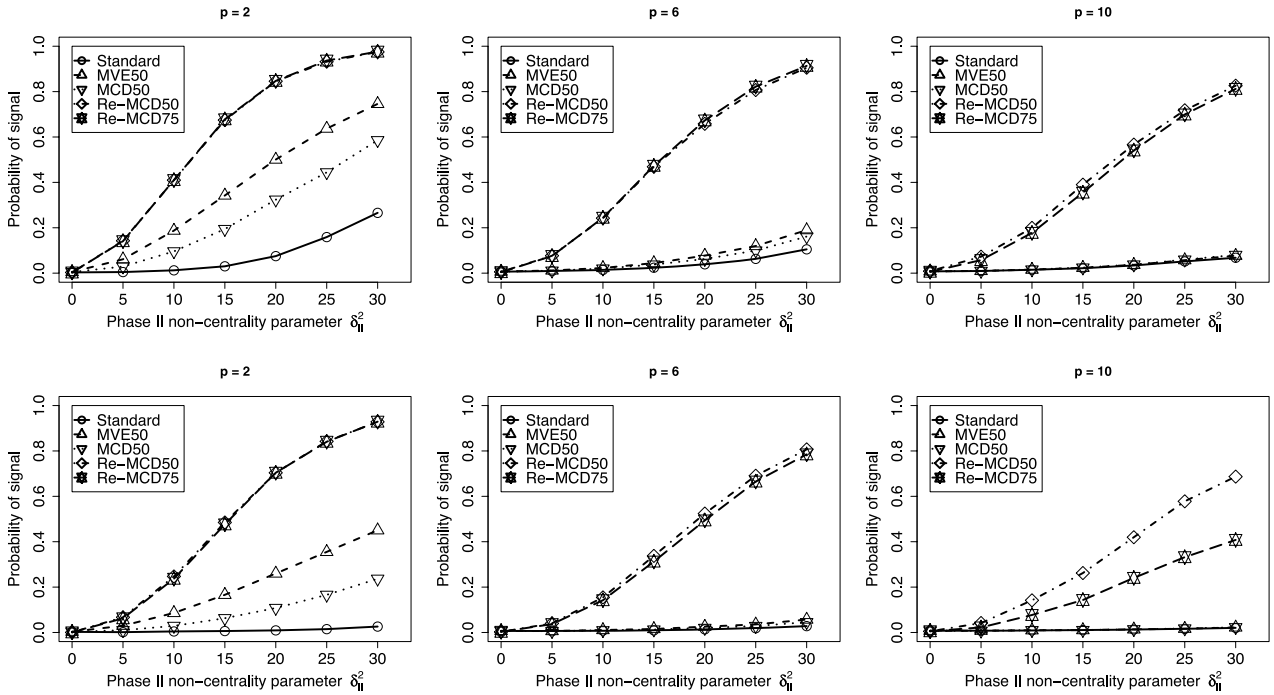

FIGURE 6. Probability of Signal When the Phase I Data Set Has 10% (Upper Panel) and 20% (Lower Panel) Outliers with $\delta_I^2 = 30$ and Sample Size $n = 150$ (See Cases 3 and 5 in Table 2).

chart based on the outlier-free Phase I data with an appropriate quantile of F distribution to monitor Phase II data. We denote these methods by Standard, MCD50, and MVE50, respectively, in Figures 2–6.

From Figure 2, we see that, when the Phase I data set is outlier free ($\pi = 0$ or $\delta_I^2 = 0$), the probability of signal is increasing as the value of the Phase II noncentrality parameter δ_{II} increases. For $n = 50$ and $p = 2$, all five methods perform similarly, but as dimensionality of data increases, ($p = 6, 10$) Re-MCD50 and Re-MCD75 methods do not perform as well as the other three methods. This is expected because, if there are no outliers in Phase I, it is best to use the efficient standard T^2 in Phase II. On the other hand, as the sample size increases ($n = 150$), the probabilities of signal for Re-MCD50 and Re-MCD75 are similar to that of the Standard, MCD50, and MVE50 charts.

Figure 3 shows the signal probabilities when $n = 50$ Phase I data are generated with $\pi = 0.10, 0.20$ and noncentrality parameter of $\delta_I^2 = 5$. As we see, for $p = 2$, Re-MCD50 and Re-MCD75 perform slightly better than the three other methods, but for $p = 6, 10$ and a sample size of $n = 50$, none of the methods work well. If we increase the sample size to $n = 150$ (Figure 4), then the Re-MCD50 and Re-MCD75 methods slightly outperform the other three methods for all dimensions $p = 2, 6, 10$. Figures 5 and 6 show that, when the noncentrality parameter in Phase I is large ($\delta_I^2 = 30$), Re-MCD50 and Re-MCD75 substantially outperform the Standard, MCD50, and MVE50 charts for both sample sizes $n = 50$ and $n = 150$.

It is worthwhile to note that the performance of Re-MCD50 in high dimensions and small sample sizes is not as good as Re-MCD75 in all cases we considered, but is still better than the standard, MCD50, and MVE50 charts for large values of the Phase I noncentrality parameter. On the other hand, when sample size is increased to 150, both Re-MCD50 and Re-MCD75 have more or less similar performance for a small percentage of outliers in the Phase I data. For a large percentage of outliers in Phase I with a high noncentrality parameter, Re-MCD50 out-performs Re-MCD75. This indicates that, if we have sufficiently large sample size, Re-MCD50 is preferable to Re-MCD75. We recommend that for a breakdown point of $1 - \gamma = 0.5$, a Phase I sample size of 10 to 15 times the dimension (p) is sufficient.

Case Study

To illustrate the use of the proposed T^2_{RMCD} control chart, we return to the automotive-alignment example discussed in the Introduction. For the Phase I data, which consist of all 186 vehicles produced during a specific time interval on January 2nd, the reweighted MCD mean vector and covariance matrix with $1-\gamma = 0.5$ (i.e., with the highest breakdown point 0.5) are

$$\overline{\mathbf{x}}_{\text{RMCD}} = (0.303, 0.431, 3.760, 4.045)'$$

$$\mathbf{S}_{RMCD} = \begin{pmatrix} 0.016 & -0.003 & 0.002 & -0.006 \\ -0.003 & 0.020 & -0.002 & 0.010 \\ 0.002 & -0.002 & 0.048 & -0.005 \\ -0.006 & 0.010 & -0.005 & 0.059 \end{pmatrix}$$

To set the control limits, we consider $\alpha = 0.01$ and 0.001. From Table 1 and Equation (12), the 99% and 99.9% control limits are given by the following functions, respectively:

$$f_{4,0.99}(n) = 13.277 + \frac{110115.9}{n^{2.420}},$$

$$f_{4,0.999}(n) = 18.467 + \frac{1897062}{n^{2.917}}.$$

Hence, for $n = 186$ the 99% and 99.9%, control limits are $f_{4,0.99}(186) = 13.63$ and $f_{4,0.999}(186) = 18.92$, respectively. As we see, these control limits are very close to the asymptotic control limits 13.277 and 18.467 (based on the chi-square distribution with 4 degrees of freedom) because the sample size is reasonably large. Using Equation (9), the individual T^2_{RMCD} values for Phase I are calculated and depicted in Figure 7. Note that, using the reweighted MCD control chart, we do not need to take any action, such as removing outlying points and reconstructing the control chart because we are using the robust control chart T^2_{RMCD} in Phase II.

Using the Phase I robust estimates, we constructed the Phase II control chart for the future observations. A control chart for the 100 vehicles produced on January12th is shown in Figure 8. From Figure 8, we see that there is a mean shift in the process (assuming that the covariance structure remains the same) because a large number of T^2_{RMCD} values fall above the 99% and 99.9% control limits. The number of out-of-control points for Re-MCD50 is 43 and 18 for the 99% and 99.9% control limits, respectively. For illustration, we also implemented standard T^2, T^2_{MCD}, and T^2_{MVE} (these three additional charts are not shown in the paper) for Phase I to identify outliers and then, after removing outliers, in Phase II, we apply the standard T^2 charts. MCD50 and MVE50 methods only identified one sample point as an outlier in Phase I. The number of data points above the 99% and 99.9% limits for all of these three charts are more or less similar (40 and 13 or 14, respectively). Because the Phase I sample is almost outlier free, the outlier detection pattern for Phase II is more or less similar for the four charts. The data are available on request from the first author.

Discussion and Conclusions

In this paper, we proposed a multivariate robust Hotelling T^2 chart based on reweighted MCD estimates as an alternative to the classical multivariate T^2 control charts for Phase II data. The proposed control chart is obtained by replacing the classical mean vector and covariance matrix of the data in the Hotelling's T^2 by the reweighted MCD estimators. These estimators are affine equivariant and highly robust with better efficiency than the ordinary MCD estimators used in Vargas (2003), Hardin and Rocke (2004, 2005) and Jensen et al. (2007) for outlier detection in Phase I. Monte Carlo simulations were carried out to obtain empirical quantiles for reweighted MCD T^2, and these quantiles were modeled to ap-

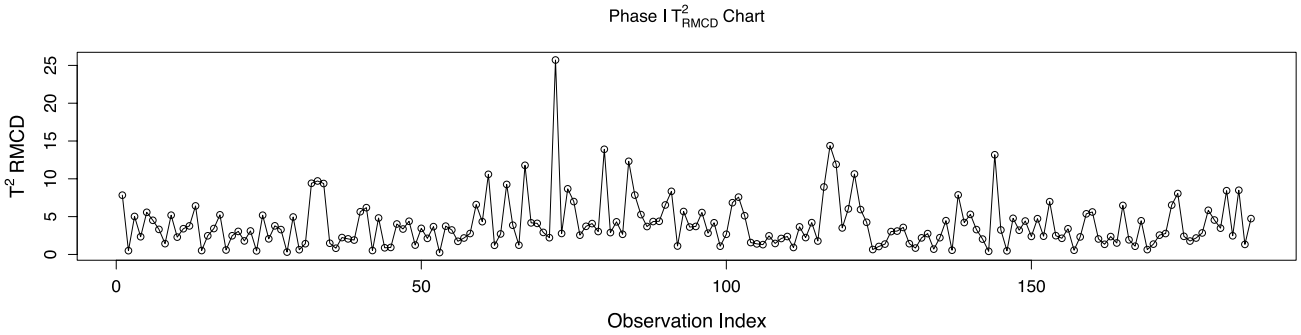

FIGURE 7. Time-Series Plot of the T^2_{RMCD} Chart for 186 Phase I Data Collected on January 2.

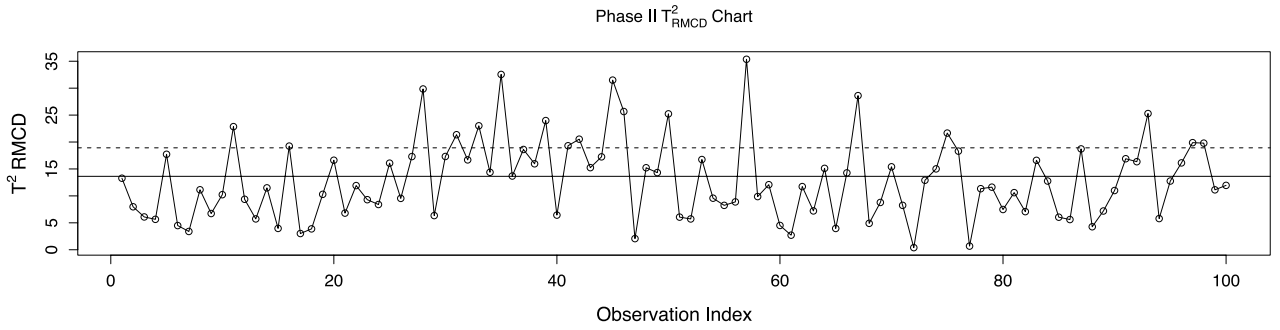

FIGURE 8. The T^2_{RMCD} Chart for 100 Phase II Observations Collected on January 12. The dashed and solid horizontal lines represent control limits based on 99.9% and 99% quantiles, respectively.

proximate control limits for any sample size. Our simulation studies showed that the proposed robust control charts (T^2_{RMCD}) are similar to standard T^2 charts in performance when the process is in-control and are more efficient than standard T^2 charts (with and without outlier removal in Phase I) when there are outliers in the process during Phase I. We illustrated our proposed method using a case study from the automotive industry.

Acknowledgments

The authors would like to thank the editor and two anonymous referees for their valuable comments, which substantially improved the overall presentation of an earlier version of the paper. The work of Drs. Chenouri and Steiner was partially supported by the Natural Sciences and Engineering Research Council of Canada.

References

Butler, R. W.; Davies, P. L.; and Juhn, M. (1993). "Asymptotics for the Minimum Covariance Determinant Estimator". *The Annals of Statistics* 21, pp. 1385–1400.

Croux, C. and Haesbroeck, G. (1999). "Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator". *Journal of Multivariate Analysis* 71, pp. 161–190.

Davies, P. L. (1987). "Asymptotic Behavior of S-Estimates of Multivariate Location Parameters and Dispersion Matrices". *The Annals of Statistics* 15, pp. 1269–1292.

Davies, P. L. (1992). "The Asymptotics of Rousseeuw's Minimum Volume Ellipsoid Estimator". *The Annals of Statistics* 20, pp. 1828–1843.

Donoho, D. L. (1982). "Breakdown Properties of Multivariate Location Estimators". Ph.D. Qualifying Paper, Harvard University.

Donoho, D. L. and Huber, P. J. (1983). "The Notion of Breakdown Point". In *A Festschrift for Erich L. Lehmann in Honor of His Sixty-Fifth Birthday*, P. J. Bickel, K. A. Doksum, and J. L. Hodges, Jr., eds., pp. 157–184. Belmont, CA: Wadsworth.

Gnanadesikan, R. and Kettenring, J. R. (1972). "Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data". *Biometrics* 28, pp. 81–124.

Hardin, J. and Rocke, D. M. (2004). "Outlier Detection in the Multiple Cluster Setting Using the Minimum Covariance Determinant Estimator". *Computational Statistics & Data Analysis* 44, pp. 625–638.

Hardin, J. and Rocke, D. M. (2005). "The Distribution of Robust Distances". *Journal of Computational and Graphical Statistics* 14, pp. 928–946.

Hawkins, D. M. and Olive, D. J. (1999). "Improved Feasible Solution Algorithm for High Breakdown Estimation". *Computational Statistics and Data Analysis* 30, pp. 1–11.

Hotelling, H. (1947). In *Techniques of Statistical Analysis*, C. Eisenhart, H. Hastay, and W. A. Wallis, eds., pp. 111–184. New York, NY: McGraw-Hill.

Huber, P. J. (1981). *Robust Statistics*. New York, NY: John Wiley and Sons.

Jensen, W. A.; Birch, J. B.; and Woodall, W. H. (2007). "High Breakdown Estimation Methods for Phase I Multivariate Control Charts". *Quality and Reliability Engineering International* 23(5), pp. 615–629.

Lopuhaä, H. P. (1989). "On the Relation Between S-Estimators and M-Estimators of Multivariate Location and Covariance. *The Annals of Statistics* 17, pp. 1662–1683.

Lopuhaä, H. P. (1999). "Asymptotics of Reweighted Estimators of Multivariate Location and Scatter". *The Annals of Statistics* 27, pp. 1638–1665.

Lopuhaä, H. P. and Rousseeuw, P. J. (1991). "Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices". *The Annals of Statistics* 19, pp. 229–248.

Maronna, R. A. (1976). "Robust M-Estimators of Multivariate Location and Scatter". *The Annals of Statistics* 4, pp. 51–67.

Pison, G.; Van Alest, S.; and Willems, G. (2002). "Small Sample Corrections for LTS and MCD". *Metrika* 55, pp. 111–123.

Rousseeuw, P. J. (1985). "Multivariate Estimation with High Breakdown Point". In *Mathematical Statistics and Applications, Section B*, W. Grossmann, G. Pflug, I. Vincze, and W. Werz, eds., pp 283–297. Dordrecht: Reidel.

ROUSSEEUW, P. J. and LEROY, A. M. (1987). *Robust Regression and Outlier Detection*. New York, NY: John Wiley and Sons.

ROUSSEEUW, P. J. and VAN DRIESSEN, K. (1999). "A Fast Algorithm for the Minimum Covariance Determinant Estimator". *Technometrics* 41, pp. 212–223.

ROUSSEEUW, P. J. and YOHAI, V. (1984). "Robust Regression by Means of S-Estimators". In *Robust and Nonlinear Time Series Analysis. Lecture Notes in Statistics* 26, pp. 256–272. Berlin: Springer.

ROUSSEEUW, P. J. and VAN ZOMEREN, B. C. (1990). "Unmasking Multivariate Outliers and Leverage Points". *Journal of American Statistical Association* 85, pp. 633–639.

SEBER, G. A. F. (1984). *Multivariate Observations*. New York, NY: John Wiley and Sons.

SERFLING, R. (1980). *Approximation Theorems of Mathematical Statistics*. New York, NY: John Wiley and Sons.

STAHEL, W. A. (1981)." Robust Estimation: Infinitesimal Optimality and Covariance Matrix Estimators". Ph.D. Dissertation, ETH, Zurich (in German).

SULLIVAN, J. H. and WOODALL, W. H. (1996). "A Comparison of Multivariate Control Charts for Individual Observations. *Journal of Quality Technology* 28, pp. 398–408.

SULLIVAN, J. H. and WOODALL, W. H. (1998). "Adapting Control Charts for the Preliminary Analysis of Multivariate Observations". *Communications in Statistics, Simulation and Computation* 27, pp. 953–979.

TRACY, N. D.; YOUNG, J. C.; and MASON, R. L. (1992). "Multivariate Control Charts for Individual Observations. *Journal of Quality Technology* 24, pp. 88–95.

VARGAS, J. A. (2003). "Robust Estimation in Multivariate Control Charts for Individual Observations". *Journal of Quality Technology* 35, pp. 367–376.

WILKS, S. S. (1962). *Mathematical Statistics*. New York, NY: John Wiley and Sons.

WILLEMS, G.; PISON, G.; ROUSSEEUW, P. J.; and VAN ALEST, S. (2002). "A Robust Hotelling Test". *Metrika* 55, pp. 125–138.

WOODRUFF, D. L. and ROCKE, D. M. (1994). "Computable Robust Estimation of Multivariate Location and Shape in High Dimension Using Compound Estimators. *Journal of American Statistical Association* 89, pp. 888-896.