A Robust Version of Hotelling's T^2 Control Chart for Retrospective Location Analysis of Individuals Using BACON Estimators

by

Richard C. Bell, Jr.

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
December 7, 2011

Keywords: multivariate statistical process control, phase I, preliminary, in-control reference sample, breakdown point, outliers

Approved by

Nedret Billor, Chair, Associate Professor of Mathematics and Statistics
L. Allison Jones-Farmer, Co-chair, Associate Professor of Management
Asheber Abebe, Associate Professor of Mathematics and Statistics

Abstract

Hotelling's T^2 chart is commonly used for Phase I analysis of individual multivariate normally distributed data. However, the presence of only a few outliers can significantly distort classical estimates of location and scale, thus rendering the resulting analysis ineffective. This poses a significant problem for the Hotelling's T^2 chart practitioner because the desired output of a Phase I analysis is an outlier-free reference sample which can be used to estimate control limits for prospectively monitoring a process in Phase II. Careful selection of a robust parameter estimation method is therefore critical when the initial reference sample is suspected to contain multiple outliers.

The purpose of this research is to propose a version of Hotelling's T^2 chart that uses the blocked adaptive computationally efficient outlier nominators (BACON) algorithm to robustly estimate location and scale parameters in Phase I. The proposed control chart, which assumes individual multivariate normally distributed data with constant covariance, is designed to detect both individual outliers and sustained mean shifts. Using Monte Carlo simulation, the proposed method is compared to Hotelling's T^2 chart using classical estimators as well as robust estimators such as the minimum volume ellipsoid (MVE), minimum covariance determinant (MCD), and clustering methods. Although the BACON-based version of Hotelling's T^2 chart turned out to be less powerful than expected, it is significantly better than the classical approach and offers some improvement over existing robust methods at a fraction of the computational expense.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ARL	average run length
BACON	blocked adaptive computationally efficient outlier nominators
CL	center line
EAP	empirical alarm probability
FAP	false alarm probability
HT2	Hotelling's T^2
IC	in control
LCL	lower control limit
MCD	minimum covariance determinant
MCUSUM	multivariate cumulative sum
MEWMA	multivariate exponentially weighted moving average
MVE	minimum volume ellipsoid
OC	out of control
RBP	replacement breakdown point
RL	run length
RMCD	reweighted minimum covariance determinant
RMVE	reweighted minimum volume ellipsoid
UCL	upper control limit

1 Introduction and Literature Review

The first multivariate quality control chart is attributed to Harold Hotelling (1947), who created the T^2 chart to monitor bombsight data during World War II. Since its introduction, many variations and refinements of Hotelling's T^2 chart have been proposed, and it remains the most familiar multivariate quality control chart in existence today [Montgomery (2005, p. 491)]. This research seeks to further broaden the appeal of Hotelling's T^2 chart for individual multivariate normally distributed data in a Phase I setting by using Billor, Hadi, and Velleman's (2000) blocked adaptive computationally efficient outlier nominators (BACON) method of robust parameter estimation to improve the T^2 statistic's robustness to outliers.

1.1 Special Considerations in Phase I Control Charting

A control charting application is typically divided into two distinct phases. In Phase I, when little is known about a process being studied, the objective is to identify an in-control (IC) reference sample. This involves retrospective analysis of a historical data set in order to eliminate any data points that do not accurately represent the routine operation of the process. The resulting data are described as in control because it is believed that all remaining variability in the process is inherent to the process itself and not due to assignable causes. Upon completion of Phase I, the in-control reference sample is used to establish control limits for Phase II, the monitoring stage of a control charting application. In Phase II, newly observed data points are successively compared to the control limits to identify significant departures from the in-control

state. Should an observation fall outside the control limits, a search for an assignable cause is immediately undertaken. If the change in process behavior can be linked to special causes or external factors, the process is deemed out of control (OC) and corrective action is implemented to fix the problem.

Prior to conducting any analysis in a control charting scenario, it is usually assumed that the unedited reference sample may contain out-of-control points and the control limits are unknown. The challenging nature of a Phase I analysis under these conditions has been recognized since the earliest days of statistical process control. Shewhart (1939, p. 76) said, "In the majority of practical instances, the most difficult job of all is to choose the sample that is to be used as the basis for establishing the tolerance range. If one chooses such a sample without respect to the assignable causes present, it is practically impossible to establish a tolerance range that is not subject to a huge error."

Phase I control charts are designed with the goal of achieving a specified overall in-control false alarm probability (FAP), defined as the probability of one or more observations plotting outside the control limits in the absence of assignable causes. Phase I usually involves iteratively comparing the reference sample to trial control limits (corresponding to the desired overall in-control FAP) estimated from the sample. At each iteration of a Phase I analysis, an out-of-control point is eliminated from the reference sample if an assignable cause is identified, and trial control limits are updated excluding the out-of-control point. This iterative process continues until all points in the reference sample are in control.

Phase I analysis requires careful consideration when it involves methods such as Hotelling's T^2 chart that compute independent control chart statistics consisting of individual observations. Provided the observations originate from random sampling, the control chart

statistics are independent of one another. However, because the control limits in Phase I are estimated from the reference sample itself, the control limits are dependent on each sample point included in their calculation. Thus, *simultaneous comparisons* of chart statistics to control limits in Phase I are statistically dependent despite the control chart statistics themselves being independent. These dependencies often make it difficult to correctly determine the overall in-control FAP for a Phase I analysis.

Phase II, on the other hand, consists of comparing *new* observations (in the form of a control chart statistic) to the control limits previously established in Phase I. Because the control limits in Phase II are fixed through conditioning, *successive comparisons* of chart statistics to control limits are independent provided the chart statistics are independent of one another as in the case of Hotelling's T^2 chart. This is in contrast to multivariate exponentially weighted moving average (MEWMA) or multivariate cumulative sum (MCUSUM) charts whose chart statistics include past observations and are therefore naturally dependent.

Chart performance in Phase II is often measured using moments of the run length (RL) distribution. The RL is the number of observations until an out-of-control signal is observed. If the comparisons of the chart statistics to the control limits are independent, the RL is a geometric random variable. The expected value of the RL is equal to $1/\alpha$, where α is equal to the probability that a single chart statistic plots outside the control limits in the absence of assignable causes. The expected value of the RL is known as the average run length (ARL) and is commonly used to describe control chart performance in Phase II.

1.2 Hotelling's T^2 Control Chart for Individuals -- the Classical Approach

Hotelling's T^2 control chart may be used to detect outliers in a *p*-dimensional multivariate process, where each observation in the form of a *p*-vector $\boldsymbol{X}_i = (X_1, ..., X_p)$ is assumed to come from a multivariate normal distribution when the process is in-control. More specifically, each $\boldsymbol{X}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is the *p*-dimensional mean vector that defines the location of the process and $\boldsymbol{\Sigma}$ is the positive definite *p* x *p* covariance matrix that specifies the dispersion of the process. Hotelling's T^2 statistic is calculated for each $\boldsymbol{X}_i$ as

$$T_i^2 = \left(\boldsymbol{X}_i - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{X}_i - \boldsymbol{\mu}\right), \tag{1.2.1}$$

and subsequently compared to either a Phase I or Phase II upper control limit (UCL).

In a Phase II control charting application when $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are known, the UCL for Hotelling's T^2 chart is given by Montgomery (2005, p. 501) as

$$\text{UCL} = \chi^2_{\alpha,p}, \text{ where } \alpha = \text{the desired in-control FAP.} \tag{1.2.2}$$

When $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown, as is usually the case in practice, they are typically replaced by the classical sample mean vector and sample covariance matrix estimated from an in-control reference sample consisting of *n* independent observations. The classical estimators are defined as follows:

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \quad \text{and} \quad \boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n} \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)'. \tag{1.2.3}$$

Using these unbiased estimators, Hotelling's T^2 statistic becomes

$$T_i^2 = \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)' \boldsymbol{S}^{-1} \left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right) \tag{1.2.4}$$

and according to Montgomery (2005, p. 501) has Phase II upper control limit

$$\text{UCL} = \frac{p(n+1)(n-1)}{n^2 - np} F_{\alpha,p,n-p}, \text{ where } \alpha = \text{the desired in-control FAP.} \quad (1.2.5)$$

If an in-control reference sample is not available, a Phase I analysis must first be conducted using Hotelling's T^2 statistic in Equation (1.2.4) and the following UCL specified by Tracy, Young, and Mason (1992):

$$\text{UCL} = \frac{(n-1)^2}{n} \beta_{\alpha,p/2,(n-p-1)/2}, \text{ where } \alpha = \text{the desired in-control FAP.} \quad (1.2.6)$$

Note that in Phase I, α represents the desired in-control FAP for each observation. In order to set α to achieve a desired *overall* in-control FAP for all *n* observations in a reference data set, the following relation must be used:

$$\alpha = 1 - \left(1 - \alpha_{overall}\right)^{1/n}. \quad (1.2.7)$$

For example, for a reference sample consisting of $n = 50$ observations and a desired *overall* in-control FAP of 0.05, $\alpha = 1 - \left(1 - 0.05\right)^{1/50} = 0.001025$ would be used in Equation (1.2.6) to determine the Phase I UCL. If the purpose of Hotelling's T^2 chart is solely to identify location shifts, the lower control limit (LCL) in both Phase I and Phase II is often defined as zero or not specified at all. This is because location shifts in all directions result in increasingly positive T^2 values, so an LCL is not necessary.

As noted by Vargas (2003), Hotelling's T^2 statistic using classical estimators in Phase I is effective in detecting a single moderately sized outlier. However, its inability to detect multiple outliers has been well documented by Vargas (2003), Jobe and Pokojovy (2009), and Yanez, Gonzalez, and Vargas (2010), and its poor performance in detecting sustained shifts in the mean vector has been demonstrated by Sullivan and Woodall (1996) and Vargas (2003). This is because the presence of even a single arbitrarily large outlier or a few moderately sized outliers

can significantly contaminate the parameter estimates $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$, thus rendering the T^2 statistic ineffective. This is precisely the problem this research seeks to address.

1.3 Previous Attempts to Improve the Robustness of Hotelling's T^2 Chart

In recent years, improving the robustness of Hotelling's T^2 chart for individual multivariate normally distributed data in Phase I has garnered much interest in multivariate quality control research. Many proposed methods have been successful under certain conditions, but none have proven to be universally superior. Thus, there is still room for improvement. Before detailing the BACON-based robust version of Hotelling's T^2 chart proposed by this research, a brief summary of other robust versions of Hotelling's T^2 chart for individual multivariate normally distributed data will be provided.

In one of the earliest attempts to improve the robustness of Hotelling's T^2 control chart, Sullivan and Woodall (1996) proposed using vector differences between successive observations to estimate the in-control covariance matrix of a process, and showed that this method used in conjunction with Hotelling's T^2 statistic results in enhanced detection ability of step (sudden) and ramp (gradual) shifts in the mean during retrospective analysis of a data set. Later, Vargas (2003) evaluated the performance of five different types of robust estimators for use with Hotelling's T^2 chart, including the minimum volume ellipsoid (MVE) estimators of Rousseeuw (1984) and Rousseeuw and Van Zomeren (1990), the minimum covariance determinant (MCD) method of Rousseeuw (1984) and Rousseeuw and Van Driessen (1999), a trimming approach based on Mahalanobis distance, the aforementioned method of vector differences proposed by Sullivan and Woodall (1996), and an outlier detection algorithm also proposed by Sullivan and Woodall (1996).

Vargas (2003) ultimately recommended the MVE estimator for detecting multiple outliers and the Sullivan and Woodall (1996) successive differences estimator for identifying sustained shifts in the mean vector. Jensen, Birch, and Woodall (2007) further detailed the advantages of the MVE and MCD methods as robust estimators and provided a detailed analysis of when to use each type of estimator with the T^2 statistic in a Phase I control chart setting. Alfaro and Ortega (2008) proposed trimming each variable to obtain robust estimates for the location vector and covariance matrix, and then using those estimates in Hotelling's T^2 chart with the Phase I UCL given in Equation (1.2.6) to provide enhanced outlier detection. The method of Alfaro and Ortega (2008) demonstrated improvement over Hotelling's T^2 chart using classical estimators, but no other performance comparisons were offered.

In one of the most comprehensive studies performed, Jobe and Pokojovy (2009) developed a computationally intensive two-step method of identifying the largest bulk of similar multivariate data from a time-ordered sequence of individual points, and used the estimated mean vector and covariance matrix from this bulk in the T^2 statistic with empirical control limits. The authors compared the performance of Hotelling's T^2 chart using their method, the classical method of parameter estimation, and the robust methods analyzed by Vargas (2003) and Jensen et al. (2007), showing that their method resulted in improved performance in detecting outliers as well as sustained shifts in location. Based on these findings, the results of Jobe and Pokojovy (2009) will be used as the standard of comparison for the BACON-based version of Hotelling's T^2 chart proposed by this research.

The most recent attempts to improve the robustness of Hotelling's T^2 chart in Phase I include Oyeyemi and Ipinyomi's (2010) proposal to robustly estimate the covariance matrix by identifying a subset of data that meets specified optimality criteria, and then iteratively

expanding the subset to a predetermined size. The method was shown to outperform the MVE and MCD methods in a limited number of cases, but only bivariate samples of size $m = 30$ were considered. Yanez, Gonzalez, and Vargas (2010) proposed a T^2 chart using biweight S estimators for location and scatter in conjunction with simulated limits, showing that it outperforms Hotelling's T^2 chart with MVE estimators for small samples.

Other authors of robust Hotelling's T^2 charts for individual multivariate normally distributed data focused their performance comparisons on Phase II rather than Phase I. Chenouri and Steiner (2009) proposed a robust Phase II Hotelling's T^2 chart with simulated limits based on reweighted MCD (RMCD) estimators, as defined by Willems, Pison, Rousseeuw, and Van Alest (2002), obtained directly from an unedited reference sample. The authors show that their version of Hotelling's T^2 chart outperforms Hotelling's T^2 chart using classical estimators, MVE estimators, and MCD estimators under certain conditions in Phase II. Chenouri and Variyath (2011) built upon the work of Chenouri and Steiner (2009) by comparing Hotelling's T^2 chart using RMCD estimators to Hotelling's T^2 chart using reweighted MVE (RMVE) and S estimators in Phase II, again favoring the RMCD estimators for location and scatter in most scenarios evaluated. Mohammadi, Midi, Arasan, and Al-Talib (2011) also explored the merits of using RMCD and RMVE estimators in a Phase II Hotelling's T^2 chart, recommending the RMVE method for small reference samples and the RMCD method for large reference samples. Assessment of a robust version of Hotelling's T^2 control chart exclusively in terms of Phase II performance inherently assumes that a practitioner desires to use Hotelling's T^2 chart in Phase II, which may or may not be the case in reality. In order to allow a practitioner the flexibility to choose another proven Phase II method (e.g. the MEWMA or MCUSUM chart) after outlier

removal is completed in Phase I, this research will focus on Hotelling's T^2 chart performance in Phase I only.

1.4 Organization of Thesis

The remainder of this document is dedicated to the detailed development and application of a BACON-based robust version of Hotelling's T^2 control chart for individual multivariate normally distributed data in Phase I. Chapter 2 explores the properties of various robust parameter estimation methods including BACON, and addresses the design of a BACON-based Hotelling's T^2 control chart. Chapter 3 provides the simulation plan for assessing the BACON-based Hotelling's T^2 control chart's in- and out-of-control performance. Chapter 4 contains the results of the simulation study and comparisons to several existing robust Hotelling's T^2 control charts. This thesis concludes in Chapter 5 with a synopsis of research conducted, recommendations to practitioners, and discussion of areas in need of further investigation.

2 Hotelling's T^2 Chart Using Robust Estimates of Location and Scatter

Numerous robust parameter estimation methods have been used in control chart research, including the MVE, MCD, and several other methods mentioned in Chapter 1. The BACON method was shown by Billor et al. (2000) to perform similarly to well known robust parameter estimation methods such as the MVE and MCD methods without the extreme computational burden, making it a seemingly ideal candidate for improving the robustness of Hotelling's T^2 chart. The BACON method demonstrates excellent balance between computational complexity and robustness to outliers while also satisfying several other important statistical properties of robust parameter estimation methods.

2.1 Desirable Properties of Robust Estimators

The performance of an estimator is commonly described by its finite-sample *replacement breakdown point* (RBP). First defined by Donoho and Huber (1983), the RBP is the minimum fraction of a sample that must be replaced by outliers in order to completely ruin an estimate, so a low RBP indicates nonrobustness and a high RBP signifies robustness to outliers. Precise definitions of RBPs for both location and scatter estimators are adapted from Donoho and Huber (1983) and Lopuhaa and Rousseeuw (1991). Let $\mathrm{X}_n = \{\boldsymbol{X}_1, ..., \boldsymbol{X}_n\}$ be a random sample of size *n* in R^p. The RBP of a location estimator *T* at X_n, or the smallest fraction *k/n* of outliers that can take the resulting estimate beyond any bound, is defined as

$$RBP(T;\mathrm{X}_n) = \min\left\{ k/n : \sup_{\mathrm{X}_{n,k}} \left\| T(\mathrm{X}_n) - T(\mathrm{X}_{n,k}) \right\| = \infty \right\}, \tag{2.1.1}$$

where $\mathrm{X}_{n,k}$ is a contaminated sample found by replacing *k* points of X_n with arbitrary values. The RBP of a scatter estimator *C* at X_n, or the smallest fraction *k/n* of outliers that can drive either the largest eigenvalue of the resulting estimate to infinity or the smallest eigenvalue of the resulting estimate to zero, is defined as

$$RBP(C;\mathrm{X}_n) = \min\left\{ k/n : \sup_{\mathrm{X}_{n,k}} M\left(C(\mathrm{X}_n), C(\mathrm{X}_{n,k})\right) = \infty \right\}, \tag{2.1.2}$$

where $\mathrm{X}_{n,k}$ is defined as before, $M(\boldsymbol{A},\boldsymbol{B}) = \max\left\{ \left|\lambda_1(\boldsymbol{A}) - \lambda_1(\boldsymbol{B})\right|, \left|\lambda_p(\boldsymbol{A})^{-1} - \lambda_p(\boldsymbol{B})^{-1}\right| \right\}$, and $\lambda_1(\boldsymbol{A}) \geq \cdots \geq \lambda_p(\boldsymbol{A})$ are the ordered eigenvalues of the matrix $\boldsymbol{A}$.

To illustrate the idea of an RBP, consider a sample of size *n* in R^1 and two common location estimators: the sample mean and the sample median. The sample mean has an RBP of only 1/*n* because a single outlier could move the sample mean to infinity, so it is considered a nonrobust location estimator. In contrast, the sample median has the highest possible RBP of 1/2 because 1/2 of the sample would have to be contaminated with outliers in order to effect a corresponding shift in the sample median. Consequently, the sample median is the preferred location estimator in R^1 from a robustness standpoint.

In addition to having a high RBP, a location or scatter estimator should also be *affine equivariant*. From Lopuhaa and Rousseeuw (1991), a location estimator *T* is affine equivariant if $T(\boldsymbol{AX}+\boldsymbol{b}) = \boldsymbol{A}T(\boldsymbol{X})+\boldsymbol{b}$ for any *p*-vector $\boldsymbol{b}$ and any *p* x *p* nonsingular matrix $\boldsymbol{A}$, and a positive definite scatter estimator *C* is said to be affine equivariant if $C(\boldsymbol{AX}+\boldsymbol{b}) = \boldsymbol{A}C(\boldsymbol{X})\boldsymbol{A}^T$ for any *p*-

vector $\boldsymbol{b}$ and any p x p nonsingular matrix $\boldsymbol{A}$. Affine equivariance means that an estimator does not depend on the location, scale, or orientation of the data.

2.2 The MVE and MCD Methods of Robust Parameter Estimation

Rousseeuw's (1984) MVE method is an affine equivariant, computationally complex, robust parameter estimation technique. The MVE method finds the ellipsoid of minimum volume that covers a subset of at least h points, and uses the geometrical center of the ellipsoid as the location estimator and the matrix defining the ellipsoid itself (multiplied by a constant) as the covariance matrix estimator. Lopuhaa and Rousseeuw (1991) showed that the integer value of $h = (n+p+1)/2$ provides the highest possible RBP of $[(n-p+1)/2]/n$, which converges to 50% as $n \to \infty$.

Due to the computational complexity of Rousseeuw's (1984) original MVE method, Rousseeuw and Leroy (1987) proposed an alternative method that approximates MVE estimates using a subsampling algorithm. With the subsampling method, a fixed number of subsets are first drawn from a data set. Rousseeuw and Leroy (1987, p. 199) recommend at least 500 subsets for small data sets in low dimensions and even more subsets for larger n and p. Next, the sample mean vector and sample covariance matrix are calculated for each subset. This determines the shape of an ellipsoid, which is then increased in size through multiplication by a constant until the ellipsoid covers at least the required h data points. Once this has been completed for each subset, the ellipsoid having the smallest volume is used to obtain the MVE estimates of location and scatter.

According to Jensen et al. (2007), the subsampling algorithm is widely used but suffers from repeatability issues. More specifically, MVE estimates of location and scatter from the

same data set can vary widely depending on the number of subsets used in the subsampling algorithm. The software package R was used in this research to calculate MVE estimates using Rousseeuw and Leroy's (1987) subsampling algorithm. The MVE function in R was employed using default input arguments, which are believed to be $h = (n+p+1)/2$ and 500 subsamples, although the R documentation is not completely clear about this. The fact that the MVE results determined by this research are somewhat different than those obtained by Vargas (2003) and Jensen et al. (2007) is likely due to the authors' use of different MVE input arguments.

Rousseeuw's (1984) MCD method, which is also robust, affine equivariant, and computationally complex, finds the subset of data that has the smallest covariance matrix determinant while covering a specified minimum number of points, h. It then uses the sample mean vector and the sample covariance matrix (as in the MVE approach, also multiplied by a constant) of the points in the subset as estimators of location and scatter. Rousseeuw and Van Driessen's (1999) FAST-MCD algorithm is a more computationally efficient version that approximates MCD estimates based on an iterative scheme involving randomly selected subsets of data. According to Jensen et al. (2007), the FAST-MCD algorithm does not suffer from the same repeatability issues as the MVE subsampling method and is therefore a better estimator. Like the MVE method, the integer value of $h = (n+p+1)/2$ provides the highest possible RBP for the MCD method of $[(n-p+1)/2]/n$, which converges to 50% as $n \to \infty$. Alternatively, h can be increased to $0.75n$ if the percentage of OC points is thought to be low. This increases the efficiency of the estimator because more IC observations from the reference sample are being used. All results in this research were derived using $h = 0.75n$ because the vast majority of scenarios evaluated involve contamination levels less than or equal to 20%, so there is no need to sacrifice statistical efficiency for a higher RBP. Using $h = 0.75n$, this research often achieved

better performance using the FAST-MCD method than Vargas (2003) and Jensen et al. (2007), especially for small samples with a low percentage of outliers.

2.3 The BACON Method of Robust Parameter Estimation

The BACON method possesses all the desired properties of robust estimators and is very computationally efficient, even for extremely large data sets and higher dimensions. It begins with a small outlier-free subset of the data and then allows this subset to grow rapidly until a stopping criteria is reached, sometimes taking only two to three iterations to converge to a satisfactory solution. Two versions of this iterative forward selection method are available -- Version 2 which is nearly affine equivariant and has RBPs exceeding 40% for various combinations of dimension and sample size, and Version 1 which is completely affine equivariant with an approximate RBP of 20%. MATLAB code for the BACON method is available from the authors upon request.

From Billor et al. (2000), the BACON algorithm is as follows:

Step 1: Identify an initial basic subset of $m > p$ observations that can safely be assumed free of outliers, where p is the dimension of the data and m is an integer chosen by the data analyst. Billor et al. (2000) suggest $m = cp$, where c is a small integer (such as 4 or 5) chosen by the data analyst.

Using Version 1 of the BACON method, Step 1 involves computing the Mahalanobis distances

$$d_i\left(\bar{\boldsymbol{X}}, \boldsymbol{S}\right) = \sqrt{\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)' \boldsymbol{S}^{-1}\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)}, \quad i = 1, \ldots, n, \tag{2.3.1}$$

where $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$ are the classical mean vector and covariance matrix, respectively, of the n observations. The $m = cp$ observations with the smallest values of $d_i\left(\bar{\boldsymbol{X}}, \boldsymbol{S}\right)$ are then nominated as a potential basic subset. According to Billor et al. (2000), this start is not robust, but it is affine equivariant and computationally easy. Furthermore, Billor et al. (2000) show that subsequent iterations make up for the nonrobust start as long as the fraction of outliers is relatively small (20% in 5 dimensions, 10% in 20 dimensions).

If Version 2 of the BACON method is used, Step 1 involves computing

$$\left\|\boldsymbol{X}_i - \boldsymbol{M}\right\|, \quad i = 1, \ldots, n, \tag{2.3.2}$$

where $\boldsymbol{M}$ is the coordinatewise median and $\|\cdot\|$ is the vector norm. The $m = cp$ observations with the smallest values of $\left\|\boldsymbol{X}_i - \boldsymbol{M}\right\|$ are then nominated as a potential basic subset. According to Billor et al. (2000), this start is robust but not affine equivariant (because the coordinatewise median is not affine equivariant) and slightly more computationally intensive because of the need to find medians in all directions. However, Billor et al. (2000) also state that because subsequent iterations are affine equivariant, the overall procedure is nearly affine equivariant, and it achieves a high RBP of approximately 40%.

Step 2: Fit an appropriate model to the basic subset, and from that model compute discrepancies for each of the observations. This involves computing

$$d_i\left(\bar{\boldsymbol{X}}_b, \boldsymbol{S}_b\right) = \sqrt{\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_b\right)' \boldsymbol{S}_b^{-1}\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}_b\right)}, \quad i = 1, \ldots, n, \tag{2.3.3}$$

where $\bar{\boldsymbol{X}}_b$ and $\boldsymbol{S}_b$ are the mean vector and covariance matrix, respectively, of the observations in the basic subset.

Step 3: Find a larger basic subset consisting of observations known (by their discrepancies) to be homogeneous with the basic subset. Generally, these are the observations

with the smallest discrepancies. This new basic subset may omit some of the previous basic subset observations, but it must be as large as the previous basic subset.

In order to accomplish this, the new basic subset will include all points with discrepancies less than $c_{npr}\chi_{p,\alpha/n}$, where $\chi^2_{p,\alpha}$ is the 1 - α percentile of the chi-square distribution with p degrees of freedom, $c_{npr} = c_{np} + c_{hr}$ is a correction factor, $c_{hr} = \max\{0, (h - r)/(h + r)\}$, $h = [(n + p + 1)/2]$, r is the size of the current basic subset, and

$$c_{np} = 1 + \frac{p+1}{n-p} + \frac{1}{n-h-p} = 1 + \frac{p+1}{n-p} + \frac{2}{n-1-3p}. \qquad (2.3.4)$$

The parameter α can be set to any number between 0 and 1, but $\alpha = 0.05$ is suggested by Billor et al. (2000) for most applications.

Step 4: Iterate Steps 2 and 3 until the size of the basic subset no longer changes.

Step 5: Nominate the observations excluded by the final basic subset as outliers.

Once the final basic subset is determined, estimates of location and scatter using the classical formulas for $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$ are computed and used in conjunction with Hotelling's T^2 chart to perform a Phase I analysis of individual multivariate normally distributed data. It is, however, important to recognize that the distribution of the T^2 statistic using BACON estimates from the outlier-free data set is not the same as the distribution of the T^2 statistic using classical estimates from the original data set. The same is true when MVE, MCD, or other estimators are used in lieu of classical estimators. In such cases, Tracy et al.'s (1992) Phase I UCL for Hotelling's T^2 chart given by Equation (1.2.6) is no longer appropriate.

Figure 2.3.1 is a graphical depiction of the results of incorrectly applying Tracy et al.'s (1992) Phase I UCL to Hotelling's T^2 chart using several non-classical estimators of location and scatter. In this example, the data are multivariate normally distributed and the desired IC FAP is

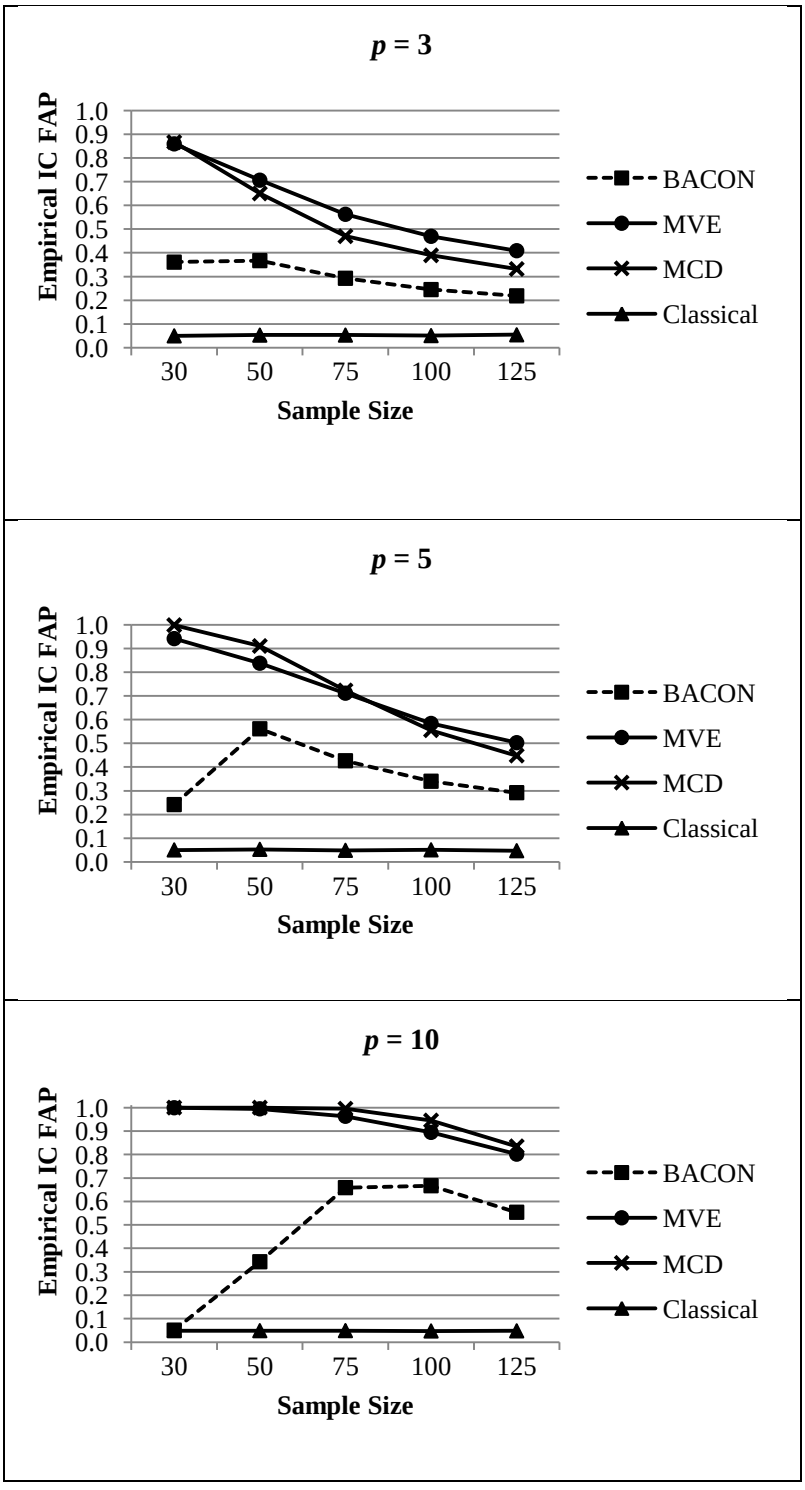

Figure 2.3.1 IC FAPs for Hotelling's T^2 Chart Using Tracy et al.'s (1992) UCL

0.05, but only Hotelling's T^2 chart using classical estimators maintains the desired IC FAP throughout the range of sample sizes and dimensions considered. The empirical IC FAPs for Hotelling's T^2 chart using BACON, MVE, and MCD estimators are in most cases significantly higher than the target of 0.05. This effect is usually most pronounced with small sample sizes and worsens with increasing dimension. If the distributions of the T^2 statistics using BACON, MVE, and MCD estimators were known, corresponding Phase I UCLs for Hotelling's T^2 chart could be derived mathematically. Since the distribution of the T^2 statistic using BACON estimators is unknown and the distribution of the T^2 statistic using MVE or MCD estimators is known only asymptotically [Jensen et al. (2009, p. 20)], Phase I UCLs must be empirically determined through simulation.

2.4 Empirical Control Limits for Hotelling's T^2 Chart with Robust Estimators

For each combination of n and p, using a desired IC FAP of 0.05, empirical Phase I UCLs for Hotelling's T^2 chart using BACON, MVE, and MCD estimators were determined using a methodology similar to the one outlined by Jensen et al. (2007):

1) Simulate 100,000 sets of individual multivariate normally distributed data with zero mean vector and identity covariance matrix. Due to the affine equivariance of the T^2 statistic, $\mathbf{0}$ and $\mathbf{I}$ may be used without loss of generality, so the resulting limits are applicable for any $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.
2) For each data set, determine robust estimates of location and scatter, compute T^2 statistics for all observations, and record the maximum T^2 statistic.

3) From the set of 100,000 maximum T^2 statistics, which represents the empirical distribution of the maximum T^2 statistic, determine the 95th percentile. This represents the empirical UCL for a desired IC FAP of 0.05.

A table of BACON, MVE, MCD, and classical control limits for all combinations of *n* and *p* used in this research is provided in Table 2.4.1, a table of input arguments for the BACON method is illustrated in Table 2.4.2, and MATLAB code for simulating additional control limits is provided in Appendix A. The input parameters in Table 2.4.2 were established through trial and error in order to achieve the highest possible alarm probabilities under a variety of OC conditions.

Desired IC FAP = 0.05					
n	***p***	**BACON UCL**	**MVE UCL**	**MCD UCL**	**Classical UCL**
30	2	21.07	41.65	27.73	10.55
30	3	24.28	58.65	41.26	12.21
50	3	22.05	35.39	28.56	14.14
100	3	21.60	26.15	24.09	16.41
30	5	34.45	84.26	69.78	14.92
50	5	29.09	46.49	43.61	17.41
100	5	27.14	32.56	31.19	20.21
30	10	20.02	217.91	192.24	20.05
50	10	49.68	95.35	103.09	23.98
100	10	40.89	51.32	52.64	28.09

Table 2.4.1 Empirical and Classical UCLs for Hotelling's T^2 Chart

BACON Input Arguments			
p	Version	α	*c*
2	2	0.10	6
3	2	0.10	6
5	2	0.10	4
10	2	0.10	3

Table 2.4.2 BACON Input Arguments for Each Dimension Evaluated

Repeating the experiment reflected in Figure 2.3.1 using the empirical UCLs for Hotelling's T^2 chart with BACON, MVE, and MCD estimators in Table 2.4.1 yields the graphs pictured in Figure 2.4.1. In this case, because the correct (empirical) UCLs were used, each robust version of Hotelling's T^2 chart maintains the desired IC FAP of 0.05 throughout the range of sample sizes and dimensions evaluated. Minor deviations from the target IC FAP of 0.05 are due to simulation noise. All simulations were conducted using the MATLAB code for Hotelling's T^2 chart provided in Appendix B, with the number of outliers set to zero.

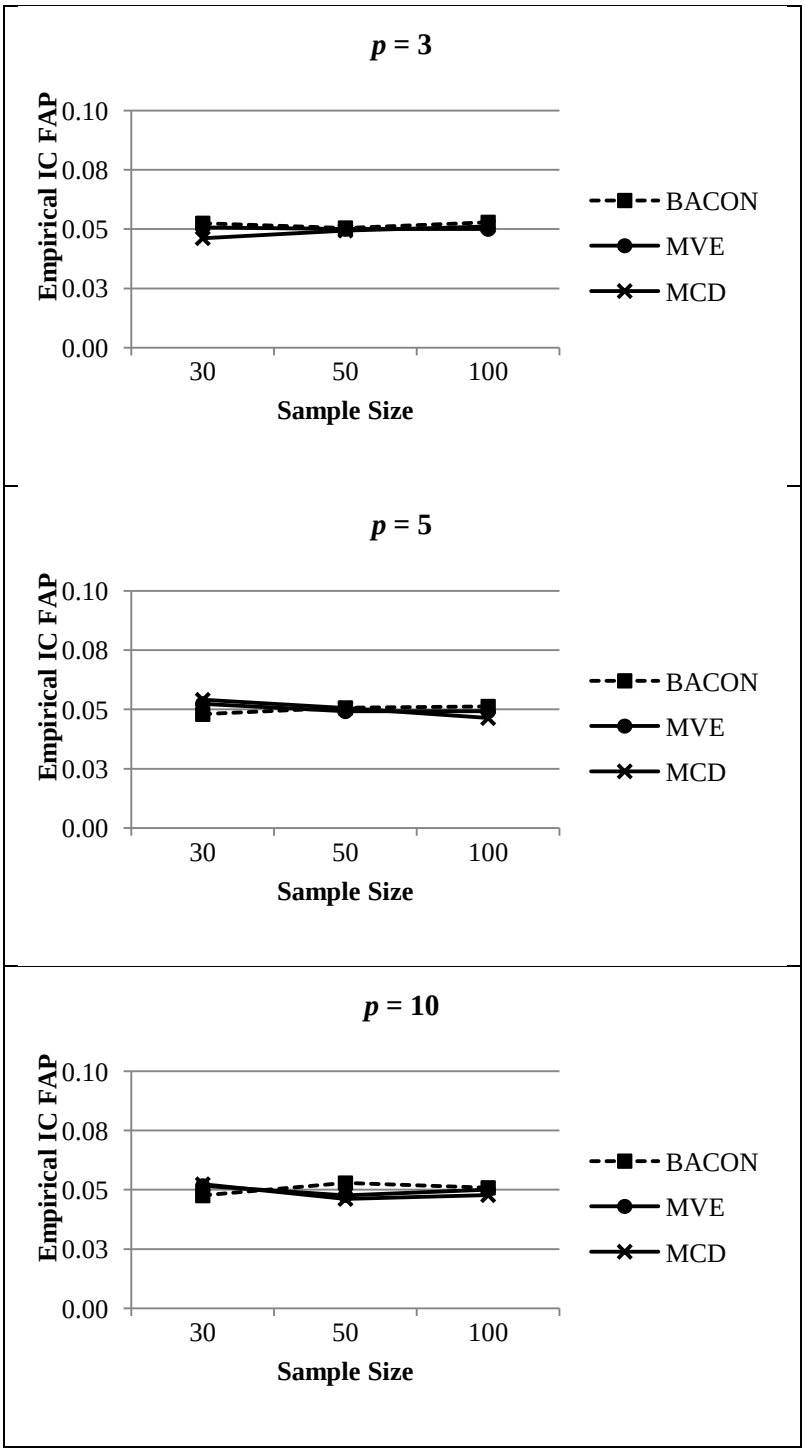

Figure 2.4.1 Hotelling's T^2 Chart IC FAPs Using Robust Estimators & Empirical UCLs

3 Evaluating Phase I Performance of Hotelling's T^2 Chart

To assess the effectiveness of Hotelling's T^2 chart using BACON estimators to establish an IC reference sample for individual multivariate normally distributed data, its performance was compared to Jobe and Pokojovy's (2009) cluster-based Hotelling's T^2 chart. This is a logical standard of performance because, as mentioned in Chapter 1, Jobe and Pokojovy's (2009) method was shown to be superior to previous attempts to improve the robustness of Hotelling's T^2 chart in most cases tested. Because the ultimate goal of this research is to match or exceed the performance of Jobe and Pokojovy's (2009) chart, the entire set of scenarios evaluated by Jobe and Pokojovy (2009) was replicated using Hotelling's T^2 chart with BACON estimators. Similar to Jobe and Pokojovy's (2009) design of experiments, Hotelling's T^2 charts using MVE, MCD, and classical estimators were also simulated for comparison purposes.

3.1 Simulating Multivariate Normally Distributed Data

Hotelling's T^2 chart was tested on individual multivariate normally distributed data under a variety of OC conditions, including situations in which simulated data in p = 2, 3, 5, and 10 dimensions contained a specified number of outliers as well as a scenario involving simulated bivariate data containing a 50% sustained shift of the mean. Due to affine equivariance of the mean vector and covariance matrix, multivariate normal data were generated without loss of generality from the standard multivariate normal distribution, $N_p(\mathbf{0}, \mathbf{I})$, where $\mathbf{0}$ is a p-dimensional mean vector of all zeros and $\mathbf{I}$ is a p x p identity matrix. The data were simulated

using MATLAB code from the MathWorks Statistics Toolbox at http://www.mathworks.com/help/toolbox/stats/. A summary of all planned simulations is illustrated in Table 3.1.1, where p = dimension, n = sample size, k = number of outliers, and NCP = shift size.

p	***n***	***k***	**NCP**
2	30	1(2)7	5(5)30
	30	15	4, 5(5)30
3, 5, 10	30	2(2)6	5(10)25
	50	2, 5, 10	5(10)25
	100	5, 10, 20	5(10)25

Table 3.1.1 Summary of Planned Experiments

3.2 Assessing Out-of-Control Performance

Hotelling's T^2 charts using BACON, MVE, MCD, and classical estimators were first evaluated in terms of their ability to detect k randomly occurring outliers, or mean-shifted data points. The k outliers were randomly arranged throughout the data in order to achieve consistency with Jobe and Pokojovy's (2009) simulation methodology. In general, the detection power of Hotelling's T^2 chart in Phase I is unaffected by the placement of outliers. This is because most parameter estimation methods (e.g. BACON, MVE, MCD, and classical methods) are unaffected by the time order of data, and Hotelling's T^2 chart in Phase I does not consider the time order of data when making simultaneous comparisons of all control chart statistics to the UCL. In contrast, the clustering method of parameter estimation used by Jobe and Pokojovy (2009) in their version of Hotelling's T^2 chart does account for the sequencing of data, hence the necessity of randomizing outliers in their simulation methodology.

Next, Hotelling's T^2 charts using BACON, MVE, MCD, and classical estimators were evaluated in terms of their ability to detect a 50% sustained shift of the mean occurring in the latter half of a data set. Sustained shifts can be induced anywhere in the data set without loss of generality, but were placed at the end of each data set for purposes of consistency with Jobe and Pokojovy's (2009) methodology. It should be noted that this level of contamination exceeds the BACON method's maximum RBP of 40%, so a significant degradation in performance of Hotelling's T^2 chart using BACON estimators was expected.

The magnitude of each shift was measured by the noncentrality parameter

$$NCP = \delta \sum^{-1} \delta' \tag{3.2.1}$$

where the process mean vector shifts from μ_o to $\mu_o + \delta$ and Σ is the process covariance matrix. Because the direction of a shift does not affect control chart performance with elliptically symmetric distributions, shifts were fixed in the direction of $\boldsymbol{e}_1 = [1, 0, ..., 0]$ without loss of generality [Stoumbos and Sullivan (2002), p. 265].

OC performance of Hotelling's T^2 charts using BACON, MVE, MCD, and classical estimators was quantified in terms of the empirical alarm probability (EAP), where EAP is defined as the estimated probability of a chart signaling at least once in an OC situation. Ideally, a control chart's EAP should be 100% for all scenarios involving outliers. The algorithm for simulating OC performance of Hotelling's T^2 chart is as follows:

1) Simulate *n* observations from a *p*-dimensional standard normal distribution.
2) Shift a specified number of randomly selected observations by the desired NCP.

3) Compute a T^2 statistic for each observation and compare it to the empirical UCL if robust estimators are used or Tracy et al.'s (1992) Phase I UCL if classical estimators are used. If at least one T^2 statistic exceeds the UCL, increment a counter by one.
4) Repeat steps 1 - 3 a total of 10,000 times.
5) Estimate the overall EAP = (final counter value)/10,000.

This process was repeated for all experiments listed in Table 3.1.1. The MATLAB program for simulating Hotelling's T^2 chart performance is provided in Appendix B.

4 Comparing the BACON-Based Hotelling's T^2 Chart to Other Robust Versions

Through a variety of scenarios involving randomly occurring outliers, a 50% sustained shift of the mean, and an example application to a bivariate data set, Hotelling's T^2 chart with BACON estimators was compared to Hotelling's T^2 chart using Jobe and Pokojovy's (2009) clustering method as well as the MVE, MCD, and classical methods included in Jobe and Pokojovy's (2009) research. Due to unavailability of complete computer code for Jobe and Pokojovy's (2009) algorithm, simulation results for Hotelling's T^2 chart using clustering were taken directly from Jobe and Pokojovy (2009).

For comparison to their cluster-based chart, Jobe and Pokojovy (2009) took tables of simulation results for Hotelling's T^2 chart using the MVE, MCD, and classical methods directly from Vargas (2003). However, this is problematic because it is unknown whether Vargas (2003) and Jobe and Pokojovy (2009) employed equivalent simulation methodologies (software, input arguments, number of iterations, etc.). As discussed in Chapter 2, the choice of input arguments can have a substantial impact on the performance of the MVE and MCD methods. In contrast, simulation results for Hotelling's T^2 chart using the MVE, MCD, and classical methods were properly recreated here using the simulation methodology outlined in Chapter 3.

4.1 Detecting Randomly Occurring Outliers

The first set of performance comparisons involved the detection of k randomly occurring outliers in a bivariate normally distributed data set of size $n = 30$. As indicated in the top panel

of Figure 4.1.1, Hotelling's T^2 chart with classical estimators is superior when only one outlier is present, although Hotelling's T^2 chart using BACON estimators and Jobe and Pokojovy's (2009) cluster-based chart offer only slightly lesser performance. As illustrated in the bottom panel of Figure 4.1.1, however, the performance of Hotelling's T^2 chart with classical estimators quickly degrades as k is increased from 1 to 3, whereas Hotelling's T^2 chart using MCD estimators, BACON estimators, and Jobe and Pokojovy's (2009) clustering method continue to perform exceptionally well.

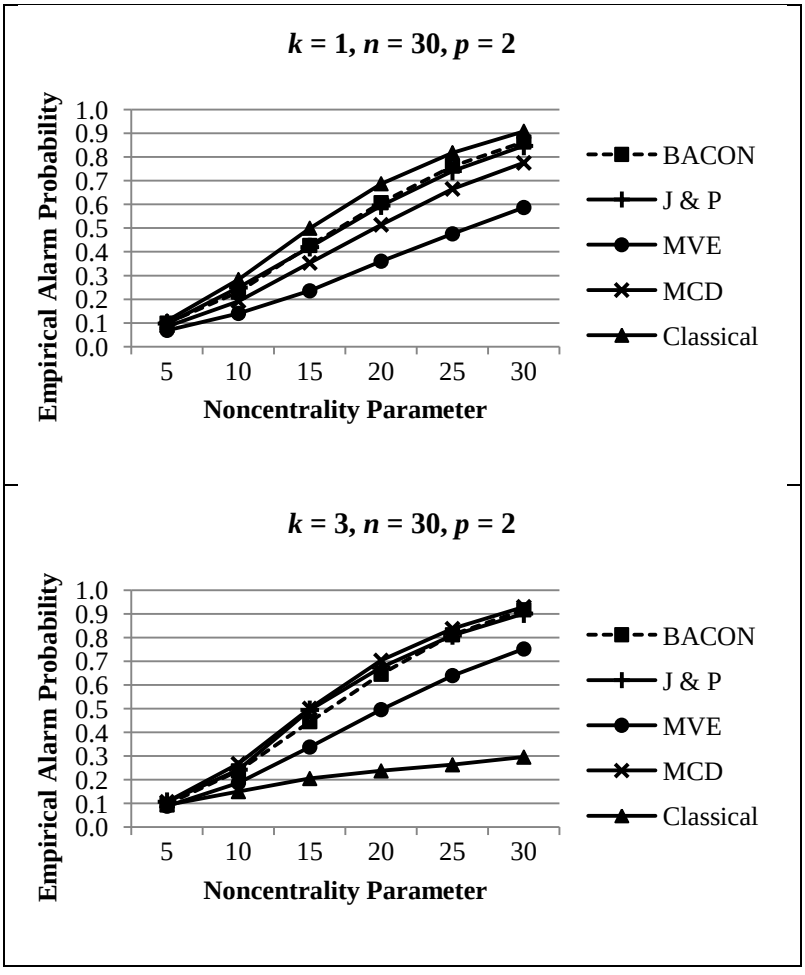

Figure 4.1.1 Control Chart Performance on Bivariate Normal Data with k = 1, 3 Outliers

As depicted in the top panel of Figure 4.1.2, when the number of outliers is increased to 5, Hotelling's T^2 chart using MCD estimators becomes the best option, although Jobe and Pokojovy's (2009) cluster-based chart remains competitive. When the number of outliers is increased to 7 as shown in the bottom panel of Figure 4.1.2, Jobe and Pokojovy's (2009) cluster-based chart is slightly better than Hotelling's T^2 chart using MVE or MCD estimators. For both k = 5 and k = 7, the performance of Hotelling's T^2 chart using BACON estimators falls behind the other robust methods, especially for small to moderate NCPs.

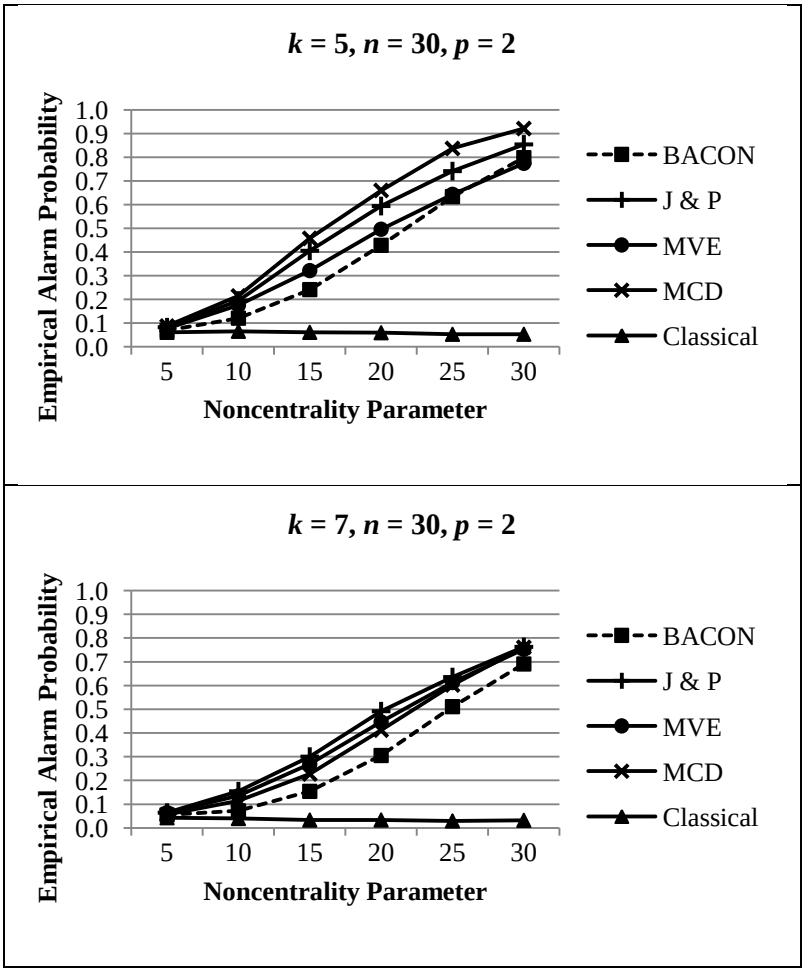

Figure 4.1.2 Control Chart Performance on Bivariate Normal Data with k = 5, 7 Outliers

Trends similar to those witnessed with bivariate normally distributed data are also observed in the three-dimensional scenarios summarized in Table 4.1.1, where the highest EAP for each scenario is in bold. Hotelling's T^2 chart using BACON estimators provides excellent performance when the number of outliers is very small. However, as the number of outliers is increased, Jobe and Pokojovy's (2009) cluster-based chart achieves superiority over all other methods.

		n = 30			n = 50			n = 100		
Method	**NCP**	k = 2	k = 4	k = 6	k = 2	k = 5	k = 10	k = 5	k = 10	k = 20
					p = 3					
BACON	5	**0.0879**	**0.0706**	0.0572	**0.1027**	**0.0908**	0.0476	0.1264	0.0904	0.0517
J & P	5	0.0710	0.0680	**0.0730**	0.0950	0.0890	**0.0850**	**0.1410**	**0.1790**	**0.1760**
MVE	5	0.0640	0.0631	0.0595	0.0770	0.0866	0.0659	0.1254	0.1221	0.0715
MCD	5	0.0782	0.0702	0.0660	0.0952	0.0907	0.0638	0.1240	0.1234	0.0588
Classical	5	0.0869	0.0663	0.0498	0.0988	0.0772	0.0462	0.1158	0.0830	0.0475
BACON	15	**0.3871**	0.2103	0.1194	**0.5239**	0.3056	0.0701	0.6671	0.2995	0.0478
J & P	15	0.3800	**0.3770**	**0.3950**	0.4860	**0.6190**	**0.6640**	0.7440	**0.8490**	**0.9170**
MVE	15	0.2031	0.2185	0.1904	0.3579	0.4550	0.3550	0.7194	0.7674	0.5733
MCD	15	0.3022	0.3162	0.2294	0.4780	0.5708	0.3362	**0.7588**	0.7864	0.3878
Classical	15	0.2627	0.0830	0.0475	0.4091	0.1391	0.0418	0.4205	0.1387	0.0435
BACON	25	**0.7476**	0.5382	0.3921	**0.8846**	0.6992	0.3191	0.9712	0.6730	0.1496
J & P	25	0.7470	**0.7830**	**0.8070**	0.8560	**0.9490**	**0.9660**	0.9870	**0.9970**	**0.9990**
MVE	25	0.4145	0.4634	0.4416	0.7044	0.8405	0.7999	0.9798	0.9928	0.9692
MCD	25	0.6084	0.6778	0.6056	0.8336	0.9260	0.8170	**0.9874**	0.9968	0.9266
Classical	25	0.4460	0.0897	0.0450	0.7469	0.1689	0.0468	0.6963	0.1633	0.0380

Table 4.1.1 Empirical Alarm Probabilities for p = 3

This trend is illustrated in Figure 4.1.3, which represents control chart performance using a three-dimensional normally distributed sample of size n = 30 with increasing k. In the case of k = 2 outliers, Hotelling's T^2 chart using BACON estimators and Jobe and Pokojovy's (2009) cluster-based chart provide nearly identical performance. Both charts are substantially better than Hotelling's T^2 chart using MCD, MVE, or classical estimators. As k is increased to 6, however, the performance of Hotelling's T^2 chart using BACON estimators drops below the performance of Hotelling's T^2 chart using MVE or MCD estimators, whereas Jobe and

Pokojovy's (2009) cluster-based chart maintains a significant performance advantage over all other methods.

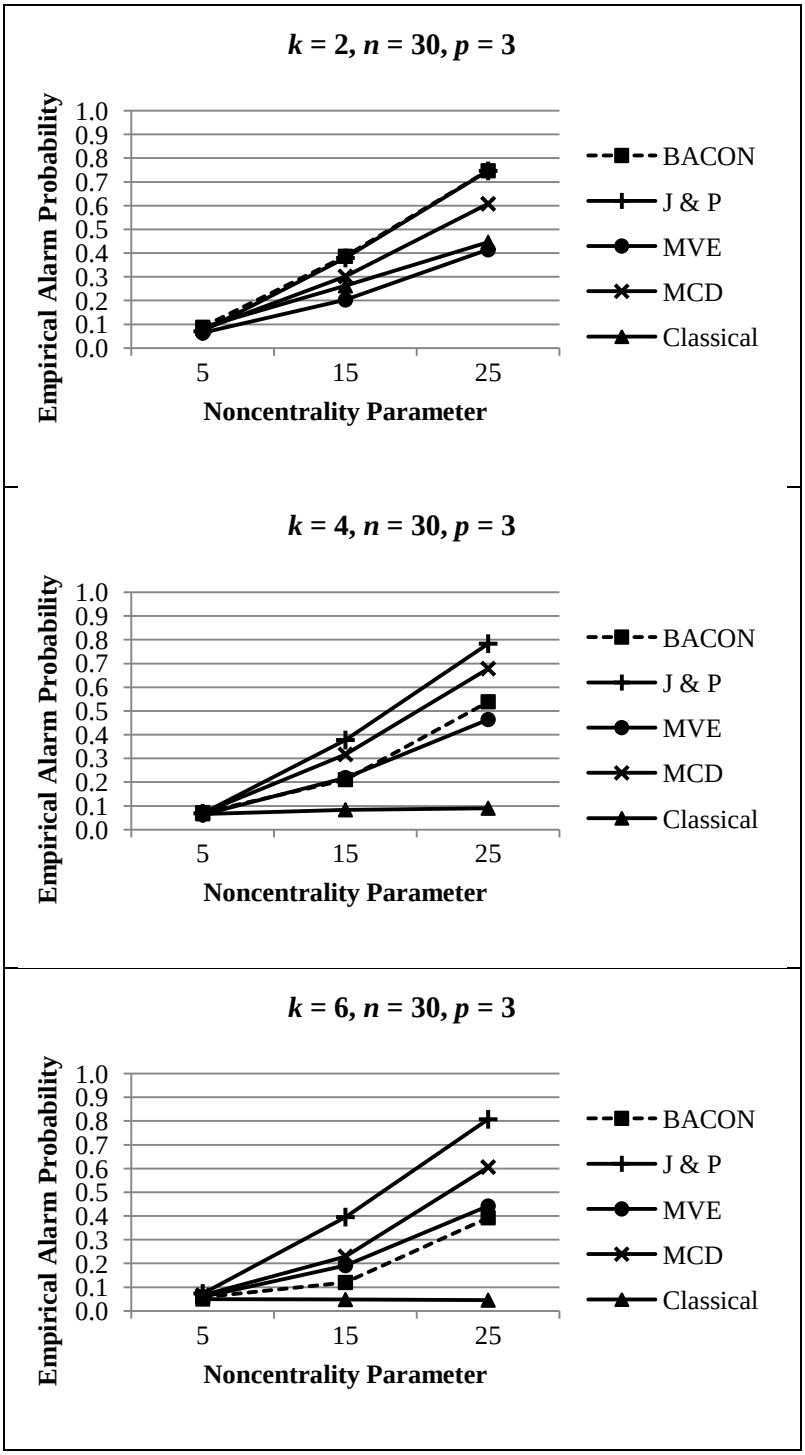

Figure 4.1.3 Control Chart Performance with *k* Outliers When *n* = 30 and *p* = 3

In all simulations performed, it was noted that for a given (k, n, p) combination, the performance of Hotelling's T^2 chart using BACON estimators improved as the NCP was increased, which is what one would expect when using a robust parameter estimation method. In general, OC points with the largest NCPs should be detected more frequently than those with small to moderate NCPs because they are farthest from the center of the data as defined by the robust mean estimate. However, as shown in Figure 4.1.4, for a given (n, p, NCP) combination, the performance of Hotelling's T^2 chart using BACON estimators actually worsened as k was increased, which is somewhat counterintuitive. One would expect the presence of more OC points to result in a higher probability of an alarm, assuming the OC points are correctly excluded from computation of the T^2 statistic by the BACON method.

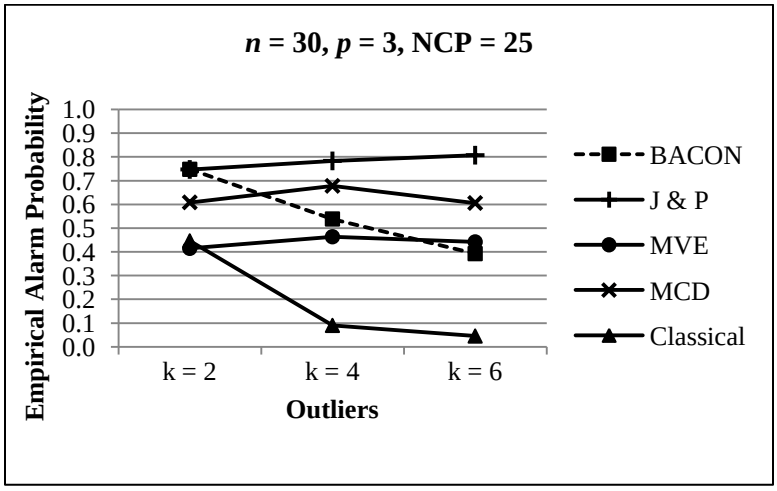

Figure 4.1.4 Effect of Increasing k on Control Chart Performance

These trends together suggest that the BACON method is unable to consistently exclude OC points with small to moderate NCPs from a data set, but rather performs optimally when the NCP corresponding to an OC point is very large. In order to confirm this, additional simulations were performed using NCPs beyond the range of those tested by Jobe and Pokojovy (2009).

Expanding the range of NCPs from NCP = 5(10)25 to NCP = 5(10)55, the performance of Hotelling's T^2 charts using BACON, MVE, MCD, and classical estimators are displayed in Figure 4.1.5. As suspected, even with a large percentage of outliers (6/30 = 20% in this case), Hotelling's T^2 chart with BACON estimators gets progressively better as the NCP is increased, and is second only to Hotelling's T^2 chart with MCD estimators when the NCP > 25. Because of its comparatively low computational burden, Hotelling's T^2 chart with BACON estimators is therefore an excellent option when it is suspected that extreme outliers are present in the data. Jobe and Pokojovy's (2009) cluster-based chart was not included in this performance comparison due to lack of availability of their simulation algorithm.

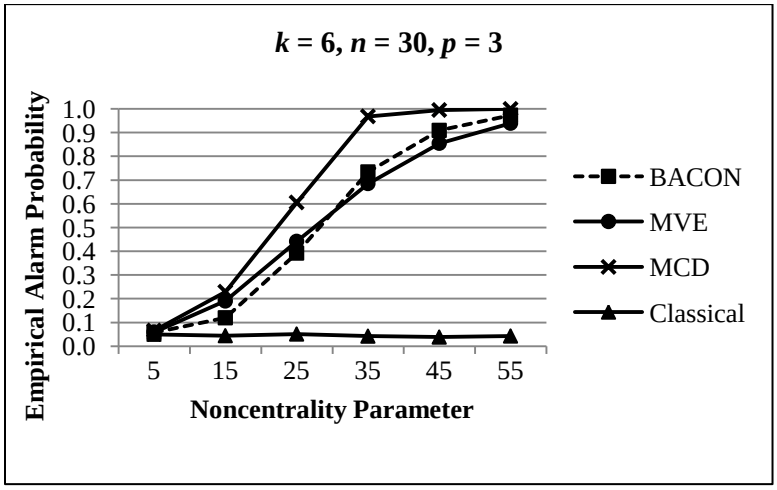

Figure 4.1.5 Control Chart Performance with Extreme Outliers

Table 4.1.2 shows empirical alarm probabilities for all charts evaluated using normally distributed data in five dimensions, with the highest EAP for each scenario in bold. As with the two- and three-dimensional cases, Hotelling's T^2 chart using BACON estimators is competitive as long as k is small relative to n, but Jobe and Pokojovy's (2009) cluster-based chart provides the best overall performance across the range of scenarios evaluated. This is further illustrated

for sample size $n = 50$ in Figure 4.1.6. In comparison to the results obtained in two and three dimensions, there is a loss of detection power observed for all control charts in five dimensions. This is to be expected because outliers are known to be more difficult to detect in higher dimensions than in low dimensions.

		$n = 30$			$n = 50$			$n = 100$		
Method	**NCP**	$k = 2$	$k = 4$	$k = 6$	$k = 2$	$k = 5$	$k = 10$	$k = 5$	$k = 10$	$k = 20$
					$p = 5$					
BACON	5	0.0696	0.0614	0.0570	0.0759	0.0669	0.0523	0.0938	0.0739	0.0489
J & P	5	**0.0830**	**0.0670**	0.0590	**0.0840**	**0.0920**	**0.0690**	0.0900	**0.1070**	**0.0840**
MVE	5	0.0568	0.0571	0.0496	0.0666	0.0644	0.0561	**0.0945**	0.0886	0.0525
MCD	5	0.0684	0.0598	**0.0598**	0.0700	0.0724	0.0548	0.0924	0.0758	0.0498
Classical	5	0.0704	0.0584	0.0541	0.0792	0.0650	0.0484	0.0886	0.0683	0.0513
BACON	15	0.2175	0.1073	0.0652	0.3428	0.1531	0.0548	0.4475	0.1507	0.0504
J & P	15	**0.2270**	**0.2550**	**0.2200**	**0.3820**	**0.3690**	**0.3450**	**0.6130**	**0.7280**	**0.7610**
MVE	15	0.1312	0.1238	0.0915	0.2386	0.2744	0.1333	0.5592	0.5305	0.1877
MCD	15	0.1916	0.1764	0.1002	0.2834	0.3384	0.1616	0.6018	0.5884	0.2012
Classical	15	0.1649	0.0673	0.0503	0.2717	0.1017	0.0501	0.2689	0.0999	0.0476
BACON	25	0.4723	0.2611	0.1918	0.7114	0.3937	0.1444	0.8363	0.3271	0.0688
J & P	25	**0.5030**	**0.5450**	**0.5480**	**0.7410**	**0.8110**	**0.8230**	**0.9690**	**0.9930**	**0.9910**
MVE	25	0.2639	0.2667	0.1896	0.5443	0.6441	0.3683	0.9302	0.9370	0.5694
MCD	25	0.3904	0.3874	0.2544	0.6122	0.7586	0.5420	0.9586	0.9740	0.8046
Classical	25	0.2638	0.0701	0.0490	0.5284	0.1190	0.0433	0.4623	0.1135	0.0461

Table 4.1.2 Empirical Alarm Probabilities for $p = 5$

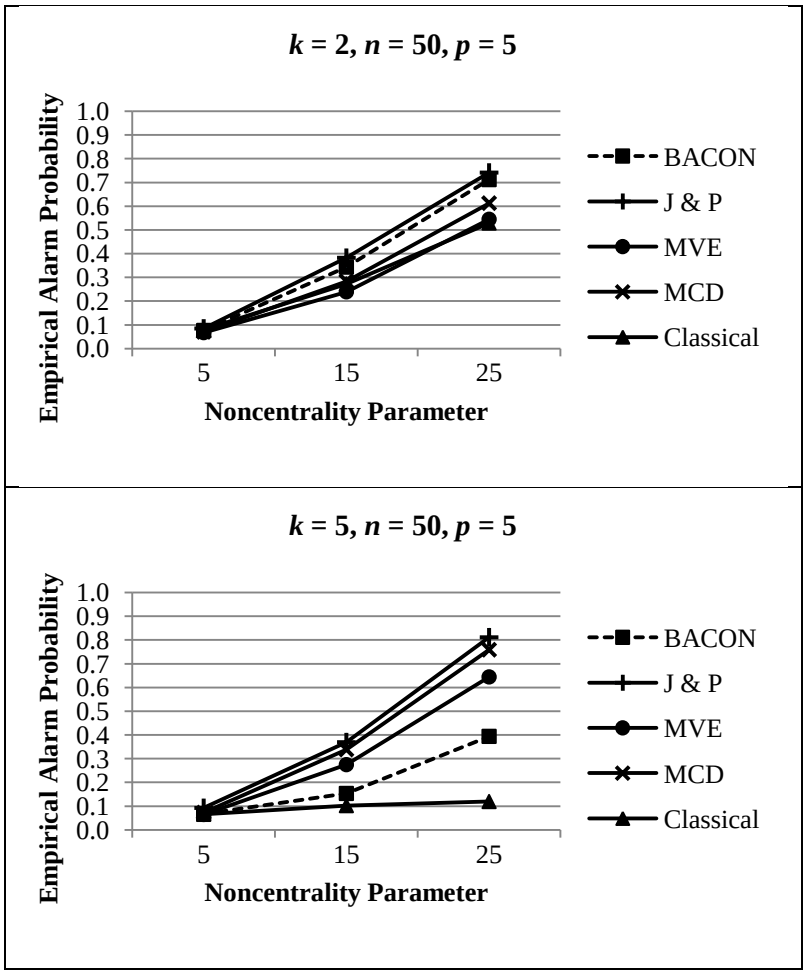

Figure 4.1.6 Control Chart Performance with *k* Outliers When *n* = 50 and *p* = 5

The loss of detection power in higher dimensions is most evident in the results for the ten-dimensional scenarios provided in Table 4.1.3. With most charts depicted in Table 4.1.3, the detection power is relatively poor. In fact, for NCP = 5, the highest EAPs for all *n* and *k* are barely above the IC FAP of 0.05. Even Jobe and Pokojovy's (2009) cluster-based chart provides acceptable performance only for the largest sample size and shift evaluated (*n* = 100, NCP = 25).

Method	NCP	n = 30			n = 50			n = 100		
		k = 2	k = 4	k = 6	k = 2	k = 5	k = 10	k = 5	k = 10	k = 20
						p = 10				
BACON	5	0.0544	0.0569	**0.0570**	0.0584	0.0588	0.0490	0.0671	0.0601	0.0516
J & P	5	**0.0610**	**0.0570**	0.0530	**0.0690**	**0.0730**	**0.0580**	**0.0680**	0.0490	**0.0530**
MVE	5	0.0555	0.0562	0.0508	0.0509	0.0550	0.0498	0.0604	0.0569	0.0493
MCD	5	0.0528	0.0496	0.0536	0.0490	0.0460	0.0422	0.0612	**0.0608**	0.0474
Classical	5	0.0585	0.0497	0.0472	0.0611	0.0620	0.0522	0.0648	0.0593	0.0460
BACON	15	**0.0874**	0.0624	0.0503	0.1443	0.0744	0.0519	0.1819	0.0871	0.0551
J & P	15	0.0790	**0.0880**	**0.0960**	**0.1810**	**0.2120**	**0.2020**	0.2190	0.2300	**0.2100**
MVE	15	0.0723	0.0578	0.0541	0.0925	0.0859	0.0542	0.2433	0.1597	0.0557
MCD	15	0.0740	0.0604	0.0550	0.0938	0.0970	0.0576	**0.2794**	**0.2662**	0.0690
Classical	15	0.0860	0.0528	0.0530	0.1320	0.0719	0.0525	0.1334	0.0772	0.0443
BACON	25	0.1150	0.0576	0.0552	0.3213	0.1021	0.0537	0.4046	0.1089	0.0500
J & P	25	**0.1600**	**0.1760**	**0.1710**	**0.4160**	**0.5010**	**0.5000**	**0.7360**	**0.8060**	**0.8140**
MVE	25	0.0979	0.0670	0.0591	0.1939	0.1546	0.0574	0.5923	0.3966	0.0754
MCD	25	0.1326	0.1022	0.0532	0.2176	0.2594	0.0830	0.7256	0.7650	0.2610
Classical	25	0.1065	0.0610	0.0519	0.2510	0.0810	0.0506	0.2240	0.0849	0.0492

Table 4.1.3 Empirical Alarm Probabilities for p = 10

The difficulty in detecting shifts in higher dimension (especially with very small samples) is further illustrated in Figure 4.1.7, which represents ten-dimensional normally distributed data with various sample sizes containing 20% outliers. Again, most charts are unable to detect the presence of the shifted data, although Jobe and Pokojovy's (2009) cluster-based chart has some success in doing so as the sample size, number of outliers, and NCP are raised.

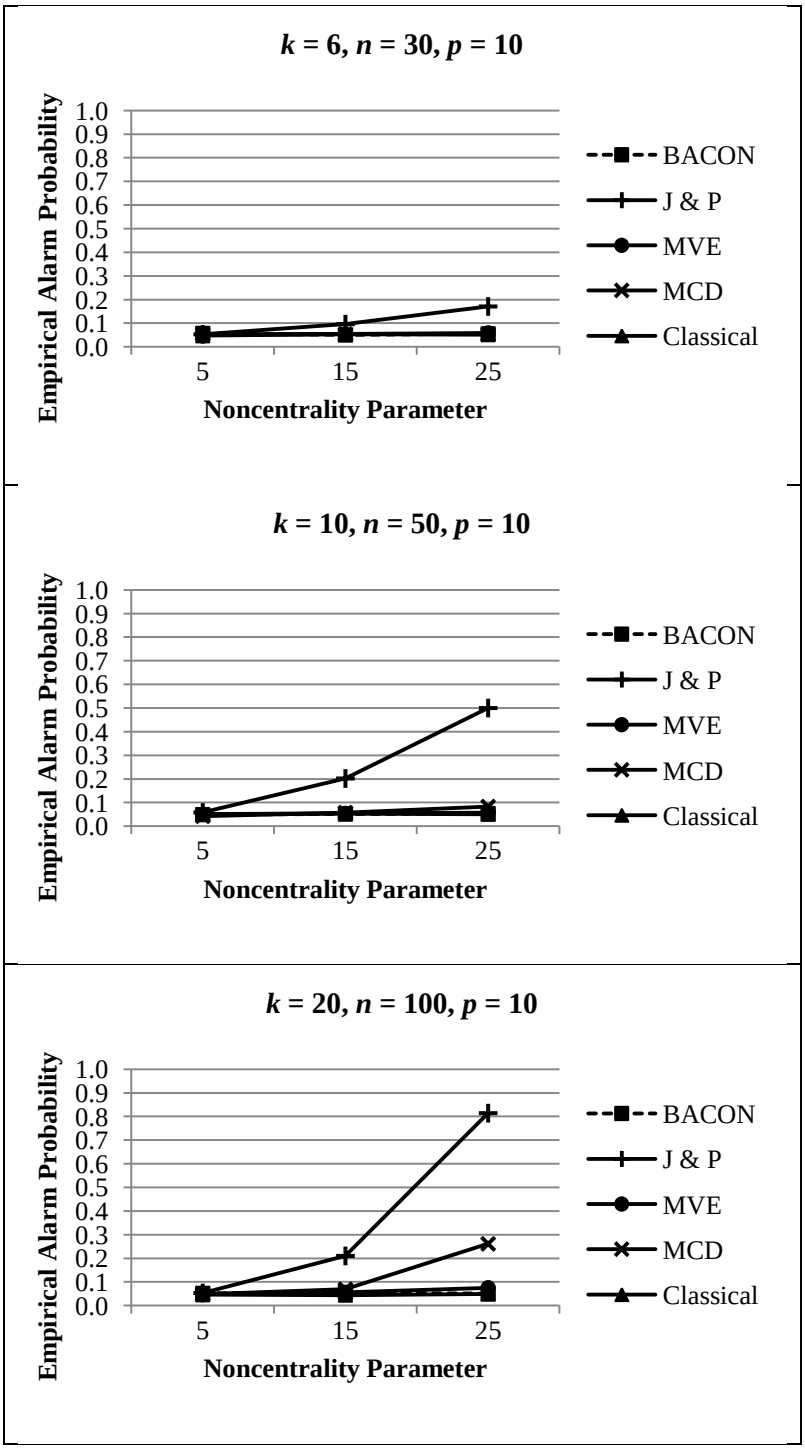

Figure 4.1.7 Control Chart Performance with 20% Outliers in Ten Dimensions

In order to visualize the changes in detection power that occur as the dimension is increased, Figure 4.1.8 shows control chart performance with $k = 5$ outliers using a sample size

of $n = 100$ in both three and ten dimensions. From the scenario depicted in Figure 4.1.8, it is evident that the detection power for each variation of Hotelling's T^2 chart is much greater when $p = 3$ than it is when $p = 10$.

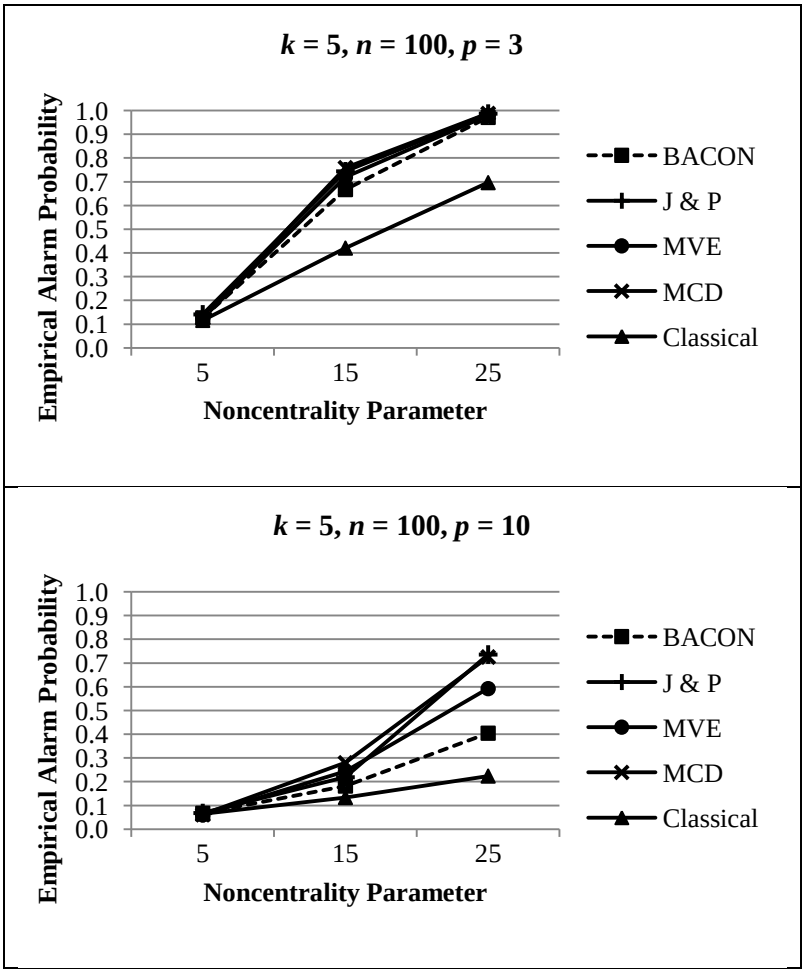

Figure 4.1.8 Effect of Increasing Dimension on Control Chart Performance

4.2 Detecting a Sustained Shift of the Mean

Following the outline of Jobe and Pokojovy (2009), the next scenario evaluated involved a sample of 30 bivariate normally distributed observations with a sustained shift of the mean occurring in the latter half of the data (so $k = 15$). The numerical results are provided in Table

4.2.1 where the highest EAP for each NCP is in bold, and the graphical results are portrayed in Figure 4.2.1. Hotelling's T^2 chart using BACON estimators was not expected to perform well since the maximum RBP of the BACON method is approximately 40%. This proved to be the case, although the BACON method did perform significantly better than the MVE, MCD, and classical methods for the larger NCPs evaluated. Jobe and Pokojovy's (2009) cluster-based chart demonstrated the best performance overall. Hotelling's T^2 chart using MVE, MCD, and classical estimators were unable to detect the presence of such a high level of contamination, producing EAPs that were in most cases even lower than the desired IC FAP of 0.05.

Method \ NCP	**4**	**5**	**10**	**15**	**20**	**25**	**30**
BACON	0.0336	0.0313	0.0287	0.0500	0.0900	0.1622	0.2483
J & P	**0.2650**	**0.3560**	**0.6930**	**0.8730**	**0.9520**	**0.9820**	**0.9860**
MVE	0.0399	0.0398	0.0443	0.0425	0.0431	0.0510	0.0483
MCD	0.0398	0.0304	0.0312	0.0320	0.0324	0.0370	0.0320
Classical	0.0316	0.0283	0.0242	0.0228	0.0215	0.0225	0.0230

Table 4.2.1 Empirical Alarm Probabilities Under a 50% Sustained Shift of the Mean

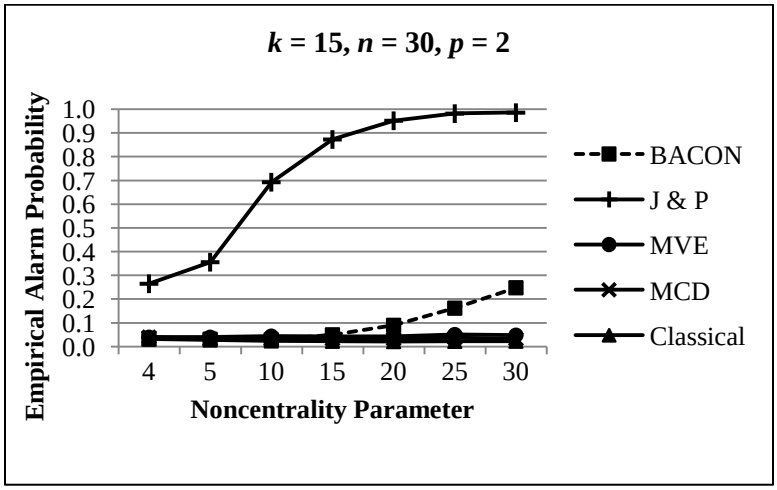

Figure 4.2.1 Control Chart Performance Under a 50% Sustained Shift of the Mean

4.3 Application to an Example Data Set

The final performance evaluation for Hotelling's T^2 chart using BACON estimators involved its application to the bivariate data set depicted in Table 4.3.1. The original data set was presented by Quesenberry (2001) in 11 variables, but later pared down to two variables by Vargas (2003) for purposes of comparing the performance of Hotelling's T^2 chart using MVE, MCD, and classical estimators as well as two methods proposed by Sullivan and Woodall (1996). Jobe and Pokojovy (2009) subsequently used the bivariate data set from Vargas (2003) to compare the effectiveness of their cluster-based chart to the five aforementioned methods evaluated by Vargas (2003).

Simulation exercises resulted in observation 2 being identified as an outlier by all versions of Hotelling's T^2 chart. This is contrary to the findings of Vargas (2003), who reported that Hotelling's T^2 chart using MCD estimators failed to identify the lone outlier. As discussed in Chapter 2, this is probably because the MCD method in this research used a higher percentage of the sample than Vargas (2003) to compute more accurate estimates of location and scatter.

i	1	**2**	3	4	5	6	7	8	9	10
x_1	0.567	**0.538**	0.530	0.562	0.483	0.525	0.556	0.586	0.547	0.531
x_2	60.558	**56.303**	59.524	61.102	59.834	60.228	60.756	59.823	60.153	60.640
i	11	12	13	14	15	16	17	18	19	20
x_1	0.581	0.583	0.540	0.458	0.554	0.469	0.471	0.457	0.565	0.664
x_2	59.785	59.675	60.489	61.067	59.788	58.640	59.574	59.718	60.901	60.180
i	21	22	23	24	25	26	27	28	29	30
x_1	0.600	0.586	0.567	0.496	0.485	0.573	0.520	0.556	0.539	0.554
x_2	60.493	58.370	60.216	60.214	59.500	60.052	59.501	58.476	58.666	60.239

Table 4.3.1 Example Bivariate Data Set

Next, imitating the analysis of Vargas (2003), the bivariate sample was altered by changing observation 16 to (0.469, 56.23) and observation 24 to (0.496, 56.08) to make them outliers. Of all the variations of Hotelling's T^2 chart evaluated by Vargas (2003) and Jobe and Pokojovy (2009), as well as the BACON-based alternative presented here, only the clustering and BACON versions correctly identified all three observations (2, 16, and 24) as outliers. The classical, MVE, MCD, and both Sullivan and Woodall (1996) versions of Hotelling's T^2 chart failed to detect at least one of the three outlying observations. Again, this is contrary to Vargas (2003), who determined that Hotelling's T^2 chart using MVE estimators correctly identified all three outliers and that Hotelling's T^2 chart using MCD estimators failed to do so. As previously stated, this is likely due to each author using different MVE and MCD input arguments.

Control chart statistics for Hotelling's T^2 chart using BACON estimators for the original and altered samples are provided in Table 4.3.2. Entries in bold indicate control chart statistics that exceed the empirical UCL of 21.07 and therefore signal a potential out-of-control condition. The corresponding control chart for the altered data is presented in Figure 4.3.1.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Original	0.92	**24.96**	0.35	2.61	1.51	0.31	1.29	0.93	0.09	1.03	0.77	0.96	0.59	6.11	0.12
Altered	0.87	**26.68**	0.51	2.62	1.87	0.34	1.25	0.80	0.06	0.99	0.65	0.83	0.54	6.09	0.10
i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Original	4.95	2.30	3.15	1.87	6.59	1.90	5.96	0.39	1.15	1.63	0.44	0.51	4.27	3.04	0.22
Altered	**30.15**	2.89	3.78	1.85	6.55	1.86	5.93	0.32	**30.94**	2.14	0.35	0.74	4.51	3.40	0.17

Table 4.3.2 T^2 Statistics for Original and Altered Samples Using BACON Estimators

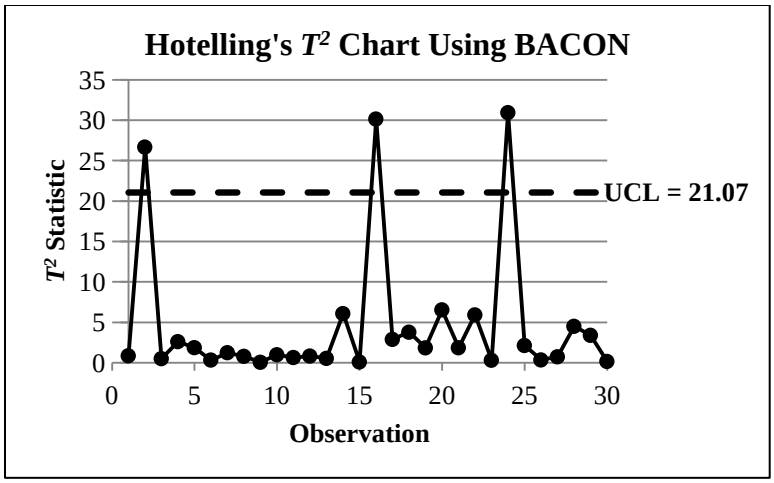

Figure 4.3.1 Application of the BACON-Based Hotelling's T^2 Chart to Altered Data

5 Summary and Conclusions

Hotelling's T^2 chart using BACON estimators, although ultimately unsuccessful in exceeding the standard of performance established by Jobe and Pokojovy's (2009) cluster-based chart, was shown to be a viable option for Phase I analysis of individual multivariate normally distributed data under certain conditions. If the level of contamination in a reference data set of size $n = 30 - 100$ is thought to be relatively small (less than 5 - 10%, depending on the sample size and dimension), or if any number of outliers are suspected to be extremely large (as measured by their NCP, and also dependent on n and p), Hotelling's T^2 chart using BACON estimators offers good performance with a low computational burden. This is in contrast to other robust parameter estimation methods such as MVE and MCD which, as discussed by Billor et al. (2000), are significantly more computationally complex.

If, on the other hand, a more universally robust procedure is desired, Jobe and Pokojovy's (2009) cluster-based chart was shown to be extremely effective across a wide range of scenarios, including the presence of a 50% sustained shift of the mean -- a level of contamination which for most methods renders IC data indistinguishable from OC data. However, Jobe and Pokojovy's (2009) control chart is not without potential drawbacks. The relatively complicated clustering algorithm may not be easy to understand for practitioners, and the computational complexity of the procedure is not detailed by the authors. Furthermore, because complete computer code for Jobe and Pokojovy's (2009) cluster-based chart is not readily available, their simulation results cannot be validated. It is therefore impossible to give their method an unqualified endorsement

at this time. While it is assumed that the general observations regarding the comparative performance of Jobe and Pokojovy's (2009) chart to other robust Hotelling's T^2 charts are valid, detailed conclusions cannot be drawn without replicating Jobe and Pokojovy's (2009) simulation results using the process outlined in Chapter 3. This is a topic for future research.

Another area that merits further investigation is the effect of sample size on control chart performance. In accordance with Jobe and Pokojovy's (2009) experimental design, only sample sizes of n = 30, 50, and 100 were evaluated in this research. Though they may be commonly encountered in Phase I scenarios, these are very small sample sizes for all but the smallest dimensions considered in this research (p = 2, 3, 5, and 10). Billor et al. (2000) originally designed the BACON method to be a computationally efficient robust parameter estimation method for extremely large data sets in higher dimensions, and the authors offered a variety of simulation results for n = 100 - 10,000 and p = 5 - 20 to demonstrate its performance. The authors noted that the BACON method matched the performance of the MVE, MCD, and other robust parameter estimation methods on all published test problems, but at a mere fraction of the computational expense. If used as originally intended (with large n and p), BACON estimators in conjunction with Hotelling's T^2 chart might become a more attractive option than Hotelling's T^2 chart using MVE or MCD estimators and perhaps even Jobe and Pokojovy's (2009) cluster-based chart, both in terms of speed and accuracy.

In conclusion, improving the robustness of Hotelling's T^2 chart applied to individual multivariate normally distributed data in Phase I has been the subject of much research over the years, yet this study has shown that many questions remain unanswered. Numerous variations of Hotelling's T^2 chart have been proposed, but none so far appear to have attained a balance of accuracy, robustness, computational complexity, and ease of implementation for a wide range of

n and *p*. The BACON-based version of Hotelling's T^2 chart presented here, despite falling short of its original performance objectives, provides a valuable contribution by demonstrating its strengths and weaknesses as a Phase I method when sample sizes are small, and in the process identifying several areas worthy of additional research.

References

Alfaro, J.L., & Ortega, J.F. (2008). A Robust Alternative to Hotelling's T^2 Control Chart Using Trimmed Estimators. *Quality and Reliability Engineering International*, 24, 601-611.

Bersimis, S., Psarakis, S., & Panaretos, J. (2007). Multivariate Statistical Process Control Charts: An Overview. *Quality and Reliability Engineering International*, 23, 517-543.

Billor, N., Hadi, A.S., & Velleman, P.F. (2000). BACON: Blocked Adaptive Computationally Efficient Outlier Nominators. *Computational Statistics & Data Analysis*, 34, 279-298.

Chenouri, S., & Steiner, S.H. (2009). A Multivariate Robust Control Chart for Individual Observations. *Journal of Quality Technology*, 41(3), 259-271.

Chenouri, S., & Variyath, A.M. (2011). A Comparative Study of Phase II Robust Multivariate Control Charts for Individual Observations. *Quality and Reliability Engineering International*, 27(7), 857-865.

Donoho, D.L. & Huber, P.J. (1983). The Notion of a Breakdown Point. In P.J. Bickel, K.A. Doksum and J.L. Hodges, Jr. (Eds.), *A Festschrift for Eric L. Lehmann* (pp. 157-184). Belmont, CA: Wadsworth.

Hotelling, H. (1947). Multivariate Quality Control – Illustrated By the Air Testing of Sample Bombsights. In C. Eisenhart, M.W. Hastay, & W.A. Wallis (Eds.), *Techniques of Statistical Analysis* (pp. 111-184). New York, NY: McGraw-Hill.

Jensen, W.A., Jones-Farmer, L.A., Champ, C.W., & Woodall, W.H. (2006). Effects of Parameter Estimation on Control Chart Properties: A Literature Review. *Journal of Quality Technology*, 38(4), 349-364.

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

Jobe, J.M., & Pokojovy, M. (2009). A Multistep, Cluster-Based Multivariate Chart for Retrospective Monitoring of Individuals. *Journal of Quality Technology*, 41(4), 323-339.

Lopuhaa, H.P., & Rousseeuw, P.J. (1991). Breakdown Points of Affine Equivariant Estimators of Location and Covariance Matrices. *The Annals of Statistics*, 19, 229-248.

Lowry, C.A., & Montgomery, D.C. (1995). A Review of Multivariate Control Charts. *IIE Transactions*, 27, 800-810.

Mahalanobis, P. C. (1936). On the Generalized Distance in Statistics. *Proceedings of the National Institute of Science of India*, 12, 49-55.

Mason, R.L., Champ, C.W., Tracy, N.D., Wierda, S.J., & Young, J.C. (1997). Assessment of Multivariate Process Control Techniques. *Journal of Quality Technology*, 29(2), 140-143.

Mason, R.L., & Young, J.C. (2002). *Multivariate Statistical Process Control with Industrial Applications*. Alexandria, VA: American Statistical Association; Philadelphia, PA: Society for Industrial and Applied Mathematics.

Mohammadi, M., Midi, H., Arasan, J., & Al-Talib, B. (2011). High Breakdown Estimators to Robustify Phase II Multivariate Control Charts. *Journal of Applied Sciences*, 11(3), 503-511.

Montgomery, D.C. (2005). *Introduction to Statistical Quality Control* (5th ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Mudholkar, G.S., & Srivastava, D.K. (2000). A Class of Robust Stepwise Alternatives to Hotelling's T^2 Tests. *Journal of Applied Statistics*, 27(5), 599-619.

Nedumaran, G., & Pignatiello, J.J. (2000). On Constructing T^2 Control Charts for Retrospective Examination. *Communications in Statistics – Simulation and Computation*, 29(2), 621-632.

Oyeyemi, G.M., & Ipinyomi, R.A. (2010). A Robust Method of Estimating Covariance Matrix in Multivariate Data Analysis. African Journal of Mathematics and Computer Science Research, 3(1), 1-18.

Rousseeuw, P.J. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association*, 79, 871-880.

Rousseeuw, P.J., & Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. New York, NY: John Wiley & Sons, Inc.

Rousseeuw, P.J., & van Driessen, K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41, 212-223.

Rousseeuw, P.J., & van Zomeren, B.C. (1990). Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association*, 85, 633-651.

Shewhart, W.A. (1939). *Statistical Method from the Viewpoint of Quality Control*. New York: Dover Publications.

Srivastava, D.K., & Mudholkar, G.S. (2001). Trimmed T^2: A Robust Analog of Hotelling's T^2. *Journal of Statistical Planning and Inference*, 97, 343-358.

Sullivan, J.H., & Woodall, W.H. (1996). A Comparison of Multivariate Control Charts for Individual Observations. *Journal of Quality Technology*, 28, 398-408.

Sullivan, J.H., & Woodall, W.H. (1998). Adapting Control Charts for the Preliminary Analysis of Multivariate Observations. *Communications in Statistics – Simulation and Computation*, 27(4), pp. 953-979.

Tiku, M.L., & Balakrishnan, N. (1988). Robust Hotelling-Type T^2 Statistics Based on the Modified Maximum Likelihood Estimators. *Communications in Statistics – Theory and Methods*, 17(6), 1789-1810.

Tiku, M.L., & Singh, M. (1982). Robust Statistics for Testing Mean Vectors of Multivariate Distributions. *Communications in Statistics – Theory and Methods*, 11(9), 985-1001.

Vargas, J.A. (2003). Robust Estimation in Multivariate Control Charts for Individual Observations. *Journal of Quality Technology*, 35(4), 367-376.

Wierda, S.J. (1994). Multivariate Statistical Process Control – Recent Results and Directions for Future Research. *Statistica Neerlandica*, 48, 147-168.

Willems, G., Pison, G., Rousseeuw, P.J., & Van Aelst, S. (2002). A Robust Hotelling Test. *Metrika*, 55, 125-138.

Woodall, W.H., & Montgomery, D.C. (1999). Research Issues and Ideas in Statistical Process Control. *Journal of Quality Technology*, 31(4), 376-386.

Yanez, S., Gonzalez, N., & Vargas, J.A. (2010). Hotelling's T^2 Control Charts Based on Robust Estimators. *Dyna*, 163, 239-247.

Appendices

Appendix A: MATLAB Code for Simulating Hotelling's T^2 Chart Empirical Control Limits

Appendix B: MATLAB Code for Simulating Hotelling's T^2 Chart Performance

Appendix A: MATLAB Code for Simulating Hotelling's T^2 Chart Empirical Control Limits

```
%=========================================================================%
%   FINDING EMPIRICAL CONTROL LIMITS FOR HOTELLING'S T^2 CONTROL CHART    %
%=========================================================================%
%  -Created by Richard Bell 8/12/2011; last updated 9/25/2011             %
%  -Finds empirical UCLs for Hotelling's T2 control chart with BACON,     %
%   MVE, or MCD location and scatter estimators                           %
%  -Uses same method as Jensen, Birch, and Woodall (2007)                 %
%=========================================================================%


clear all  % clear all objects in the MATLAB workspace
clc  % clear the output screen


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%% INPUT SIMULATION PARAMETERS %%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

iterations=100000;  % number of simulation iterations to be performed
input=xlsread('c:\Users\Rich\Documents\InputFile.xlsx','Sheet1','a1:b10');
inputRows=length(input(:,1));  % determine the number of rows of data in the
      input file
empUCLtable=zeros(inputRows,1);  % initialize the table of empirical UCLs to
      all zeros

% NOTE:  USE ONLY IF MVE METHOD IS EMPLOYED
[status,msg] = openR;
if status ~= 1
    disp(['Problem connecting to R: ' msg]);
end

for row=1:inputRows  % perform the simulation below for each scenario in the
      input file
    n=input(row,1);  % read in the sample size
    p=input(row,2);  % read in the number of variables

    alpha=.05;  % desired overall false alarm probability (FAP) for the chart
    percentile=(1-alpha)*100;  % percentile corresponding to the desired
      alpha level
    maxT2vector=zeros(iterations,1);  % initialize the vector of maximum T2
      statistics to all zeros


    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    %%%%%%%%%% GENERATE DATA AND CONSTRUCT HOTELLING'S T2 CHART %%%%%%%%%%
    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

    count=0;  % initialize the counter for the number of iterations performed

    while count < iterations  % run the entire loop for a set number of
      iterations
```

```
    %=====> SIMULATE MULTIVARIATE NORMAL DATA

    mu=zeros(1,p);  % set the mean vector to all zeros
    sigma=eye(p);  % set the covariance matrix equal to the identity
        matrix
    X=mvnrnd(mu,sigma,n);  % generate multivariate normal data


    %=====> COMPUTE ROBUST ESTIMATES OF LOCATION AND SCATTER (must code
        out 2 of the 3 methods listed using % symbols)

    % BACON METHOD
    out=baconV(X,2,.10,6);  % compute BACON estimate for location using
        Mahalanobis distance, alpha=0.05, and c=4; use version 2
        (Euclidean distance) if expected contamination exceeds 20 percent
    Xbar_robust=out.center3;  % BACON estimate for mean vector
    S_robust=out.cov3;  % BACON estimate for covariance matrix

    % MVE METHOD
    evalR('library(MASS)');
    putRdata('X',X);
    putRdata('n',n);
    putRdata('p',p);
    Xbar_robust=evalR('cov.mve(X, cor=FALSE, quantile.used=floor((n + p +
        1)/2), nsamp = "best")$center');
    S_robust=evalR('cov.mve(X, cor=FALSE, quantile.used=floor((n + p +
        1)/2), nsamp = "best")$cov');

    % MCD METHOD
    [rew,raw]=mcdcov(X,'plots',0);  % compute MCD estimates for location
        and scatter using default parameter values; suppress plot output
        by adding the arguments ('plots',0)
    Xbar_robust=rew.center;  % MCD estimate for mean vector
    S_robust=rew.cov;  % MCD estimate for covariance matrix


    %=====> COMPUTE HOTELLING'S T2 STATISTICS AND COMPARE TO UCL

    T2vector=zeros(n,1);  % initialize the vector of T2 statistics

    for i=1:n  % compute T2 control statistic for each observation
        T2stat=(X(i,:)-Xbar_robust)/S_robust*(X(i,:)-Xbar_robust)';
        T2vector(i)=T2stat;  % store the T2 statistics in a vector
    end

    count=count+1;  % increment the counter for the total number of
        iterations performed

    maxT2=max(T2vector);  % identify the maximum T2 statistic
    maxT2vector(count,1)=maxT2;  % store the maximum T2 statistic in a
        vector

end
```

```
    empUCL=prctile(maxT2vector,percentile);  % compute the empirical UCL for
      the current scenario IAW Jensen, Birch, and Woodall (2007)

    % store the results in a table and display the table on the output screen
    empUCLtable(row,1)=empUCL;
    disp(empUCLtable);

    % send the results to an Excel file

xlswrite('c:\Users\Rich\Documents\OutFile.xlsx',empUCLtable,'Sheet1','A1');

end

%%%%%%%%%%%%%%%%%%%%%%%%%%%% END OF PROGRAM %%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Appendix B: MATLAB Code for Simulating Hotelling's T^2 Chart Performance

```
%=========================================================================%
%               HOTELLING'S T^2 CONTROL CHART PROGRAM FILE                %
%=========================================================================%
%  -Created by Richard Bell on 9/15/2010; last updated on 9/25/2011       %
%  -Based on Hotelling's T2 chart with classical or empirical UCLs        %
%  -File is set up to run multiple scenarios; before using, undesired     %
%   sections must be commented out using "%"                              %
%=========================================================================%


clear all  % clear all objects in the MATLAB workspace
clc  % clear the output screen


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%% INPUT SIMULATION PARAMETERS %%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% read in m, n, UCL, shift size, and p from an Excel file
iterations=10000;  % number of simulation iterations to be performed
input=xlsread('c:\Users\Rich\Documents\InputFile.xlsx','Sheet1','a1:e10');
inputRows=length(input(:,1));  % determine the number of rows of data in the
      input file
APtable=zeros(inputRows,1);  % initialize the table of alarm probabilities to
      all zeros

% NOTE:  USE ONLY IF MVE METHOD IS EMPLOYED
[status,msg] = openR;
if status ~= 1
    disp(['Problem connecting to R: ' msg]);
end

for row=1:inputRows  % perform the simulation below for scenario in the input
      file
    n=input(row,1);  % read in the sample size
    p=input(row,2);  % read in the number of variables
    UCL=input(row,3);  % read in the empirical upper control limit
    shiftSize=input(row,4); % read in the size of the shift (as defined by
      the NCP)
    numOC=input(row,5);  % read in the number of out-of-control points

    count=0;  % initialize the counter for the number of iterations performed
    alarmCount=0;  % initialize the alarm counter

    while count < iterations  % run the entire loop for a set number of
iterations
```

```
%=====> SIMULATE MULTIVARIATE NORMAL DATA

% NOTE:  USE THIS UCL ONLY FOR NORMALLY DISTRIBUTED DATA!
alpha=.05;  % desired overall false alarm probability (FAP) for the
    chart
alphaAdjusted=1-(1-alpha)^(1/n);  % desired FAP for each individual
    comparison
UCL=((n-1)^2/n)*betainv(1-alphaAdjusted,p/2,(n-p-1)/2)  % Tracy,
    Young, and Mason's (1992) Phase I UCL

mu=zeros(1,p);  % set the mean vector to all zeros
sigma=eye(p);  % set the covariance matrix equal to the identity
    matrix
X=mvnrnd(mu,sigma,n);  % simulate multivariate normally distributed
    data
shift=zeros(1,p);  % initialize the shift vector to all zeros
shift(1)=sqrt(shiftSize);  % place the desired shift in the first
    position of the shift vector

% check the NCP to ensure it equals the desired value
NCP=shift/eye(p)*shift';
if abs(NCP-shiftSize) > 0.0001  % display error message if calculated
    NCP is significantly different than shift size (they should be
    equal since the theoretical covariance matrix of X is I)
    disp('ERROR in NCP!')
end

% add the desired shift to randomly selected points
i=1;
randIndex=randperm(n)';
while i <= numOC
    X(randIndex(i),:)=X(randIndex(i),:)+shift;
    i = i + 1;
end


%=====> COMPUTE ROBUST ESTIMATES OF LOCATION AND SCATTER (must code
    out 2 of the 3 methods listed using % symbols)

% BACON METHOD
out=baconV(X,2,.10,3);  % compute BACON estimate for location using
    Mahalanobis distance, alpha=0.05, and c=4; use version 2
    (Euclidean distance) if expected contamination exceeds 20 percent
Xbar_robust=out.center3;  % BACON estimate for mean vector
S_robust=out.cov3;  % BACON estimate for covariance matrix

% MVE METHOD (requires code for R interface)
evalR('library(MASS)');  % call the R library named "MASS"
putRdata('X',X);  % send sample data to R
Xbar_robust=evalR('cov.mve(X)$center');  % use R to find MVE estimate
    for mean vector
S_robust=evalR('cov.mve(X)$cov');  % use R to find MVE estimate for
    covariance matrix
```

```
        % MCD METHOD
        [rew,raw]=mcdcov(X,'plots',0);  % compute MCD estimates for location
            and scatter using default parameter values; suppress plot output
            by adding the arguments ('plots',0)
        Xbar_robust=rew.center;  % MCD estimate for mean vector
        S_robust=rew.cov;  % MCD estimate for covariance matrix

        % CLASSICAL METHOD
        Xbar_robust=mean(X);
        S_robust=cov(X);


        %=====> COMPUTE HOTELLING'S T2 STATISTICS AND COMPARE TO UCL

        alarm=0;  % initialize indicator variable representing an alarm (=1)
            or no alarm (=0)
        T2vector=zeros(n,1);  % initialize vector of T2 statistics

        for i=1:n  % perform loop for all observations in the sample
            if alarm==0  % continue loop as long as no false alarms occur
                T2stat=(X(i,:)-Xbar_robust)/S_robust*(X(i,:)-Xbar_robust)';
                  % compute T2 control statistic for each observation
                T2vector(i)=T2stat;  % store T2 control statistics in a
                  vector
                if T2stat > UCL
                    alarm=1;  % issue an alarm if the current T2 control
                        statistic exceeds the UCL
                end
            end
        end

        if alarm==1
            alarmCount=alarmCount+1;  % if a control chart issues an alarm,
            increment the counter representing total alarms for all
            iterations
        end

        count=count+1;  % increment the counter for the total number of
            iterations performed

    end

    AP=alarmCount/iterations;  % estimate the alarm probability (AP) for the
      current scenario
    APtable(row,1)=AP;  % store the current AP in a table
    disp(APtable);  % display AP table for Hotelling's T2 chart on screen

    % send the estimated APs to an Excel file

    xlswrite('c:\Users\Rich\Documents\OutFile.xlsx',APtable,'Sheet1','A1');

end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% END OF PROGRAM %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```