Robust Multivariate Control Chart based on Shrinkage for Individual Observations

Elisa Cabana [*1,2] and Rosa E. Lillo[2,3]

[1]IMDEA Networks Institute, Madrid, Spain
[2]uc3m-Santander Big Data Institute, Madrid, Spain
[3]Department of Statistics, University Carlos III of Madrid, Spain

Abstract

A robust multivariate quality control technique for individual observations is proposed, based on the robust reweighted shrinkage estimators. A simulation study is done to check the performance and compare the method with the classical Hotelling approach, and the robust alternative based on the reweighted minimum covariance determinant estimator. The results show the appropriateness of the method even when the dimension or the Phase I contamination are high, with both independent and correlated variables, showing additional advantages about computational efficiency. The approach is illustrated with two real data-set examples from production processes.

Keywords: multivariate process control, reweighted shrinkage estimator, Hotelling T2, reweighted MCD

[*]This research was partially supported by research grants and Project PID2019-104901RB-I00 from Ministerio de Ciencia e Innovacion.

1 Introduction

Statistical process control charts are the most popular tool for monitoring production quality in the industry since they allow to search for abnormal or out-of-control behavior. In the univariate case, the classical approach is to use control charts that study the behavior of the mean and the variability of the process at the same time. These are known as the Shewhart control charts [Shewhart, 1941, Company, 1956, Nelson, 1984, Montgomery, 1997]. These control charts allow defining a quality control process that consists of two distinct phases. In Phase I, historical observations are used to create the control charts for retrospectively testing whether the process is in control detecting abnormal behavior. By removing the out-of-control data, the parameters of the in-control process can be estimated. Once this is accomplished, in Phase II, future samples obtained during the manufacturing process can be monitored.

Nowadays, it is very common that the dataset available is characterized by more than one variable. In this case, monitoring the variables one by one using the univariate control charts can be misleading and the variability due to their relationship would not be taken into account. Therefore, multivariate statistical process control techniques are more appropriate. Harold Hotelling introduced the first approach within this area in his pioneer paper [Hotelling, 1947]. It was basically an extension of the Shewhart control charts to the multivariate case. The problem divides into two cases: (i) when there are groups or subsamples in the data, and (ii) when the data consists of individual observations. In this paper, we focus on the latter case. For the multivariate sample $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ in Phase I, where $\mathbf{x}_i$ represents a $p-$dimensional vector of measurements, the Hotelling T^2 statistic is defined as:

$$T_i^2 = (\mathbf{x}_i - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}), \qquad i = 1, ..., n\,. \tag{1}$$

When the process is in control, the sample is assumed to be independent and follows a multivariate normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is the location vector and $\boldsymbol{\Sigma}$ is the covariance matrix. The interpretation of the T^2 statistic is the weighted distance of any point from the process mean, under stable conditions. A large value of T^2 indicates that the process has shifted in some way.

This definition has a close resemblance to the Mahalanobis distance. Therefore, the appropriate probability limits may be obtained using the known distribution of the corresponding statistic. In the case of $p > 1$ dimensions, the limits are ellipsoids that depend on the paired correlations between the variables [Tracy et al., 1992, Mason and Young, 2002, Mitra, 2008]. The idea is to calculate the value of the Hotelling T^2 statistics for each observation, and if any value falls out of the $p-$dimensional ellipsoid contour, then it will be considered as out-of-control.

In the definition of the Hotelling T^2 statistic in Equation 1, since $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are not known, the sample mean vector and the sample covariance matrix are used to estimate location and dispersion. These estimators, however, are sensitive to the presence of special causes. In fact, Sullivan and Woodall [1996, 1998] showed that a T^2 chart based on the classical sample estimators is not

effective in detecting a shift or a trend in the mean vector. That is the main reason why in this paper, the focus of attention is centered around the idea of using robust estimators of location and covariance in the definition of the Hotelling T^2 control chart.

In this paper, we propose an alternative robust Hotelling T^2 procedure, using the robust shrinkage reweighted (SR) estimators, introduced in Cabana et al. [2019] and Cabana et al. [2020] Section 2, introduces the background about the classical Hotelling T^2 and the RMCD control chart. Section 3 describes the proposed approach based on SR estimators. Also, the quantiles of the proposed robust statistic are estimated through the Monte Carlo technique and the control limit of the proposed control chart is described. Next, Section 4 studies the performance of the new alternative method and it is compared with the classical Hotelling T^2 approach and the RMCD under several simulation schemes. Section 5 describes the performance through a real dataset example and finally, Section 6 provides some conclusions.

2 Multivariate Hotelling T^2

In practice, it is known that the assumption that the Phase I data comes from an in-control process is not always valid. Abnormal observations in Phase I can inflate the control limits and reduce the power to detect process changes in Phase II. For this reason, the first step in Phase I is investigate those observations that turned to be outside the control limits. If the out-of-control observations are found to be due to an identified assignable cause that can be removed, they should be eliminated and the control limits recalculated. The iterative re-estimation procedure in Phase I can eliminate the effect of a small number of very extreme observations. However, this process will fail to detect more moderate outliers, which is known as the masking effect. This fact motivates the use a robust estimators of the mean vector and the covariance matrix, in order to determine an appropriate control limit for Phase II data.

In the case of multivariate individual observations, the Hotelling T^2 statistic is defined using in Equation 1 the classical sample estimators:

$$T_i^2 = (\mathbf{x}_i - \overline{\mathbf{x}})^t \mathbf{C}^{-1} (\mathbf{x}_i - \overline{\mathbf{x}}) \,, \tag{2}$$

where $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^t$, for $i = 1, ..., n$, are the $p-$variate Phase I observations with sample mean $\overline{\mathbf{x}} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i$ and sample covariance matrix $\mathbf{C} = (n-1)^{-1} \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^t$.

For each individual observation, the value of the T^2 statistic is compared with a control limit usually derived by assuming the data are independent multivariate normal, i.e., $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The observations are assumed not to be correlated over time. Under these normality and independence assumptions, the control limit for Phase I data is based on a *Beta* distribution and for a Phase II observation that is independent of the Phase I data, is based on an F distribution [Tracy et al., 1992]. Note that the

definition in Equation 2 uses the classical estimators sample mean and sample covariance matrix that are very sensitive to the presence of outliers. Thus, this classical definition of the Hotelling T^2 statistic can be affected if abnormal observations are present in Phase I [Rousseeuw and Van Zomeren, 1990].

Sullivan and Woodall [1996, 1998] proposed a procedure based on estimating the mean vector and the covariance matrix using successive differences, which is more effective than the classical estimators in detecting the process shift when certain conditions apply. However it is not effective in detecting multiple multivariate outliers because its breakdown point tends to zero, thus it will not be considered here [Jensen et al., 2007]. Following this idea, some alternatives have been proposed in the literature to avoid the negative effect of outliers, such as the Trimmed approach introduced by Alfaro and Ortega [2008]. Although, this method has the disadvantage of being computed only up to dimension $p = 5$. Motivated by this, Vargas [2003] proposed a robust control chart in order to identify outliers in Phase I based on two robust estimators: the minimum covariance determinant (MCD) and the minimum volume ellipsoid (MVE) estimators both proposed by Rousseeuw [1985]. However, when different estimators are used in the T^2 statistic, the assumption about the distribution does not have to be true. Moreover, the exact distribution of the statistic based on the alternative robust estimators is not available. Then, the control limits for the robust T^2 control chart in Phase I should be obtained empirically. Vargas [2003] and Jensen et al. [2007] proposed to estimate the control limits based on simulations.

The later study of Chenouri et al. [2009], followed some of these ideas. The authors proposed to use the reweighted MCD (RMCD) estimator [Rousseeuw and Van Zomeren, 1990, Lopuhaa and Rousseeuw, 1991, Willems et al., 2002] to compute the robust control chart for the Phase II data. The motivation is that the RMCD estimators inherit the nice properties of the initial MCD estimators, such as affine equivariance, robustness, and asymptotic normality, while achieving a higher efficiency. Because of the reweighting step in the calculation of the estimates, the result is not unduly influenced by outliers and there is no need to identify outliers in the Phase I data. The authors concluded that their proposed multivariate robust Hotelling T^2 chart based on RMCD estimates is a good alternative to the classical multivariate T^2 control chart for Phase II data.

Although, there are some disadvantages in the case of using the MCD or the RMCD estimators. The MCD estimators of location and scatter are determined by the subset of size $h = \lfloor n\gamma \rfloor$ (where $0.5 \leq \gamma \leq 1$), for which the covariance matrix has the smallest possible determinant. Thus, the method depends on the parameter h which depends on γ, that should be defined by the user or practitioner in practice. This parameter directly affect the performance of the method, the breakdown value and the efficiency. Here, $1 - \gamma$ represents the (asymptotic) breakdown point of the MCD estimator and thus, of the RMCD. The MCD estimator has its highest possible finite sample breakdown point (bdp) when $h = \lfloor (n + p + 1)/2 \rfloor$ [Leroy and Rousseeuw, 1987]. It is important to note that the computation of the estimates is very challenging, especially in case of large data-sets or high dimension. That is why *approximate* algorithms have to be used for this task. The problem is that this results in worse performance

about consistency and bdp, than the exact theoretical estimator would have had. And it gets worse with the increase of the sample size n and/or the dimension p of the sample (Stromberg et al. [2000], Hawkins and Olive [2002]). Actually, in the statistical software packages, the MCD only runs for dimension $p \leq 30$. Also, the approximate algorithm Fast-MCD still requires substantial runing times for large p, because the number of candidate solutions grows exponentially with the dimension p of the sample and, as a consequence, the procedure becomes computationally expensive for even moderately sized problems.

3 Robust shrinkage reweighted control chart

Cabana et al. [2019] proposed a robust estimator for location and covariance matrix to deal with the problem of outliers in multivariate data. They were based on a notion frequently used in finance and portfolio optimization (see Ledoit and Wolf [2004], Chen et al. [2011], Couillet and McKay [2014] and Steland [2018]), known as *shrinkage*, defined as follows:

$$\hat{E}_{Sh} = (1-\eta)\hat{E} + \eta\hat{T}\,, \tag{3}$$

where $\hat{E}$ is the prior estimator of a parameter θ to which the shrinkage is done, towards a target estimator $\hat{T}$. This way, $\hat{E}_{Sh}$ is a convex combination that relies on the fact that "shrinking" $\hat{E}$ towards $\hat{T}$, would help to reduce the estimation error, because although the target estimator is usually biased, it also contains less variance than the estimator $\hat{E}$. Therefore, under general conditions, there exists a *shrinkage intensity* η, a number between 0 and 1, such that that the resulting shrinkage estimator would contain less estimation error than $\hat{E}$ [James and Stein, 1961]. Cabana et al. [2019] proposed to use the shrinkage to overcome the positive definite condition that the covariance estimator used in the Mahalanobis distance should meet. Since the shrinkage estimator has the advantages of providing a trade-off between bias and variance, and in the case of covariance matrices the shrinkage is always positive definite and well conditioned, we propose to use the shrinkage estimator in our proposal as well.

Let $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ be the $n \times p$ data matrix with n being the sample size and p the number of variables. For location, the authors defined a shrinkage over the multivariate median $\hat{\boldsymbol{\mu}}_{MM}$ called $L_1-median$ which is a robust and highly efficient estimator of central tendency (Lopuhaa and Rousseeuw [1991], Vardi and Zhang [2000], Oja [2010]). The $L_1-median$ is defined as:

$$\hat{\boldsymbol{\mu}}_{MM} = argmin_{\mathbf{x}_m} \frac{1}{n} \sum_{i=1}^{n} ||\mathbf{x}_m - \mathbf{x}_i||_1\,.$$

The shrinkage estimator over $\hat{\boldsymbol{\mu}}_{MM}$ was proposed considering it as the sample estimator $\hat{E}$ and $\nu_{\boldsymbol{\mu}}\mathbf{e}$ as the target estimator $\hat{T}$ in the definition of Equation (3), where $\mathbf{e}$ is the $p-$dimensional vector of ones. Then, the shrinkage estimator over the multivariate L_1-median is:

$$\hat{\boldsymbol{\mu}}_{Sh} = (1-\eta)\hat{\boldsymbol{\mu}}_{MM} + \eta\nu_{\boldsymbol{\mu}}\mathbf{e}\,. \tag{4}$$

The scaling factor $\nu_{\boldsymbol{\mu}}$ and the intensity η should minimize the expected quadratic loss:

$$\begin{array}{cc} \min_{\nu_{\boldsymbol{\mu}},\eta} & E\left[\|\hat{\boldsymbol{\mu}}_{Sh}-\boldsymbol{\mu}\|_2^2\right] \\ \text{s.t.} & \hat{\boldsymbol{\mu}}_{Sh}=(1-\eta)\hat{\boldsymbol{\mu}}_{MM}+\eta\nu_{\boldsymbol{\mu}}\mathbf{e}, \end{array}$$

where $\|\mathbf{x}\|_2^2=\sum_{j=1}^{p}x_j^2$. The parameters $\nu_{\boldsymbol{\mu}}$ and η are estimated robustly and optimally minimizing the estimation error, as proposed in Proposition 2 (page 6) of Cabana et al. [2019]. The authors propose to use the asymptotic distribution for the L_1-median $\hat{\boldsymbol{\mu}}_{MM}$, which can be approximated by $N_p\left(\boldsymbol{\mu},\frac{1}{n}\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}\right)$, [Bose and Chaudhuri, 1993, Bose, 1995, Möttönen et al., 2010], where $\hat{\mathbf{A}}(\mathbf{x}_i)=\frac{1}{||\mathbf{x}_i||_2}\left(\mathbf{I}_p-\frac{\mathbf{x}_i\mathbf{x}_i^T}{||\mathbf{x}_i||_2^2}\right)$ and $\hat{\mathbf{B}}(\mathbf{x}_i)=\frac{\mathbf{x}_i\mathbf{x}_i^T}{||\mathbf{x}_i||_2^2}$, with $\mathbf{x}_i\in\mathbb{R}^p$, for each $i=1,...,n$. Therefore, the parameters $\nu_{\boldsymbol{\mu}}$ and η are estimated in practice as follows:

$$\hat{\nu}_{\boldsymbol{\mu}}=\frac{\hat{\boldsymbol{\mu}}_{MM}\mathbf{e}}{p}\quad\text{and}\quad\hat{\eta}=\frac{trace\left(\frac{1}{n}\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}\right)}{\|\hat{\boldsymbol{\mu}}_{MM}-\hat{\nu}_{\boldsymbol{\mu}}\mathbf{e}\|^2}\,.$$

On the other hand, the authors also propose an adjusted special comedian matrix $\hat{\mathbf{S}}_{Sh}$, based on the classical definition of comedian from Falk [1997]:

$$\hat{\mathbf{S}}_{Sh}=2.198\cdot(median((\mathbf{x}_j-(\hat{\boldsymbol{\mu}}_{Sh})_j)(\mathbf{x}_t-(\hat{\boldsymbol{\mu}}_{Sh})_t))\,.$$

The comedian matrix is a robust alternative for the covariance matrix, but in general it is not positive (semi-)definite (see Falk [1997]). Thus, Cabana et al. [2019] used the shrinkage approach applied to the comedian, and a robust and well-conditioned estimate is obtained (Ledoit and Wolf [2003a], Ledoit and Wolf [2003b], Ledoit and Wolf [2004], DeMiguel et al. [2013]):

$$\hat{\boldsymbol{\Sigma}}_{Sh}=(1-\eta)\hat{\mathbf{S}}_{Sh}+\eta\nu_{\boldsymbol{\Sigma}}\mathbf{I}\,. \tag{5}$$

In this case, the target is the scaled identity matrix, as suggested in the literature. The optimal expression for the parameters η and $\nu_{\boldsymbol{\Sigma}}$ is described in Cabana et al. [2019] in Proposition 3 (page 10). In this paper, the robust estimators of location $\hat{\boldsymbol{\mu}}_{Sh}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_{Sh}$ based on shrinkage are used to define a robust Mahalanobis distance that has the ability to discover outliers with high precision in the vast majority of cases in the simulation scenarios studied in the paper, with both gaussian data and with skewed or heavy-tailed distributions. The behavior under correlated and transformed data showed that the approach was approximately affine equivariant. With highly contaminated data it is shown that the method had high breakdown value even in high dimension or large level of contamination. Furthermore, the significantly low computational times show the advantages of the proposal.

These robust estimators are also studied in Cabana et al. [2020], and are used to define a robust regression approach, for which they used the reweighted version of $\hat{\boldsymbol{\mu}}_{Sh}$ and $\hat{\boldsymbol{\Sigma}}_{Sh}$ (Equations 4 and 5). To define those reweighted estimators, the associated robust squared Mahalanobis distance must be defined,

for each observation $\mathbf{x}_i$, $i = 1, ..., n$:

$$d^2(\mathbf{x}_i) = (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{Sh})^t \hat{\boldsymbol{\Sigma}}_{Sh}^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{Sh}) \,.$$

Cabana et al. [2020] defined $w_i = w(d^2(\mathbf{x}_i))$, a weight function depending on the robust squared Mahalanobis distance, as an indicator function which assigns weight 1 to the $\mathbf{x}_i$ having a distance less than certain quantile of the chi-square distribution with $p+1$ degrees of freedom: $q_{\delta_l} = \chi^2_{p+1,1-\delta_l}$, $\delta_l = 0.025$.

$$w(d^2(\mathbf{x}_i)) = I(d^2(\mathbf{x}_i) \leq q_{\delta_l}) \,.$$

Then, the *shrinkage reweighted (SR)* estimators $\hat{\boldsymbol{\mu}}_{SR}$ and $\hat{\boldsymbol{\Sigma}}_{SR}$ for location and covariance matrix, respectively, are:

$$\hat{\boldsymbol{\mu}}_{SR} = \frac{\sum_{i=1}^n w_i \mathbf{x}_i}{\sum_{i=1}^n w_i}, \quad \hat{\boldsymbol{\Sigma}}_{SR} = \frac{\sum_{i=1}^n w_i (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{SR})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{SR})^t}{\sum_{i=1}^n w_i} \,. \tag{6}$$

The reweighting step in the SR estimators provides the advantage of keeping the properties of the initial shrinkage estimators, such as robustness, approximate affine equivariance and high breakdown value, while achieving a higher efficiency.

In the statistical process control procedure, with the proposed estimators there is no need to identify outliers in the Phase I data because the estimators are not influenced by outliers. Therefore, using them will provide us with robust Phase I estimates of location and covariance that will serve to compute a robust control chart for Phase II data. Next, the procedure for the main contribution of this paper: defining a robust quality control approach based on the SR estimators, is described in detail.

Let $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ be the $n \times p$ data matrix in Phase I, with n being the sample size and p the number of variables. Let us assume that it follows a multivariate normal distribution $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Since we are in the case of individual observations, it is well known [Wilks, 1962] that for a new observation in Phase II, $\mathbf{x}_k$ we have:

$$T_k^2 \sim \frac{p(n+1)(n-1)}{n(n-p)} F_{p,n-p} \,,$$

where T_k^2 is defined as in Equation 2. In this paper, we propose to robustify the definition of the Hotelling T^2 control chart based on Phase I data, replacing the classical estimators with the robust shrinkage reweighted (SR) estimators $\hat{\boldsymbol{\mu}}_{SR}$ and $\hat{\boldsymbol{\Sigma}}_{SR}$, defined in Equation 6. The proposed robust Hotelling T^2 for the Phase II observation $\mathbf{x}_k$ is:

$$T^2_{SR}(\mathbf{x}_k) = (\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{SR})^t \hat{\boldsymbol{\Sigma}}_{SR}^{-1} (\mathbf{x}_k - \hat{\boldsymbol{\mu}}_{SR}) \,. \tag{7}$$

Note that this definition is similar to the definition of a squared robust Mahalanobis distance. It is well known that the distribution of the classical

squared Mahalanobis distance, i.e. with sample mean and covariance matrix, is a chi-square with p degrees of freedom. When different estimators are used, this assumption does not have to be true. Although, in the literature the 0.975 quantile of the chi-square distribution is usually used to determine a threshold for the distance to detect multivariate outliers. In this paper we apply Monte Carlo simulations to estimate the quantiles of the distribution of T^2_{SR} for several combinations of sample sizes n and dimensions p, similar as in Chenouri et al. [2009], but considering a wider range of n and p. For each dimension, a smooth curve between the sample size and quantiles of T^2_{SR} is fitted and used to estimate appropriate quantiles of T^2_{SR} for different Phase I sample sizes.

The simulation study is organized as follows:

- In order to estimate the 99% and 99.9% quantiles of T^2_{SR} for a given Phase I sample size n and a dimension p, M = 10.000 samples of size n from a standard multivariate normal distribution $N_p(\mathbf{0}, \mathbf{I}_p)$ are generated.

- For each data set of size n, the SR location and covariance matrix estimates are computed, $\hat{\boldsymbol{\mu}}_{SR}(m)$ and $\hat{\boldsymbol{\Sigma}}_{SR}(m)$, $m = 1, ..., M$.

- In addition, for each data set, we randomly generate a new observation $\mathbf{x}_{k,m}$ from $N_p(\mathbf{0}, \mathbf{I}_p)$, which is treated as a Phase II observation, and calculate the corresponding $T^2_{SR}(\mathbf{x}_{k,m})$ value as given by Equation 7.

  The empirical distribution function of T^2_{SR} is based on the simulated values:

  $$T^2_{SR}(\mathbf{x}_{k,1}),\ T^2_{SR}(\mathbf{x}_{k,2}),\ T^2_{SR}(\mathbf{x}_{k,3}),\ ...,\ T^2_{SR}(\mathbf{x}_{k,M})\,.$$

- With this values, the empirical distribution function of T^2_{SR} can be obtained, for different combinations of $p = 2, ..., 30$ and $n = 20, ..., 300$.

- By inverting the empirical distribution function, the Monte Carlo estimates of the 99% and 99.9% quantiles of T^2_{SR} can be obtained.

Figures 2-7 from the Appendix A show the scatterplots of the empirical 99% and 99.9% quantiles of T^2_{SR} against the sample n for the selected dimensions $p = 5, 10, 30$ to illustrate the results. The graphical results for different dimensions are similar and they suggest that the quantiles could be modeled using a family of regression curves of the form:

$$f(n) = a_3 + \frac{a_1}{n^{a_2}}\,, \tag{8}$$

where $a_3 = \chi^2_{(p,1-\alpha)}$ as recommended in Chenouri et al. [2009], and a_1 and a_2 are constants that depend on p, and $1-\alpha$ as well. Fitting the curve to the data in each case, will help predict the quantiles of $T^2_{SR}(\mathbf{x}_k)$ for any Phase I sample size n. Note that, as n increases, $f(n)$ approaches $\chi^2_{(p,1-\alpha)}$, the quantile used in the literature.

Table 1: Least-squares estimates of the regression parameters a_1, a_2 and R^2, for dimensions $p = 2, ..., 30$ and confidence levels $1 - \alpha = 0.99, 0.999$, where $a_3 = \chi^2_{p;1-\alpha}$

	$1-\alpha = 99\%$				$1-\alpha = 99.9\%$			
p	a_1	a_2	a_3	R^2	a_1	a_2	a_3	R^2
2	204	0.850	9.210	0.9976	113000	2.827	13.816	0.9909
3	1082	1.146	11.345	0.9958	42040	2.152	16.266	0.9830
4	2804	2.302	13.277	0.9967	1174000	2.880	18.467	0.9968
5	122000	2.086	15.086	0.9981	426100	2.458	20.515	0.9988
6	327500	2.253	16.812	0.9985	1152000	2.607	22.458	0.9987
7	919800	2.435	18.475	0.9986	1903000	2.651	24.322	0.9989
8	1609000	2.503	20.090	0.9990	3092000	2.691	26.125	0.9992
9	2867000	2.600	21.666	0.9992	5607000	2.778	27.877	0.9993
10	4319000	2.644	23.209	0.9992	6902000	2.777	29.588	0.9993
11	48000000	3.172	24.725	0.9970	53680000	3.207	31.264	0.9970
12	3974000000	4.497	26.217	0.9977	318900000000	5.893	32.910	0.9982
13	4045000000	4.418	27.688	0.9975	4553000000	4.457	34.528	0.9976
14	375200000000	5.625	29.141	0.9977	300000000000	5.562	36.123	0.9977
15	1007000000	3.729	30.578	0.9951	31160000000	4.525	37.697	0.9953
16	10630000	2.647	32.000	0.9991	9566000	2.624	39.252	0.9994
17	49530000	2.972	33.409	0.9978	44240000	2.943	40.790	0.9983
18	62590000	3.004	34.805	0.9974	54730000	2.971	42.312	0.9976
19	122100000	3.140	36.191	0.9964	114700000	3.121	43.820	0.9964
20	263000000	3.295	37.566	0.9953	217600000	3.233	45.315	0.9963
21	1354000000	3.696	38.932	0.9886	654900000	3.515	46.797	0.9897
22	24740000	2.695	40.289	0.9978	13240000	2.544	48.268	0.9990
23	44500000	2.807	41.638	0.9965	33560000	2.734	49.728	0.9970
24	171900000	3.098	42.980	0.9925	113700000	2.993	51.179	0.9933
25	156400000	3.055	44.314	0.9933	111000000	2.967	52.620	0.9933
26	326400000	3.208	45.642	0.9911	222600000	3.108	54.052	0.9910
27	450600000	3.257	46.963	0.9918	250600000	3.110	55.476	0.9926
28	52120000	2.740	48.278	0.9959	37440000	2.653	56.892	0.9965
29	11060000	2.392	49.588	0.9964	5046000	2.218	58.301	0.9983
30	14890000	2.438	50.892	0.9976	11220000	2.362	59.703	0.9981

Table 1 shows the resulting least-square estimates of the parameters of the curves a_1 and a_2 for each dimension $p = 2, ..., 30$ considered and both 99% and 99.9% quantiles. In practice, using Equation 8 and the values of the constants from Table 1, the 99% and 99.9% quantiles can be computed for any dimension between 2 and 30 and any Phase I sample size. The estimated curves fit well to the data for all cases yielding R^2 values of at least 98%. Furthermore, the significance tests yielded p-values less than 0.05, which allows to conclude that there is a statistically significant linear relationship between the quantiles and $\frac{1}{n}$. For dimensions $p > 30$ the practitioner can simulate the control limits following

the described steps, but a similar behavior is expected. Since the method is going to be compared with another that has this dimension restriction, the results of this paper only consider dimensions $p \leq 30$, but the proposed method does not have any restriction with respect to the dimension. The simulations were done in Matlab in a PC with a 3.40 GHz Intel Core i7 processor with 32GB RAM.

The implementation in practice to construct the T^2_{SR} control chart is as follows:

Phase I:

1. Save the sample size n and the dimension p of the data. Select the confidence level $1-\alpha$, with $\alpha = 0.01,\ 0.001$.
2. Using the Phase I data, compute the shrinkage reweighted (SR) estimates of location and covariance matrix.
3. For the desired α and p values, choose the least square estimates a_1 and a_2 from Table 1 and compute the control limit using Equation 8.

Phase II:

4. Compute the value $T^2_{SR}(\mathbf{x}_k)$ for each new observation $\mathbf{x}_k$ defined in Equation 7, and plot each value on a control chart with the limit derived in Phase I (step 3).
5. Interpret the chart and look for out-of-control observations or patterns. Diagnose the process if needed.

4 Performance simulation study

The performance of the proposed SR control chart is compared to the classical Hotelling T^2 and the RMCD control charts. The Phase I data structure is considered under a multivariate normal distribution, considering the presence of outliers. The behavior is studied through the shifts in the process mean vector in the Phase II data, for both cases of independent or correlated variables. The shifts are defined through the non-centrality parameter (ncp) ξ^2:

$$\xi^2 = (\boldsymbol{\mu} - \boldsymbol{\mu}_A)^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_A)\,,$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_A$ represent the in-control and the out-of-control mean vectors, respectively. The larger the value of ξ^2 is, the more extreme the outliers are. The in-control (outlier-free) Phase I data are generated from a multivariate normal distribution $N_p(\mathbf{0}, \boldsymbol{\Sigma})$. Then, some abnormal observations are included, generated from $N_p(\boldsymbol{\mu}_I, \boldsymbol{\Sigma})$. The Phase II data are generated from $N_p(\boldsymbol{\mu}_{II}, \boldsymbol{\Sigma})$. Since there is no change in the covariance structure, the ncp in Phase I is $\xi^2_I = \|\boldsymbol{\mu}_I\|$ and in Phase II is $\xi^2_{II} = \|\boldsymbol{\mu}_{II}\|$. Different scenarios are considered

varying the following elements: number of observations (n), number of variables (p), proportion of outliers (π), proportion of dimensions to be shifted (ϑ) and the parameters in the distributions, i.e. $\boldsymbol{\mu}_I$, $\boldsymbol{\mu}_{II}$ and $\boldsymbol{\Sigma}$.

This simulation study is analogous as the one from Chenouri et al. [2009], with three main improvements: (i) the authors studied the problem up to dimension $p = 10$, and here we will consider higher dimensions, (ii) they considered the identity matrix $\boldsymbol{\Sigma} = \mathbf{I}_p$ as the covariance structure of both the clean data and the outliers, and here we propose to study also correlated variables as in the comparison study from Alfaro and Ortega [2009], and (iii) we additionally consider to study not only the magnitude of the shift but also where it happens.

The Phase I and II data generation models we propose are the following:

- **Case A**:
  Phase I: $(1-\pi)N_p(\mathbf{0}, \mathbf{I}_p) + \pi N_p(\boldsymbol{\mu}_I, \mathbf{I}_p)$.
  Phase II: $N_p(\boldsymbol{\mu}_{II}, \mathbf{I}_p)$

- **Case B**:
  Phase I: $(1-\pi)N_p(\mathbf{0}, \boldsymbol{\Sigma}) + \pi N_p(\boldsymbol{\mu}_I, \boldsymbol{\Sigma})$.
  Phase II: $N_p(\boldsymbol{\mu}_{II}, \boldsymbol{\Sigma})$

The Phase I sample sizes considered are: $n = 50, 150$ with the respective dimensions $p = 2, 6, 10, 30$. The confidence level studied is set to be $\alpha = 0.01$. The level of contamination in the Phase I data is characterized by $\pi = 0, 0.1, 0.2$. The process mean shifts considered in Phase I are $\xi_I^2 = 5,\ 30$ and in Phase II are $\xi_{II}^2 = 0,\ 5,\ 10,\ 15,\ 20,\ 25,\ 30$. Two covariance structures $\boldsymbol{\Sigma}$ are considered: (Case A) the identity matrix to compare the methods when there are different-sized changes in the average of all the variables if the variables are independent, and (Case B) the matrix of size p with 1 on the main diagonal and 0.9 elsewhere, to analyze whether the correlation level affects the detection probability of each alternative. Table 2 summarizes the different combinations for the other parameters in the simulations study, for both cases A and B.

Table 2: Combinations in the simulation study for both cases A and B

Combination	Phase I	Phase II
1	No outliers ($\pi = 0$)	Shift $\xi_{II}^2 = 0, 5, 10, 15, 20, 25, 30$
2	10% Outliers $\pi = 0.10$ with $\xi_I^2 = 5$	Shift $\xi_{II}^2 = 0, 5, 10, 15, 20, 25, 30$
3	10% Outliers $\pi = 0.10$ with $\xi_I^2 = 30$	Shift $\xi_{II}^2 = 0, 5, 10, 15, 20, 25, 30$
4	20% Outliers $\pi = 0.20$ with $\xi_I^2 = 5$	Shift $\xi_{II}^2 = 0, 5, 10, 15, 20, 25, 30$
5	20% Outliers $\pi = 0.20$ with $\xi_I^2 = 30$	Shift $\xi_{II}^2 = 0, 5, 10, 15, 20, 25, 30$

The following results show the performance when the process mean shifts happen in all dimensions. We also further studied the case in which the shifts happens in some percentage of the corresponding dimension, instead of happening in all, to see if the performance of the control chart is not only affected by the magnitude of the shift but also where it happened. We defined a parameter ϑ for this purpose. The values of ϑ are $100\%, 75\%, 50\%$ and 25%, each representing

the scenarios in which the shift happens in all dimensions, most dimension, half of the dimensions and some dimensions, respectively. Note that this is done for each dimension considered in the simulations, i.e. $p = 2, 6, 10, 30$, each value of the sample size $n = 50, 150$, each combination from the Table 2 considering the percentage of outliers π and the process mean shifts. And of course, this is also studied in both cases A and B, i.e. with no correlations and with correlations. Since the results with $\vartheta = 75\%, 50\%, 25\%$ are very similar to when it is 100%, and to avoid unnecessary extension of the paper, we include only the results when $\vartheta = 100\%$, i.e., the shift happens in all dimensions. For both cases A and B described as follows, and all the combinations of the dimension, sample size, percentage of outliers and shift, the proposed control chart behaves similar, not only when the shift happens in all dimensions, but also when it happens in some percentage of the dimensions, set by the parameter ϑ. The results for when $\vartheta = 75\%, 50\%, 25\%$ can be found in the Supplementary Material.

4.1 Case A

The performance of the control charts is studied by means of the probability of signal that is estimated as the proportion of $T^2_{SR}(\mathbf{x}_k)$ values (Equation 7), that fall above the control limit based on 10.000 simulations. The probability of signal is also obtained for the classical Hotelling T^2 control chart and the robust procedure based on RMCD proposed by Chenouri et al. [2009]. Figures 8-17 from the Appendix A show the results from all the different scenarios in the simulation study, in which the probability of signal is obtained for each Phase II non-centrality parameter. Figure 8 shows the resulting probability of signal when the Phase I dataset is outlier free, i.e. $\pi = 0$ and the sample size is $n = 50$. In all cases the probability of detecting the Phase II shifts increases with the increase of the shift, i.e. the non-centrality parameter ξ^2_{II}. In case of low dimension $p = 2$, all methods behave similarly. However, when the dimension of the data increases, method T^2_{RMCD} decrease its performance. Meanwhile, in these cases, the proposed method T^2_{SR} behaves similarly to the classical Hotelling T^2 in Phase II, which is efficient since we are in the case of no Phase I outlier presence.

On the other hand, Figure 9 shows the results when the sample size increases to $n = 150$ and there are still no outliers in Phase I. For the three methods the probabilities of signal are similar, although the lowest values are those of T^2_{RMCD}, and when the dimension increases it can be noted a slightly better performance of T^2_{SR}.

When outliers are included in the Phase I simulated data, and they are near the center of the distribution ($\xi^2_I = 5$), while the sample size is low ($n = 50$), the performance of the robust alternatives is slightly better than the classical approach in low dimension $p = 2$, both for 10% and 20% contamination, as shown by Figures 10 and 11. But when dimension increases, T^2_{RMCD} starts to worsen its performance, specially when 20% outliers are considered. In the mean time, the probabilities of signal of T^2_{SR} remains higher than the classical T^2 and the other robust alternative, even in high dimension or high contamination.

Figures 12 and 13 show the behavior of the methods when the sample size increases to $n = 150$. In this case, although it is difficult to detect outliers when they are near the center of the data ($\xi_I^2 = 5$), the robust alternatives perform better than the classical approach, for all dimensions considered. However, when dimension increases, T^2_{RMCD} decreases the probability of signal. T^2_{SR} has a good behavior in this case because the increase in dimension or level of contamination does not greatly affect its performance.

Figures 14 and 15 show the performance when the non-centrality parameter in Phase I is large ($\xi_I^2 = 30$) and the sample size is $n = 50$. In this case, both robust alternatives substantially outperform the classical approach. But with the increase of dimension, T^2_{RMCD} decreases its performance, while T^2_{SR} remains robust, even in presence of a high level of contamination.

When the dimension is increased to $n = 150$ (Figures 16 and 17), the performance of the robust approaches is better than with lower sample size, but T^2_{RMCD} still worsen its performance with the increase of the dimension or the level of contamination and it is outperformed by T^2_{SR}.

The overall outcomes are that if the variables are independent, i.e. when the covariance structure is the identity matrix, the outliers and mean shifts have a negative impact on the classical Hotelling T^2 approach, and the use of robust alternatives is needed. The robust T^2_{RMCD} approach has good performance only when the dimension or the level of contamination are low, while the proposed approach T^2_{SR} has better performance in most cases, even when the dimension or the level of contamination increases.

4.2 Case B

Figures 18 - 20 in Appendix A, show the resulting probabilities of signal with respect to the Phase II non-centrality parameter. The results do not greatly change with different sample sizes, that is why we show only the results associated with $n = 150$. The same happens in case of Phase I contamination with the different values of π, we show the most significant results with $\pi = 20\%$. In case of outlier-free Phase I data, the three methods behave similarly, but when a certain percentage of outliers are introduced, the classical method T^2 and the robust alternative T^2_{RMCD} start worsening their performance. With the increase of the contamination level, the results are more or less the same, but their behavior gets worse with the increase of the dimension p. Meanwhile, the proposed method T^2_{SR} remains robust to the presence of outlying observations, even when dimension increases. In general, comparing with the results when the variables are considered independent (Case A), in presence of correlated features (Case B), a decrease in the probability of signal is noted for all methods, but mostly for the classical Hotelling T^2 and the T^2_{RMCD}.

4.3 Computational times

Table 3 shows the resulting computational times in seconds depending on the dimension p and the sample size n considered, for the three methods. The results

contain the average time between both simulation schemes, i.e. considering independent or correlated variables.

Table 3: Average computational times for both simulation schemes A and B.

p	n	T^2	T^2_{RMCD}	T^2_{SR}
2	50	1.0335	1.0950	0.0088
	150	1.6891	1.7344	0.0122
6	50	2.3467	2.9710	0.0410
	150	3.4989	3.0498	0.0628
10	50	6.2387	8.7894	0.4206
	150	7.8698	8.9657	0.9804
30	50	9.6511	9.1864	1.4589
	150	11.4671	12.7358	2.0013

The results show that when the dimension of the data or the sample size increase, the computations become more expensive, but the running time of T^2_{SR} is very low compared to the classical method and the other robust alternative, specially in high dimension, where T^2 and T^2_{RMCD} are 6 times slower than the proposed method T^2_{SR}.

5 Case study

To show the performance of the proposed method in practice, two real examples are selected. The first is also studied in Chenouri et al. [2009]. The data consists on four variables that measure proper alignment in the final assembly of automobiles, a characteristic that is known to influence the customer perception of quality. The alignment depends on four angles, namely front-right, and front-left wheel camber and caster. In this example, large volumes of data are available because all final inspections include the measurement of the four angles. Those automobiles having any of the characteristics out of specification, or out-of-control, should be reworked before shipment. To monitor the alignment, a multivariate Hotelling T^2 control chart could be used. However, the process is known to produce occasional outliers that should be reworked. Therefore, it is better to use a robust control chart procedure. In this example, the Phase I data consist of 186 vehicles produced during the time interval of January 2nd.

The first step of the proposed method is to compute the Phase I robust SR estimators of location and covariance matrix, using Equation 6:

$$\hat{\boldsymbol{\mu}}_{SR} = (0.256, 0.641, 4.978, 5.257)^t$$

$$\hat{\boldsymbol{\Sigma}}_{SR} = \begin{pmatrix} 0.028 & -0.101 & 0.012 & 0.019 \\ -0.101 & 0.025 & -0.018 & 0.021 \\ 0.012 & -0.018 & 0.036 & -0.009 \\ 0.019 & 0.021 & -0.009 & 0.070 \end{pmatrix}.$$

Next step is to compute the control limits using Equation 8 and Table 1. The 99% and 99.9% control limits are, respectively:

$$f_{4,0.990}(n) = 13.277 + \frac{2804}{n^{2.302}}$$

$$f_{4,0.999}(n) = 18.467 + \frac{1174000}{n^{2.88}}\,.$$

For $n = 186$, the resulting values of the control limits are $f_{4,0.990} = 13.293$ and $f_{4,0.999} = 18.809$. As we see, these control limits are near the asymptotic classical control limits 13.277 and 18.467, from the chi-square distribution with 4 degrees of freedom, because the sample size $n = 186$ is reasonably large. Also, they are slightly lower than the limits estimated by the approach based on RMCD estimators. Since the robust SR estimates are not influenced by outliers, we do not need to take any action, such as removing outlying points and reconstructing the control limits and the chart. Using the Phase I robust SR estimates, the Phase II control chart for the future observations can be obtained. Figure 1 shows the control chart for the 100 vehicles produced on January 12, the Phase II data. It can be seen that there is a mean shift in the process, assuming that the covariance structure remains the same, because a large number of T^2 values exceed the 99% and 99.9% control limits.

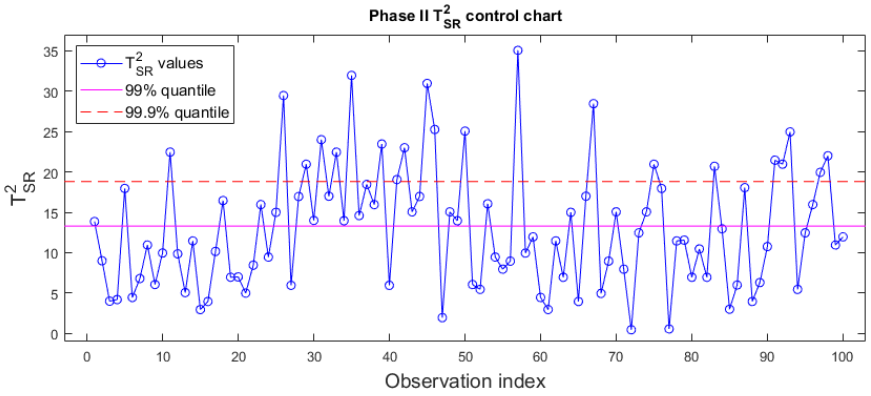

Figure 1: T^2_{SR} control chart for the 100 Phase II observations collected on January 12. The dashed and solid horizontal lines represent the control limits based on 99% and 99.9% quantiles, respectively.

More specifically, 45 Phase II observations exceed the 99% control limit, whereas 21 observations exceed the more restrictive 99.9% control limit. Table 4 contains the results. It also shows the results based on this dataset example using the classical T^2 control chart and the RMCD based control chart, studied in Chenouri et al. [2009]. With the RMCD based control chart, there is no outlier detection in the Phase I because of the reweighting step of the RMCD estimates, and 43 (99%) and 18 (99.9%) observations resulted to be out-of-

control in Phase II. The proposed method detected a few more out-of-control observations than the classical approach and the robust alternative. Since the Phase I sample is almost outlier free, the Phase II pattern is more or less similar for the three control charts.

Table 4: Phase I and Phase II out-of-control observations with aligment dataset.

Method	Phase I	Phase II (99%)	Phase II (99.9%)
Classical T^2	1	40	13
RMCD		43	18
SR		45	21

A second case, which is also used in Quesenberry [2001], Vargas [2003] and Alfaro and Ortega [2008], is studied. The original data consist of 11 quality variables measured on 30 products from a production process. In Vargas [2003], the authors consider the first two variables and use it to compare six different methods. On the other hand, Alfaro and Ortega [2008] use the same example to illustrate the performance of their proposed approach. Almost all methods proposed by the authors detect observation 2 as out-of-control in Phase I. Then, they propose to modify two of the observations (16 and 24) and turn them into outliers to test the performance of the methods in the task of detecting them in Phase I. Let us call this modified dataset that contains three outliers X_new. Only one of the six methods from Vargas [2003] and the proposed method from Alfaro and Ortega [2008] were able to detect in X_new the three outliers, through the computation of the control limits. All the other methods including the classical approach failed. Since their proposed methodologies are rather different than ours (because they perform the detection on Phase I), we propose to use X_new in Phase I and also generate $n = 100$ Phase II observations containing 10% and 20% of randomly generated outliers, to check the performance of the classical T^2, RMCD and SR. We repeated the simulation process $M = 1000$ times and the average True Positive Rates (FPR) and False Positive Rates (FPR) in Phase II are shown in Table 5.

Table 5: TPR and FPR with the randomly generated Phase II observations from the production dataset with 10% (case i) and 20% (case ii) out-of-control observations.

Method	TPR (i)	FPR (i)	TPR (ii)	FPR (ii)
Classical T^2	0	0	0	0
RMCD	0.71	0	0.54	0
SR	0.98	0	0.92	0

The classical T^2 does not detect any of the three outliers in Phase I (as reported by Vargas [2003] and Alfaro and Ortega [2008]), that is, these outliers are not detected due to the masking effect. This obviously worsens its performance in Phase II, failing to detect the out-of-control observations in both cases

(10% and 20%). For both RMCD and SR the procedure is different. We do not need to remove outlying points in Phase I as in the classical T^2. Using the Phase I corresponding robust estimates, the Phase II control chart for the future observations can be obtained. We compute the corresponding 99% control limit for both methods which is more sensitive for the FPR (with 99.9% control limits the FPR will be less or equal). The results show the minimum possible FPR for both methods, i.e. no observation is incorrectly signaled as out-of-control. Regarding the TPR, i.e. the rate of correctly signaled observations, it is lower with the RMCD compared to our proposed method SR, and it significantly reduces when the level of contamination increases, while the TPR of SR maintains high even with higher contamination. These results show an advantageous performance of our proposed method SR over the other two.

6 Conclusions

In this paper, a multivariate robust control chart is proposed as an alternative to the classical Hotelling T^2 approach. The computation of the chart is based on the robust shrinkage estimators introduced by Cabana et al. [2019] and further studied by Cabana et al. [2020]. The proposed control chart is obtained by replacing these robust estimators of location and covariance matrix in the definition of the classical approach. Since the estimators have good properties like robustness, approximate affine equivariance and high breakdown value, the results in this area is advantageous.

The quantiles for the proposed robust statistic are computed by means of Monte Carlo simulations considering a wide range of dimensions and sample sizes, extending the previous study of Chenouri et al. [2009] up to dimension $p = 30$. The estimated quantiles are modeled to approximate the control limits for any sample size, in order to ease the use of the proposed methodology in practice.

A simulation study is carried out, considering different Phase I scenarios in which the data is contaminated to see the effect in the proposed control chart computations and the capacity of detecting Phase II process shifts. The method is compared with the classical Hotelling T^2 approach and the robust alternative T^2_{RMCD}. Two different correlation schemes are considered: independent or highly correlated variables. Other parameters are also considered to explore different scenarios, such as the dimension, the sample size, the percentage of contamination, the magnitude of the contamination and where it happens, i.e. in all dimensions or in some percentage of the corresponding dimension. The results under these scenarios showed that the classical Hotelling T^2 is highly sensitive to the presence of Phase I outliers. Meanwhile, the robust alternative T^2_{RMCD} shows in general good results but only when the dimension p or the level of contamination π is low, but the performance worsens when p or π start to increase. On the other hand, the proposed approach T^2_{SR} outperforms the classical method and the other robust alternative in the vast majority of cases, showing special advantages in high dimension or when the level of

contamination increases, even in presence of high correlations. The proposed robust control chart performs robustly not only when the shift happens in all dimensions, but also when it happens in some percentage of the dimensions. Furthermore, T^2_{SR} shows additional advantages with respect to the competitive computational time.

Finally, the proposed method is illustrated using two case studies: one from the automotive industry consisting on four variables measured in 186 vehicles and the other from a production process consisting of two variables measured in 30 products. The robust SR control chart is computed for the Phase II data of the first real example, using the robust shrinkage reweighted estimators of location and covariance matrix computed from the Phase I data. The control limits based on the quantiles estimated in this paper suggest the presence of several outliers for both confidence levels considered 99% and 99.9%. The results are similar to the obtained by the other methods because the Phase I data was almost outlier-free, and the proposed T^2_{SR} control chart detects a few more observations as out-of-control. In the second example, additional out-of-control observations are included in Phase I to test the method's performance, as in Vargas [2003] and Alfaro and Ortega [2008]. Phase II observations are generated containing 10% and 20% of outliers, and the process is repeated 1000 times to obtain the TPR and FPR of the methods in the task of detecting the abnormal behavior. The classical method seems to be influenced by the masking effect and it fails to detect any of the Phase II out-of-control observations. The other two robust approaches show the lowest possible FPR. Although, the TPR is significantly lower for RMCD compared to SR, and it worsens with the increase of the contamination level, while the TPR of SR remains high in both cases. These results demonstrate the good performance of the robust SR control chart in real applications.

In this paper, the case of individual observations is considered, analogously as in Chenouri et al. [2009]. The case when the data is divided in subgroups with equal or different sample sizes is left for future research work to study if the proposed approach may have advantages in this scenario as well. On the other hand, as stated in Ledoit and Wolf [2004], when the dimension p is larger than the number n of observations available, the sample covariance matrix is not even invertible. When the ratio p/n is less than one but not negligible, the sample covariance matrix is invertible but numerically ill-conditioned,which means that inverting it amplifies the estimation error dramatically. However, in this case, their shrinkage estimator is asymptotically well-conditioned and more accurate than the classical estimator. This fact can motivate the further study of the proposed approach in the "large p, small n" scenario. Although, since this case was not the focus of the present paper, and the classical approach based on the sample covariance matrix and the method RMCD are not suitable in the case of $n < p$, we propose to explore the performance of the robust control chart based on shrinkage in this alternative scenario, in a future research.

7 Acknowledgments

The authors are grateful to the editor and the referees for their constructive and valuable comments that led to considerable improvement in this paper.

Appendices

A Figures

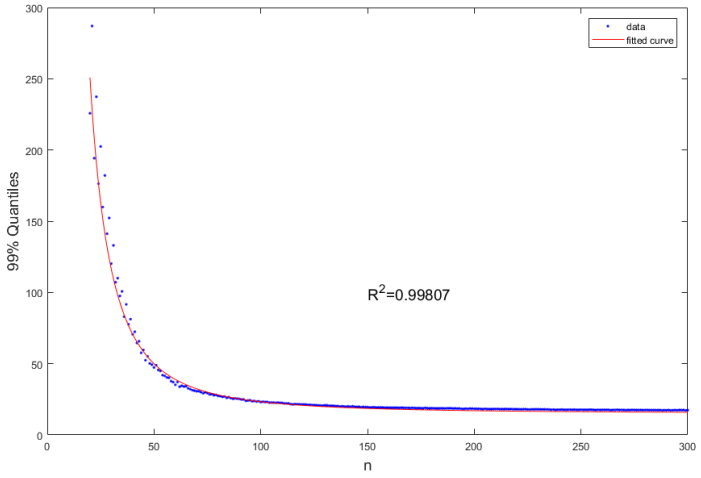

Figure 2: Simulated quantiles of T^2_{SR} and fitted curve for $p = 5$ and $\alpha = 0.01$

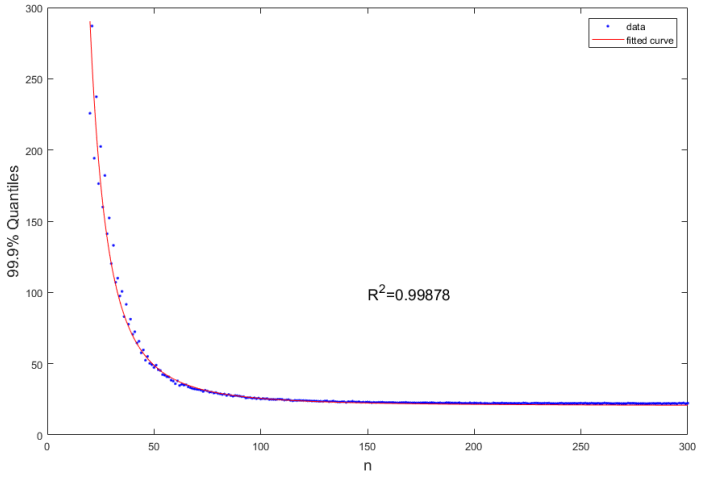

Figure 3: Simulated quantiles of T^2_{SR} and fitted curve for $p = 5$ and $\alpha = 0.001$

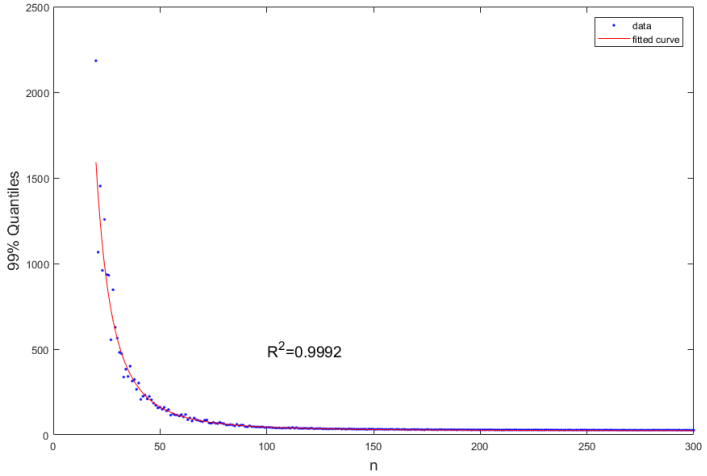

Figure 4: Simulated quantiles of T^2_{SR} and fitted curve for $p = 10$ and $\alpha = 0.01$

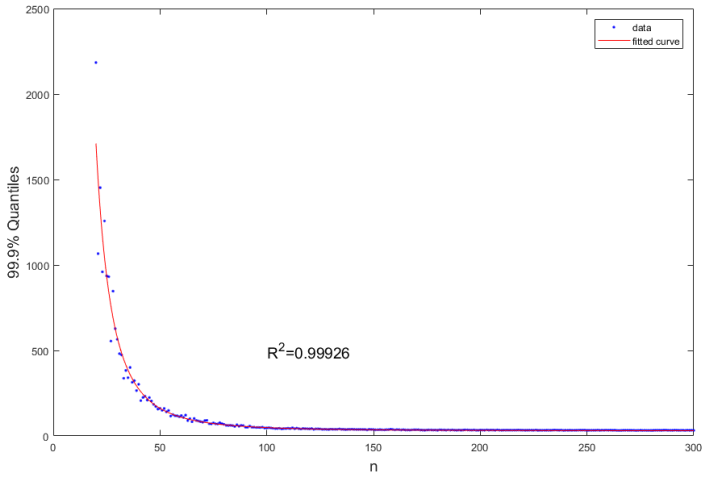

Figure 5: Simulated quantiles of T^2_{SR} and fitted curve for $p = 10$ and $\alpha = 0.001$

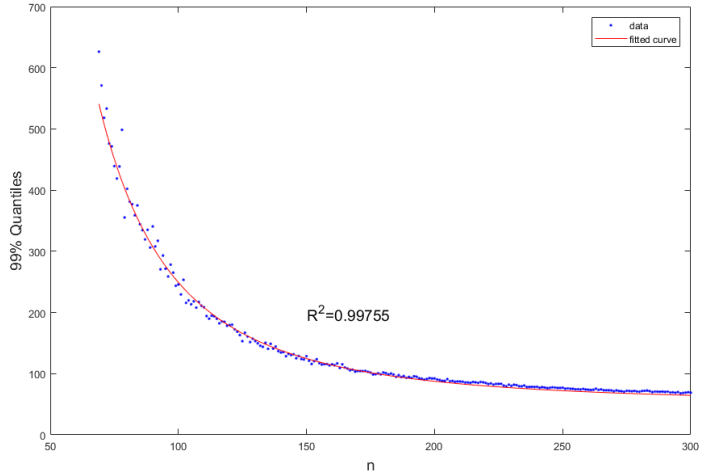

Figure 6: Simulated quantiles of T^2_{SR} and fitted curve for $p = 30$ and $\alpha = 0.01$

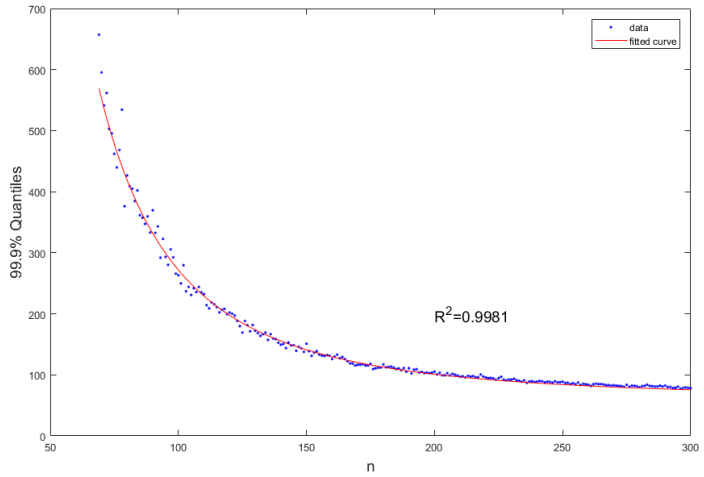

Figure 7: Simulated quantiles of T^2_{SR} and fitted curve for $p = 30$ and $\alpha = 0.001$

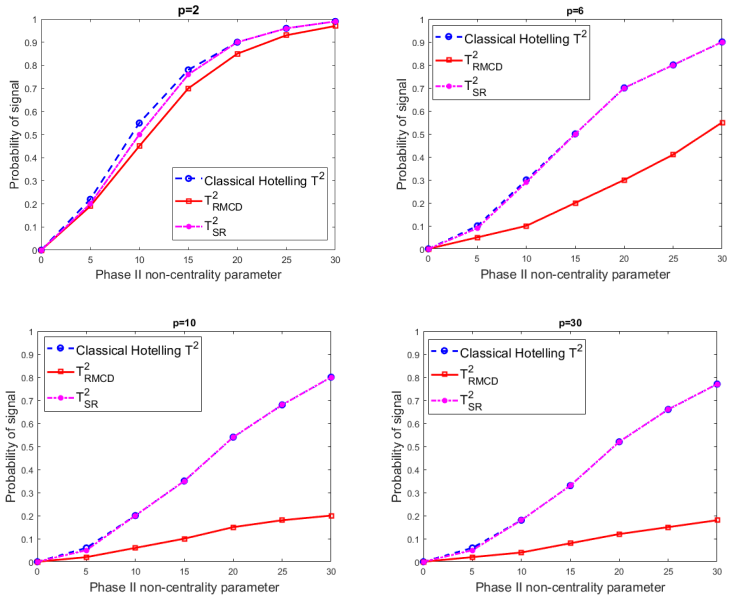

Figure 8: Probability of signal when Phase I data is of size $n = 50$ and outlier free (Combination 1 of Case A in Table 2).

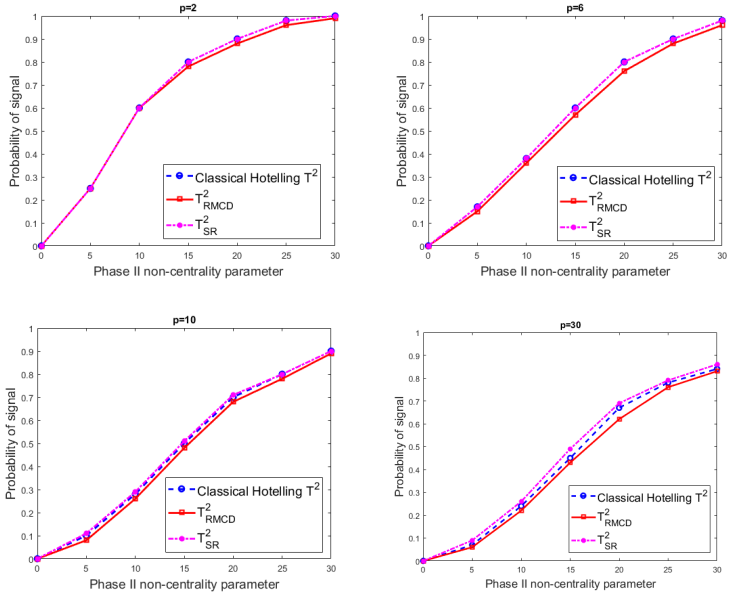

Figure 9: Probability of signal when Phase I data is of size $n = 150$ and outlier free (Combination 1 of Case A in Table 2).

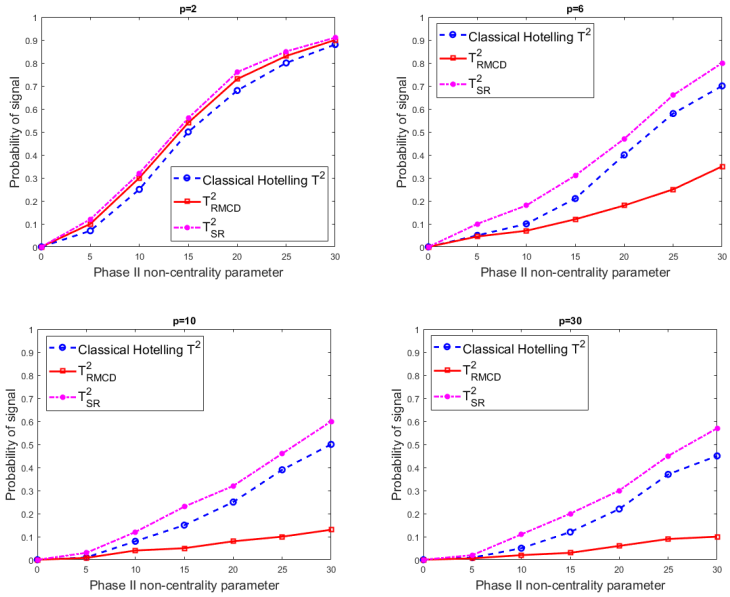

Figure 10: Probability of signal when Phase I data is of size $n = 50$ and has 10% of outliers with $\xi_I^2 = 5$ (Combination 2 of Case A in Table 2).

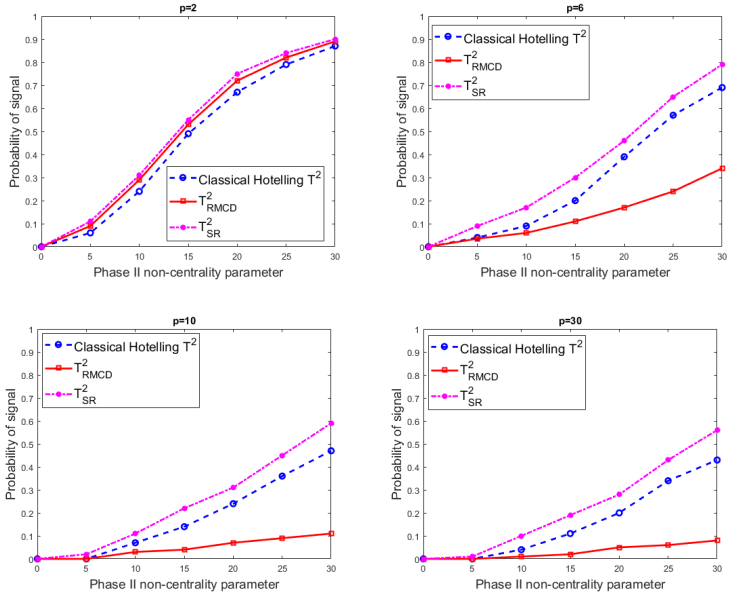

Figure 11: Probability of signal when Phase I data is of size $n = 50$ and has 20% of outliers with $\xi_I^2 = 5$ (Combination 4 of Case A in Table 2).

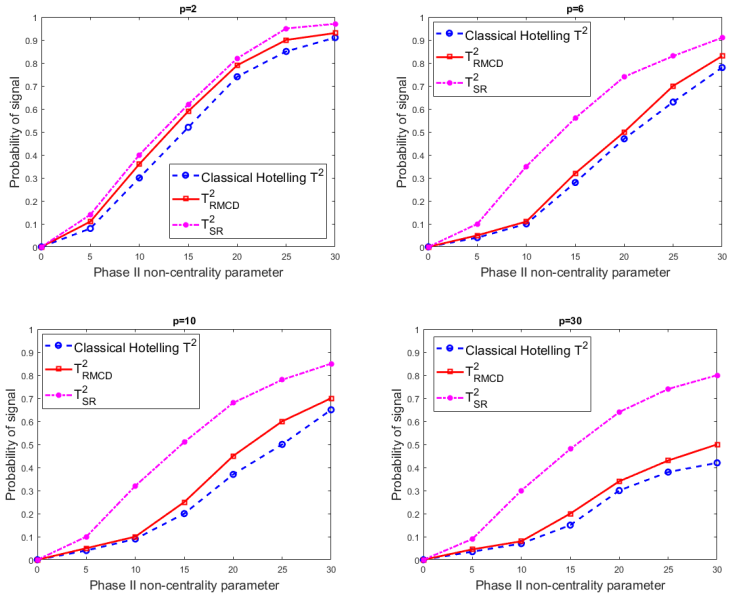

Figure 12: Probability of signal when Phase I data is of size $n = 150$ and has 10% of outliers with $\xi_I^2 = 5$ (Combination 2 of Case A in Table 2).

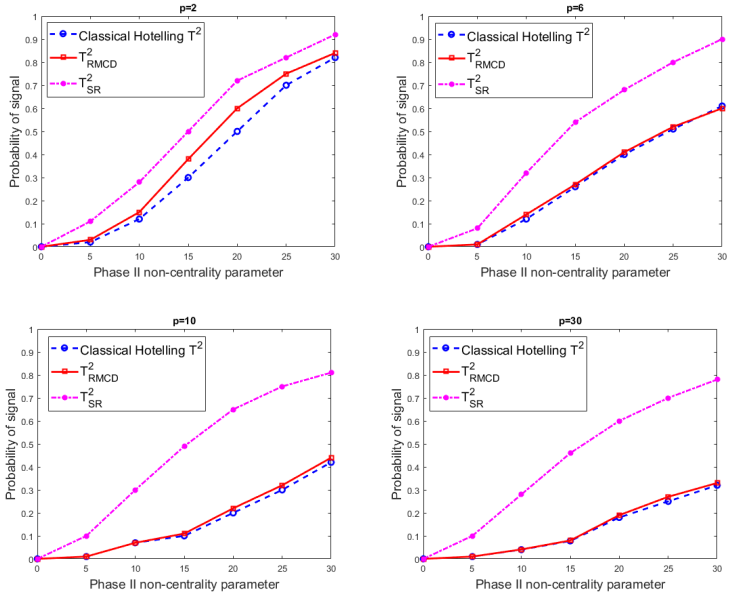

Figure 13: Probability of signal when Phase I data is of size $n = 150$ and has 20% of outliers with $\xi_I^2 = 5$ (Combination 4 of Case A in Table 2).

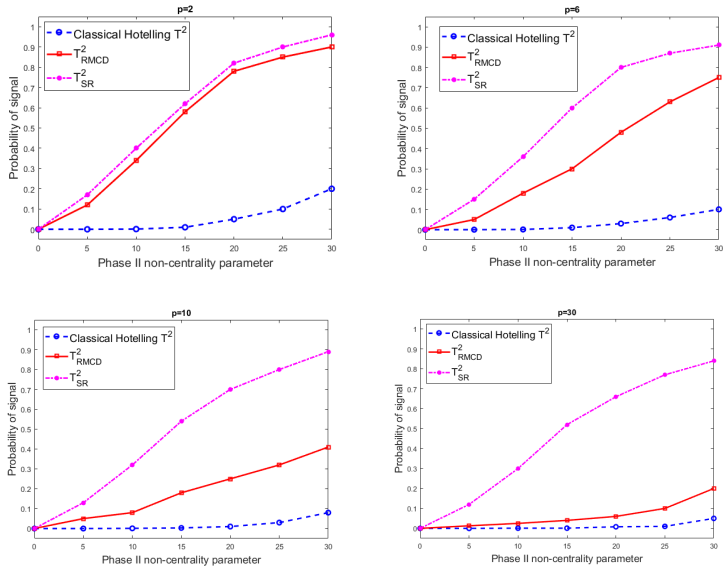

Figure 14: Probability of signal when Phase I data is of size $n = 50$ and has 10% of outliers with $\xi_I^2 = 30$ (Combination 3 of Case A in Table 2).

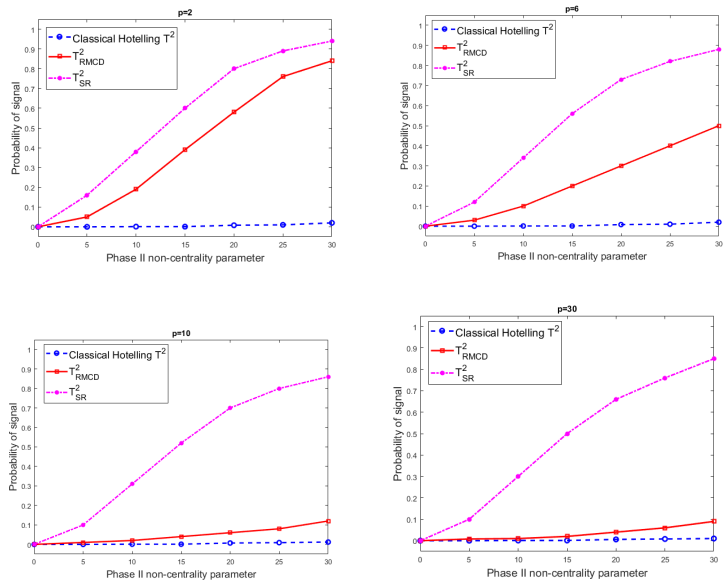

Figure 15: Probability of signal when Phase I data is of size $n = 50$ and has 20% of outliers with $\xi_I^2 = 30$ (Combination 5 of Case A in Table 2).

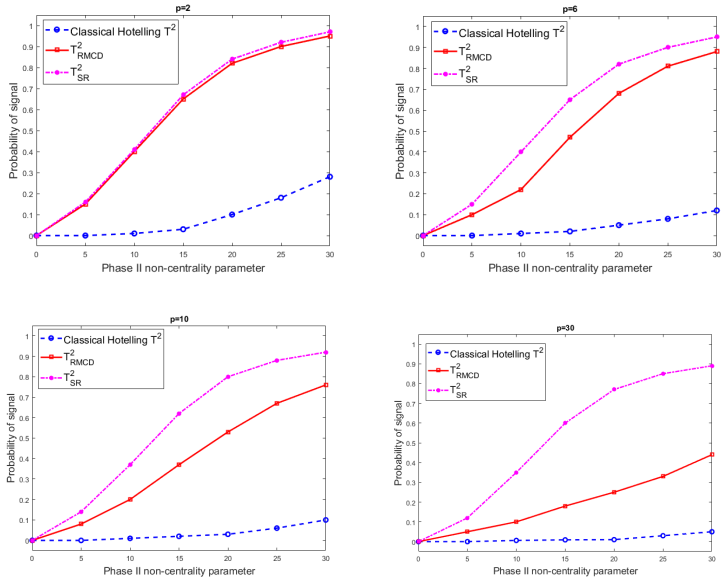

Figure 16: Probability of signal when Phase I data is of size $n = 150$ and has 10% of outliers with $\xi_I^2 = 30$ (Combination 3 of Case A in Table 2).

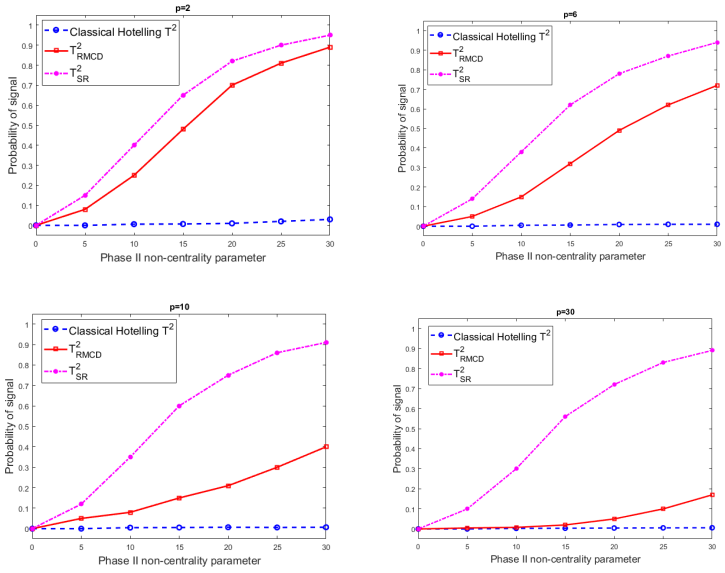

Figure 17: Probability of signal when Phase I data is of size $n = 150$ and has 20% of outliers with $\xi_I^2 = 30$ (Combination 5 of Case A in Table 2).

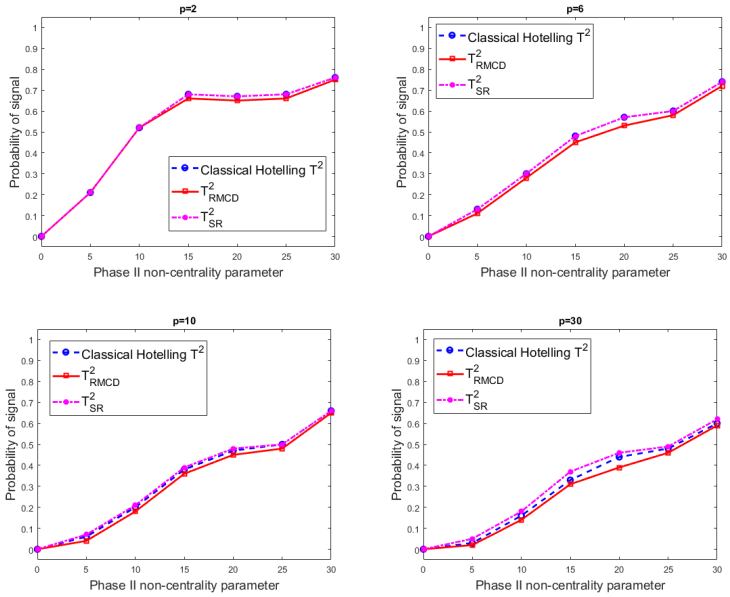

Figure 18: Probability of signal when Phase I data is of size $n = 150$ and is outlier-free (Combination 1 of Case B in Table 2).

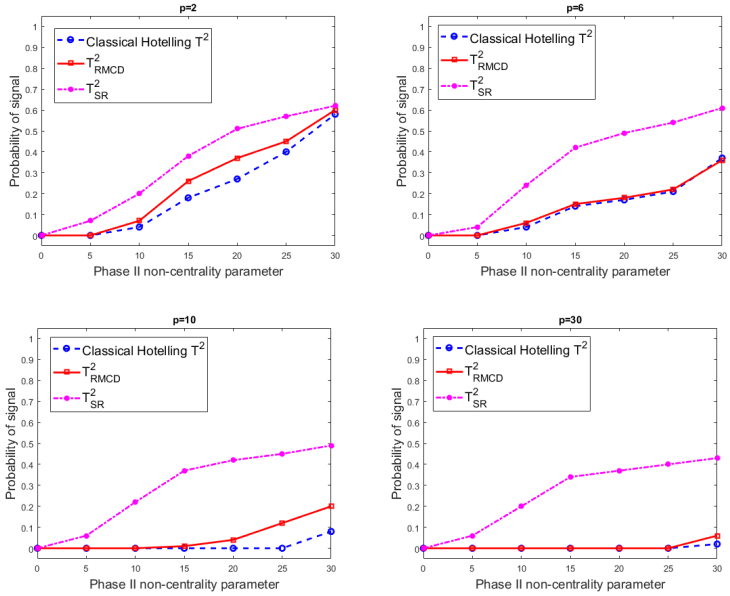

Figure 19: Probability of signal when Phase I data is of size $n = 150$ and has 20% of outliers with $\xi_I^2 = 5$ (Combination 4 of Case B in Table 2).

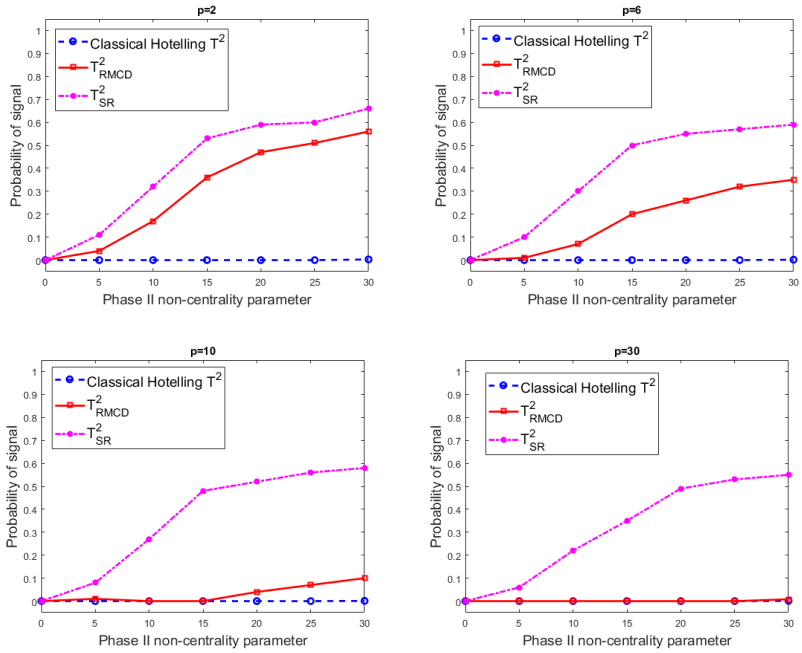

Figure 20: Probability of signal when Phase I data is of size $n = 150$ and has 20% of outliers with $\xi_I^2 = 30$ (Combination 5 of Case B in Table 2).

References

J. Alfaro and J. F. Ortega. A comparison of robust alternatives to hotelling's t 2 control chart. *Journal of Applied Statistics*, 36(12):1385–1396, 2009.

J. L. Alfaro and J. F. Ortega. A robust alternative to hotelling's t2 control chart using trimmed estimators. *Quality and Reliability Engineering International*, 24(5):601–611, 2008.

A. Bose. Estimating the asymptotic dispersion of the l1 median. *Annals of the Institute of Statistical Mathematics*, 47(2):267–271, 1995.

A. Bose and P. Chaudhuri. On the dispersion of multivariate median. *Annals of the Institute of Statistical Mathematics*, 45(3):541–550, 1993.

E. Cabana, R. E. Lillo, and H. Laniado. Multivariate outlier detection based on a robust mahalanobis distance with shrinkage estimators. *Statistical Papers*, pages 1–27, 2019.

E. Cabana, R. E. Lillo, and H. Laniado. Robust regression based on shrinkage with application to living environment deprivation. *Stochastic Environmental Research and Risk Assessment*, 34(2):293–310, 2020.

Y. Chen, A. Wiesel, and A. O. Hero. Robust shrinkage estimation of high-dimensional covariance matrices. *IEEE Transactions on Signal Processing*, 59(9):4097–4107, 2011.

S. Chenouri, S. H. Steiner, and A. M. Variyath. A multivariate robust control chart for individual observations. *Journal of Quality Technology*, 41(3):259–271, 2009.

W. E. Company. *Statistical quality control handbook*. Western Electric Company, 1956.

R. Couillet and M. McKay. Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators. *Journal of Multivariate Analysis*, 131:99–120, 2014.

V. DeMiguel, A. Martin-Utrera, and F. J. Nogales. Size matters: Optimal calibration of shrinkage estimators for portfolio selection. *Journal of Banking & Finance*, 37(8):3018–3034, 2013.

M. Falk. On mad and comedians. *Annals of the Institute of Statistical Mathematics*, 49(4):615–644, 1997.

D. M. Hawkins and D. J. Olive. Inconsistency of Resampling Algorithms for High-Breakdown Regression Estimators and a New Algorithm. *Journal of the American Statistical Association*, 97(457):136–148, 2002.

H. Hotelling. Multivariate quality control. *Techniques of statistical analysis*, 1947.

W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379, 1961.

W. A. Jensen, J. B. Birch, and W. H. Woodall. High breakdown estimation methods for phase i multivariate control charts. *Quality and Reliability Engineering International*, 23(5):615–629, 2007.

O. Ledoit and M. Wolf. Honey, i shrunk the sample covariance matrix. *UPF economics and business working paper*, (691), 2003a.

O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5):603–621, 2003b.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

A. M. Leroy and P. J. Rousseeuw. *Robust regression and outlier detection*, volume 589. John Wiley & Sons, 1987.

H. P. Lopuhaa and P. J. Rousseeuw. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, 19(1):229–248, 1991.

R. L. Mason and J. C. Young. *Multivariate Statistical Process Control with Industrial Applications*, volume 9. American Statistical Asociation - Society for Industrial and Applied Mathematics, 2002.

A. Mitra. *Fundamentals of Quality Control and Improvement*. John Wiley & Sons, Inc, 3rd edition, 2008.

D. C. Montgomery. Introduction to statistical quality control, john wiley&sons. *Inc., New York*, pages 313–343, 1997.

J. Möttönen, K. Nordhausen, H. Oja, et al. Asymptotic theory of the spatial median. In *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková*, pages 182–193. Institute of Mathematical Statistics, 2010.

L. S. Nelson. The shewhart control chart—tests for special causes. *Journal of Quality Technology*, 16(4):237–239, 1984.

H. Oja. *Multivariate nonparametric methods with R: an approach based on spatial signs and ranks*. Springer Science & Business Media, 2010.

C. P. Quesenberry. The multivariate short-run snapshot q chart. *Quality Engineering*, 13(4):679–683, 2001.

P. J. Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8(283-297):37, 1985.

P. J. Rousseeuw and B. C. Van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical association*, 85(411):633–639, 1990.

W. Shewhart. Contributions of statistics to the science of engineering'in university of pennsylvania bicentennial conference. volume on fluid mechanics and statistical methods. pages 97–124, 1941.

A. Steland. Shrinkage for covariance estimation: asymptotics, confidence intervals, bounds and applications in sensor monitoring and finance. *Statistical Papers*, pages 1–22, 2018.

A. J. Stromberg, O. Hössjer, and D. M. Hawkins. The Least Trimmed Differences Regression Estimator and Alternatives. *Journal of the American Statistical Association*, 95(451):853–864, 2000.

J. H. Sullivan and W. H. Woodall. A comparison of multivariate control charts for individual observations. *Journal of Quality Technology*, 28(4):398–408, 1996.

J. H. Sullivan and W. H. Woodall. Adapting control charts for the preliminary analysis of multivariate observations. *Communications in Statistics-Simulation and Computation*, 27(4):953–979, 1998.

N. D. Tracy, J. C. Young, and R. L. Mason. Multivariate control charts for individual observations. *Journal of quality technology*, 24(2):88–95, 1992.

Y. Vardi and C.-H. Zhang. The multivariate l1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426, 2000.

N. J. A. Vargas. Robust estimation in multivariate control charts for individual observations. *Journal of Quality Technology*, 35(4):367–376, 2003.

S. S. Wilks. Mathematical statistics. *John Wiley and Sons, New York*, 260, 1962.

G. Willems, G. Pison, P. Rousseeuw, and S. Van Aelst. A robust hotelling test. *Metrika*, 55(1-2):125–138, 2002.