A Cluster-Based Outlier Detection Scheme for Multivariate Data

J. Marcus Jobe and Michael Pokojovy

Detection power of the squared Mahalanobis distance statistic is significantly reduced when several outliers exist within a multivariate dataset of interest. To overcome this masking effect, we propose a computer-intensive cluster-based approach that incorporates a reweighted version of Rousseeuw’s minimum covariance determinant method with a multi-step cluster-based algorithm that initially filters out potential masking points. Compared to the most robust procedures, simulation studies show that our new method is better for outlier detection. Additional real data comparisons are given. Supplementary materials for this article are available online.

KEY WORDS: Bandwidth; Breakdown point; Kernel estimator; Size and power.

1. INTRODUCTION

Let a sample of \( n \) \( p \)-dimensional data values be denoted as \( x = (x_1, \ldots, x_n)' \). The squared Mahalanobis distance statistic for \( x_i \) is

\[
T_i^2 = (x_i - \bar{x})S^{-1}(x_i - \bar{x}), \tag{1}
\]

where \( \bar{x} \) and \( S \) are the usual sample mean vector and covariance matrix based on \( x \). If \( T_i^2 \) exceeds the limit \( L_{1-\alpha} \) given by

\[
L_{1-\alpha} = \frac{(n-1)\gamma}{n}B_{1-\alpha; \frac{np}{2}, \frac{np}{2}} \tag{2}
\]

\( x_i \)'s is determined to be an outlier. Equation (2) assumes the \( x_i \) to be independent with the same multivariate normal distribution. Assuming no outliers, if \( n \) tests are performed using Equation (2) and each test is of size \( \alpha \) where \( \alpha > 0 \) is chosen small, the probability at least one test incorrectly identifies an outlier is \( \gamma = 1 - (1 - \alpha)^n \). Even for moderately sized samples, \( \gamma \) can be unacceptably close to 1. So, to address this issue, for any \( n, \gamma > 0 \) \( n \) is chosen small and \( \alpha \) in Equation (2) becomes

\[
\alpha = 1 - (1 - \gamma)^{1/n}. \tag{3}
\]

Thus, \( B_{1-\alpha; \frac{np}{2}, \frac{np}{2}} \) is the \( (1 - \alpha) \)-quantile of the beta distribution (Traacy, Young, and Mason 1992; Wierda 1994). Tracy, Young, and Mason (1992) also identified the \( T_i^2 \)'s in Equation (1) to be dependent. Hence, the approach implementing the beta random variable described above is approximate. We address this deficiency later in the article.

When only one outlier exists, the above-described method is effective at detecting its existence. The primary objection to the use of Equations (1) and (2) comes from the fact that when multiple outliers are present, significant masking occurs and the power to detect them becomes undesirable small.

A natural approach to overcome the masking effect is to substitute into Equation (1) robust estimators of the mean vector and covariance matrix. Many authors have proposed robust estimators of the mean and covariance matrix from multivariate data. The minimum volume ellipsoid (MVE) method, the minimum covariance determinant (MCD) method, and estimators generated using trimming approaches are the most commonly adopted for detecting multivariate outliers.

Jobe and Pokojovy (2009) developed an original multi-step, cluster-based algorithm for detecting outliers from individuals multivariate data collected sequentially over time. In this article, we propose a new multi-step, cluster-based algorithm for detecting multivariate outliers from an unordered sample \( x \). Our new method implements the recently developed high breakdown estimators of the multivariate mean vector \( \mu \) and covariance matrix \( \Sigma \) described by Cerioli (2010). These new high breakdown estimators, referred to as finite sample reweighted MCD (FSRMCD) estimators, are a modification of the reweighted MCD estimators of Hawkins and Olive (1999) and Rousseeuw and van Driessen (1999).

An overview of our new algorithm can be characterized by the following. From a sample of interest \( x \), FSRMCD estimators are used to identify a bandwidth to nonparametrically estimate the multivariate probability density with the well-known kernel method by Parzen (1962). Application of the modal EM (MEM) algorithm, outlined by Li, Ray, and Lindsay (2007), to the nonparametrically estimated multivariate probability density function identifies local maxima. MEM determines an ascending path for each \( x_i \) to a local maximum. All \( x_i \) assigned to the same local maximum are considered to be members of the same cluster. Li, Ray, and Lindsay (2007) developed and named this cluster assignment modal association clustering (MAC). The largest cluster is identified and filtered using a symmetry measure. Remaining data points from this cluster are used to construct reweighted estimators of the mean vector \( \mu \) and covariance matrix \( \Sigma \). With these robust estimators, squared Mahalanobis distances \( d_i^2 \), whose form is given in Equation (9) are calculated for each \( x_i \) from \( x \). If the maximum \( d_i^2 \) exceeds a limit \( L_1 \), then at least one outlier is detected and each \( d_i^2 \) is then compared to a second limit \( L_2 \). If the \( d_i^2 \) exceeds \( L_2 \), \( x_i \) is identified as an outlier. We determine the limits \( L_1 \) and \( L_2 \) by simulation for selected sample size \( n \) and dimension \( p \) such that the family-wise and per-comparison rates are controlled at a
selected level $\gamma > 0$. We refer to our new method as FSRMCD-MAC. The ability to detect outliers with the FSRMCD-MAC method is compared to that of the highly recommended IRMCD method developed by Cerioli (2010). For selected $(n, p)$ combinations, the proportions of good points identified as outliers by the FSRMCD-MAC method, typically referred to as swamping rates, are compared to the rates produced by the IRMCD method. Outlier detection performance of both the IRMCD and FSRMCD-MAC methods are compared for real data given by Flury and Riedwyl (1988).

2. THEORETICAL FOUNDATIONS

2.1 Finite Sample Reweighted MCD

For multivariate data $\mathbf{x}$, the MCD method attempts to find the subset of $h$ points such that the determinant of the covariance matrix from that subset is minimal among all determinants of the covariance matrix derived from any other set of $h$ points contained in $\mathbf{x}$. The MCD robust estimators then become the corresponding mean and appropriately scaled covariance matrix calculated from the identified subset $\hat{\mathcal{T}}_{\text{MCD}}$:

$$\hat{\mu}_{\text{MCD}} = \frac{1}{h} \sum_{i \in \hat{\mathcal{T}}_{\text{MCD}}} \mathbf{x}_i,$$

$$\hat{\Sigma}_{\text{MCD}} = \frac{k_{\text{MCD}}}{h-1} \left( \sum_{i \in \hat{\mathcal{T}}_{\text{MCD}}} (\mathbf{x}_i - \hat{\mu}_{\text{MCD}})(\mathbf{x}_i - \hat{\mu}_{\text{MCD}})^\prime \right),$$

where $k_{\text{MCD}}$ is a proportionality constant assuring for unbiasedness and consistency of the estimators if the data $\mathbf{x}$ are iid normal (Croux and Haesbroeck 1999). Davies (1987), Rousseeuw and Leroy (1987), and Lopuhaä and Rousseeuw (1991) showed that $h = \lfloor (n + p + 1)/2 \rfloor$ should be used to obtain the largest breakdown point for the MCD estimators, where $\lfloor \cdot \rfloor$ denotes the integer part.

We briefly outline the FSRMCD algorithm here. First, raw estimates $\hat{\mu}_{\text{MCD}}$ and $\hat{\Sigma}_{\text{MCD}}$ given by Equation (3) are computed using the proportionality constant $k_{\text{MCD}} = \frac{\sigma_{\text{MCD}}}{\sqrt{h(n/h)}}$ with $\sigma_{\text{MCD}}$ as proposed by Pison, Van Aelst, and Willems (2002). Letting

$$d^2_{\text{MCD},i} = (\mathbf{x}_i - \hat{\mu}_{\text{MCD}})^\prime \hat{\Sigma}_{\text{MCD}}^{-1} (\mathbf{x}_i - \hat{\mu}_{\text{MCD}}),$$

the weights $w_i$ are computed using the cut-off value $D_{p,0.975}$, that is, $w_i = 1$, $d^2_{\text{MCD},i} \leq D_{p,0.975}$, or $w_i = 0$, otherwise, where $D_{p,0.975}$ stands for an approximation of the 0.975th quantile of the distribution of $d^2_{\text{MCD},i}$ obtained by Hardin and Rocke (2005). The FSRMCD estimators are then the resulting reweighted estimates

$$\hat{\mu}_{\text{FSRMCD}} = \frac{1}{m} \sum_{i = 1}^{n} w_i \mathbf{x}_i,$$

$$\hat{\Sigma}_{\text{FSRMCD}} = \frac{k_{\text{FSRMCD}}}{m-1} \sum_{i = 1}^{n} w_i (\mathbf{x}_i - \hat{\mu}_{\text{FSRMCD}})(\mathbf{x}_i - \hat{\mu}_{\text{FSRMCD}})^\prime,$$

where $k_{\text{FSRMCD}} = \frac{\sigma_{\text{MCD}}}{\sqrt{h(n/h)}}$ as proposed by Cerioli (2010) and $m = \sum_{i = 1}^{n} w_i$.

In the present article, we apply the FSRMCD estimators implemented in the mcd-routine of the FSDA toolbox (www.riani.it/matlab.htm) under Matlab. We have chosen this method over the fast-MCD algorithm of Rousseeuw and van Driessen (1999) since it is less biased both for small and large samples due to incorporating a more accurate choice of small sample bias-correction factor for the raw MCD obtained by Pison, Van Aelst, and Willems (2002) and an asymptotic adjustment by Hardin and Rocke (2005).

2.2 FSRMCD-Based Bandwidth Selection for Multivariate Density Estimation

Given a dataset $\mathbf{x} = (x_1, \ldots, x_p)'$ selected independently from some underlying distribution with an unknown density function $f$ in a $p$-dimensional Euclidean space, a nonparametric estimate of $f$ at each point $y$ is desired. We have chosen to develop our multivariate density estimation based on the well-known method by Parzen (1962). Throughout the following sections $H$ will denote the bandwidth or window size matrix.

Let the kernel $K : \mathbb{R}^p \to \mathbb{R}$ be a function satisfying certain regularity and integrability properties. The estimator of $f$ at $y$ is then given by

$$\hat{f}_H(y) = \frac{1}{n \det H} \sum_{i = 1}^{n} K \left( H^{-1}(y - x_i) \right).$$

The quality of the estimator depends thus on the choice of the kernel $K$ and the bandwidth $H$. A detailed treatment of this problem can be found in Scott (1992, pp. 125–195). For our purposes, we chose $K$ to be the $p$-variate standard normal density.

Selecting an optimal bandwidth matrix is a more difficult problem than estimating the probability density function $f$ itself and it is especially complex in multivariate settings. Since in our density-based clustering algorithm we are looking for the modes of $f$, that is, points $y$ where local maximum of $f$ is achieved satisfying $\nabla f(y) = 0$ if $f$ is smooth, it is important that a good estimation of $\nabla f$ is achieved. Chacón, Duong, and Wand (2011) established a general theorem that justifies the following choice under the multivariate normality assumption

$$H = \left( \frac{4}{4 + p} \right)^{\frac{p}{2}} n^{-\frac{1}{2p}} \hat{\Sigma}^{1/2}.$$

This is analogous to the usual Scott’s rule of thumb bandwidth for estimating $f$.

Because the underlying distribution can substantially differ from the normal distribution due to a contamination, $\hat{\Sigma}$ should be replaced with a robust covariance estimator. In this article, we use the $\hat{\Sigma}_{\text{FSRMCD}}$ described earlier, which produces a robust normal reference rule of thumb given by Equation (6).

3. FSRMCD-MAC ALGORITHM

3.1 Introduction to the Method

Our approach for identifying multivariate outliers rests on an assumption that a random sample of $n$ $p$-dimensional data vectors comes from a finite mixture of multivariate normal densities with the biggest component representing the bulk of points. All data vectors not in this bulk, whether they are scattered or form clusters, are identified as masking points or potential outliers. The purpose of implementing multivariate probability density
estimation and cluster identification is to be able to distinguish between the bulk and other components of the mixture. We propose to exploit a high breakdown robust estimator to address the oversmoothing effect that usually arises if Scott’s rule-of-thumb bandwidth based on the standard covariance matrix is employed. The MAC method presented by Li, Ray, and Lindsay (2007) uses multiple bandwidths, each a function of the $S$ matrix from the complete set of data to estimate the multivariate probability density function. Even though $S$ has an undesirably low breakdown point, their method is effective in the practical sense of identifying modes and associated clusters. We modify the method of Li, Ray, and Lindsay (2007) by choosing the FSRMCD estimator instead of $S$ because it has a much higher breakdown point and does not lead to oversmoothed density estimators even if the whole set of data is not multivariate normal. We multiply the FSRMCD covariance matrix estimator by a correction factor to form a single bandwidth $H$. Once the filtered central bulk is found, $d_i^2$’s for each of the $n$ data vectors are produced. Based on these $d_i^2$’s, a decision is made as to whether a point is an outlier or not. Details of our proposed outlier detection algorithm follow.

3.2 Steps of the Algorithm

In this section, we describe our FSRMCD-MAC Algorithm.

Step 1. Choose $h = [(n + p + 1)/2]$ and compute FSRMCD estimators $\hat{\mu}_{\text{FSRMCD}}$ and $\hat{\Sigma}_{\text{FSRMCD}}$ according to Equation (4).

Step 2. Similar to Equation (6), define the bandwidth matrix

$$H = \left(\frac{4}{4 + p}\right)^{\frac{1}{p}}n^{-\frac{1}{p+1}} \left(\frac{s_{n,p}}{\hat{\Sigma}_{\text{FSRMCD}}} \right)^{1/2},$$

where $s_{n,p}$ is a correction factor that improves performance. See Table S1 in the supplement. For example, $s_{5,5}^2 = 0.7532 + 37.51n^{-0.9834}$ for $p = 5$ in supplemental Table S1. Use linear interpolation to compute $s_{n,p}$, if the particular $p$ is not given in the table.

Step 3. Estimate the multivariate probability density via $\hat{f}_H$ given by Equation (5) using $H$ from Equation (7) and kernel $K$ as the $p$-variate standard normal density.

Step 4. Apply the Mode Association Clustering (MAC) described by Jobe and Pokojovy (2009) to $\hat{f}_H$. Among the clusters of points from $x$ indexed by $I_1, \ldots, I_c \subset \{1, \ldots, n\}$, the biggest one, say $I_{\text{max}}$, and corresponding mode $m$ are selected. In case of ambiguity, we pick that cluster with a smaller index $i$. If the largest cluster contains less than $p + 1$ points, take $x$ as a single cluster, that is, $I_{\text{max}} = \{1, \ldots, n\}$.

Step 5. For every $i \in I_{\text{max}}$, determine a mirror point of $x_i$ with respect to $m$ by $\tilde{x}_i = 2m - x_i$. Compute a density symmetry measure $\beta(x_i) = \min\{\hat{f}(x_i), \hat{f}(\tilde{x}_i)\}/\max\{\hat{f}(x_i), \hat{f}(\tilde{x}_i)\}$. This will be used to filter out skewness, which may dilute the effectiveness of our method. Note that $0 < \beta(x_i) \leq 1$.

Step 6. All $i \in I_{\text{max}}$ with $\beta(x_i) < \beta_0$ are peeled. The threshold $\beta_0$ is selected as

$$\beta_0 = \min\left\{\frac{\phi_{x_{i_{max}}}(-\beta) + \varphi(-z\sqrt{\text{var}}\phi_{x_{i_{max}}}(-\beta))}{\phi_{x_{i_{max}}}(\beta) + \varphi(z\sqrt{\text{var}}\phi_{x_{i_{max}}}(\beta))}, 1\right\},$$

where $\phi_{x_{i_{max}}}(x)$ is standard $p$-variate normal probability density function,

$$\text{Bias} = \frac{k^2}{2} \phi_{x_{i_{max}}}(\xi) \left(\frac{\alpha_{x_{i_{max}}}(0.95)}{2}\right)^{1/2} - p,\quad \text{var} = \frac{2^{-p} - p/\pi}{\pi n\kappa p} \phi_{x_{i_{max}}}(\xi),$$

$$\kappa = \left(\frac{4}{\pi p}\right)^{\frac{1}{p+1}} \pi n^{-\frac{1}{p+1}} \left(\frac{s_{n,p}}{\hat{\Sigma}_{\text{FSRMCD}}} \right)^{1/2} \text{and } \xi \text{ is an arbitrary point on the hypersphere } \|x\|^2 = \chi^2_{p,0.05}. \text{ Note that } \phi_{x_{i_{max}}}(\xi) = (2\pi)^{-p/2} exp(-\alpha_{x_{i_{max}}}(0.95)/2). \beta_0 \text{ is thus based on a } \alpha \text{-} \text{confidence interval for } \phi_{x_{i_{max}}}(\xi). \text{ Bias and var are the asymptotic bias and variance of } \hat{f}(\xi) \text{ from (5)} \text{ if the underlying density } f(x) \text{ is standard normal and } H = \kappa I_{p \times p}, \kappa > 0 \text{ (see H"ardle et al. 2004, p. 71). Simulations suggested taking } z = 1. \text{ Remaining points are then indexed by } B \subset I_{\text{max}}.$$

Step 7. Based on $\hat{\mu}_B = \frac{1}{|B|} \sum_{i \in B} x_i$, $\hat{\Sigma}_B = \frac{1}{|B| - 1} \sum_{i \in B} (x_i - \hat{\mu}_B) (x_i - \hat{\mu}_B)'$ compute squared Mahalanobis distances $d_i^2_B = (x_i - \hat{\mu}_B)' \hat{\Sigma}_B^{-1} (x_i - \hat{\mu}_B)$ and weights $w_i = 1$, if $d_i^2_B \leq \chi^2_{0.975, p}$, and $w_i = 0$, otherwise.

Step 8. Finally, letting $m = \sum_{i=1}^n w_i$, define

$$\hat{\mu}_{\text{robust}} = \frac{1}{m} \sum_{i=1}^n w_i x_i,$$

$$\hat{\Sigma}_{\text{robust}} = \frac{1}{m-1} \sum_{i=1}^n w_i (x_i - \hat{\mu}_{\text{robust}}) (x_i - \hat{\mu}_{\text{robust}})'$$

and obtain robust squared Mahalanobis distances $d_i^2 = (x_i - \hat{\mu}_{\text{robust}})' \hat{\Sigma}_{\text{robust}}^{-1} (x_i - \hat{\mu}_{\text{robust}}).$

See Section 3.4 for the decision rule using $d_i^2$’s. A Matlab implementation of the algorithm can be found in compute_signals.m file of the supplementary materials.

3.3 Correction Factor, Unbiasedness, and Implementation

To improve performance of the FSRMCD-based bandwidth, we introduced a correction factor $c_{n,p}$, see Equation (7). For $p = 1, 5, 10, 15$, simulations were conducted for different values of $n$ to obtain particular values of the scaling factor that lead to the optimal performance of the FSRMCD-MAC algorithm. For each $p$, the method of least squares was used to fit a curve of the form $\alpha_1 + \alpha_2n^{\alpha_3}$ to observed scaling factors that lead to an optimal performance. The best-fit curves are listed in supplemental Table S1. For example, selecting $p = 5$, $c_{5,5} = 0.7532 + 37.51n^{-0.9834}$.

Since the FSRMCD-MAC-estimators given in Equation (8) are based on $\hat{\mu}_{\text{FSRMCD}}$ and $\hat{\Sigma}_{\text{FSRMCD}}$, which are affine equivariant, $\hat{\mu}_{\text{robust}}$ and $\hat{\Sigma}_{\text{robust}}$ have the property of affine equivariance. Hence, $\hat{\mu}_{\text{robust}}$ is unbiased. On the other hand, simulations that $\hat{\Sigma}_{\text{robust}}$ is slightly biased. This is not essential in the framework of outlier detection. However, if $\hat{\Sigma}_{\text{robust}}$ is used to estimate the covariance, $\hat{\Sigma}_{\text{robust}}$ should be multiplied by a bias correction factor $c_{\text{FSRMCD-MAC}}$ given in supplemental Table S2. For example, choosing $n = 60$ and $p = 5$, $c_{\text{FSRMCD-MAC}} = 1.0407$. 
3.4 FSRMCD-MAC Decision Rule

It is helpful to frame outlier detection in a hypothesis testing mode. The null hypothesis is that no outliers exist, that is, for some \( x = (x_1, \ldots, x_n)' \) all \( x_i \) come from the same distribution and are independent. Thus, \( H_0 : x_i \sim N(\mu, \Sigma) \) for all \( i = 1, \ldots, n \). By accepting \( H_0 \) for every \( i \), we conclude no contamination. Likewise, if at least one of the \( H_0 \) is rejected, at least one \( x_i \) from the data matrix \( x \) is deemed to be an outlier. If each \( H_0 \) is tested individually, we can control the per-comparison error rate by choosing the same test size \( \alpha \) close to 0 for each test. Hence, a high proportion of contaminated \( x_i \) will be detected with a small per-comparison error rate. However, when all \( x_i \sim N(\mu, \Sigma) \) where \( \mu \) and \( \Sigma \) are known, the test statistics are independent and thus the probability at least one \( H_0 \) is rejected becomes \( \gamma = 1 - (1 - \alpha)^n \). This probability is frequently referred to as the family-wise error rate and if the per-comparison error rate \( \alpha \) is controlled to be small, \( \gamma \) can be unacceptably large, perhaps close to 1 even for moderately sized samples. Similar to what Cerioli (2010) proposed, the FSRMCD-MAC decision rule has two testing steps. The first testing step controls the family-wise error rate and the second testing step controls the per-comparison error rate. This two-step testing procedure is much like Fisher’s protected \( F \)-test. These two testing steps are not to be confused with Steps 1 to 8 of the algorithm. Consider each test statistic \( d_{ij}^2 \) given in Equation (9) for \( x = (x_1, \ldots, x_n)' \). In testing Step 1 of the FSRMCD-MAC decision rule, using an identified threshold \( L_{1,1-\gamma} \) that controls the family-wise error rate, if \( \max_{i=1,\ldots,n} d_{ij}^2 > L_{1,1-\gamma} \), at least one \( x_i \) is determined to be an outlier and the family-wise error rate is controlled at a reasonably selected level \( \gamma \). Then, testing Step 2 is executed. If the maximum does not exceed \( L_{1,1-\gamma} \), the FSRMCD-MAC decision rule concludes no outliers exist in the sample \( x \). Consider testing Step 2. Each \( d_{ij}^2 \) given in Equation (9) for \( x \) is compared to a threshold \( L_{2,1-\gamma} \) constructed to control the per-comparison error rate. If \( d_{ij}^2 > L_{2,1-\gamma} \), \( x_i \) is determined to be an outlier, otherwise, \( x_i \) is judged not to be an outlier. The determination of \( L_{1,1-\gamma} \) and \( L_{2,1-\gamma} \) is described below.

For selected \((n, p)\) combinations, 5000 sets of \( n \)-dimensional multivariate standard normal data vectors were generated. Let \( d_{ij}^2 \) be the squared Mahalanobis distance statistic produced by Equation (9) from the \( i \)th item in the \( j \)th sample of size \( n \). For each of the \( j = 1, \ldots, 5000 \) sets of \( n \) data vectors, \( d_{ij}^2 = \max_{i=1,\ldots,n} d_{ij}^2 \) was recorded and \( L_{1,1-\gamma} \) identified as \((1 - \gamma)5000\)th ranked \( d_{ij}^2 \). Jensen, Birch, and Woodall (2007) recommended this approach for determining an initial threshold \( L_{1,1-\gamma} \) that guarantees a family-wise error rate of \( \gamma \) because of dependencies among the \( d_{ij}^2 \) within the \( j \)th set of \( n \) data points. Second, for each of the \( j = 1, \ldots, 5000 \) sets of \( n \) independently generated data vectors, a single \( d_{ij}^2 \) denoted by \( d_{\text{rand},j}^2 \) was randomly selected. Note that since in each set of \( n \) independently generated data vectors, the \( d_{ij}^2 \) are dependent, a single \( d_{ij}^2 \) was randomly selected from each different set. Thus, the threshold \( L_{2,1-\gamma} \) was identified as the \((1 - \gamma)\)th quantile for the distribution of \( d_{\text{rand},j}^2 \) when no outliers exist. See supplemental Table S3 for the limits \( L_1 \) and \( L_2 \). For example, choosing \( n = 60 \) and \( p = 5 \), \( L_{1,0.983} = 33.8628 \) and \( L_{2,0.983} = 18.1691 \).

The simulations outlined here and assessed in Sections 4.1 and 4.2 were carried out on a computer cluster with 32 dual quad-core nodes, GigE connectivity for internode communication, 768 GB of distributed memory, and 15 TB of shared storage using Matlab R2010A under a 64-bit Linux kernel. Pseudo-random numbers were produced using the ziggurat-algorithm implemented in the function \texttt{randn} of Matlab. The maximum average run time over all selected scenarios ranged between 0.78 sec and 7.87 sec depending on \( p \) and \( n \).

4. METHOD EVALUATION AND COMPARISONS

4.1 Outlier Detection Power

We implemented extensive simulations to calculate outlier detection power and swamping rates over a wide range of outlier counts and shift magnitudes for each of the \((n, p)\) combinations in supplemental Table S3. We denoted \( r \) as the count of \( p \)-dimensional “bad” or “outlying” points in a sample of size \( n \) and \( \lambda \) as a shift size indicator. For each of the \( 18 \times 12 = 216 \) combinations, simulations were run for the \( 72 \) \((r, \lambda)\) combinations letting \( r = \tau n \), where \( \tau = 0.05, 0.10, \ldots, 0.30 \) and \( \lambda = 1.6, 1.8, \ldots, 3.8 \). For a given \((n, p)\) combination, each of the \( 72 \) simulations generated 5000 independent sets of \( n \)-dimensional random vectors. In a given set of random vectors, \( n - r \) of the vectors were generated from \( N(0_{p \times 1}, I_{p \times p}) \) and \( r \) of the vectors were generated from \( N(\lambda e_{p \times 1}, I_{p \times p}) \), where \( e_{p \times 1} \) is a \( p \times 1 \) vector of ones and a \( p \times p \) identity covariance matrix \( I_{p \times p} \). Using the two-step FSRMCD-MAC method and Cerioli’s IRMCD method, the percent of \( r \) vectors identified as outliers and percent of the \( n - r \) vectors incorrectly determined to be outliers were recorded for each set of \( n \)-dimensional random vectors. The average percent of \( r \) vectors identified as outliers based on 5000 sets of \( n \)-dimensional vectors is the detection power for selected \((n, p, r, \lambda)\). Likewise, swamping rate is the average percent of \( n - r \) vectors incorrectly identified as outliers based on the same 5000 sets. So, \( 18 \times 6 \times 12 = 1296 \) simulations, each of size 5000, were performed for both methods producing corresponding swamping and detection power rates. Choosing the mean vector \( \lambda e_{p \times 1} \) is our way of identifying a shift in the distribution from a mean vector of \( 0_{p \times 1} \). The noncentrality parameter \( \text{ncp} = (\mu_1 - \mu_0)' \Sigma^{-1} (\mu_1 - \mu_0) \) is sometimes used to index a shift producing outliers. Here, \( \mu_0, \mu_1, \Sigma = \Sigma_0 = \Sigma_1 \) are the mean vectors and covariance matrices of the first and second group of points. It can be easily seen that \( \text{ncp} = \lambda^2 p \). So, for a given \( p \), a large \( \lambda \) corresponds to a large \( \text{ncp} \), that is, a large shift.

Since we desired to compare the FSRMCD-MAC outlier detection performance with that of the IRMCD method Cerioli (2010) proposed, our comparisons were based on the same family-wise error rate denoted as \( \gamma \), where \( \gamma \) is close to zero. Cerioli (2010) noted the published IRMCD performance used a stated \( \gamma = 0.01 \). For notational purposes, we refer to this \( \gamma \) as \( \gamma_{\text{nomin}} \). Because the IRMCD method is based on approximate probability distributions and the test statistics are not independent, the actual family-wise error rate \( \gamma_{\text{actual}} \) for the IRMCD method is usually larger than \( \gamma_{\text{nomin}} = 0.01 \). Based on simulations, the actual \( \gamma_{\text{actual}} \) values for a stated \( \gamma_{\text{nomin}} = 0.01 \) are given in our Table 1 and can also be found in Table 1 of Cerioli’s article. To legitimately compare outlier detection performances of the FSRMCD-MAC to the IRMCD method, \( \gamma_{\text{actual}} \) from Table 1 for the FSRMCD-MAC method and \( \gamma_{\text{nomin}} = 0.01 \) for
the IRMCD method were chosen. This ensures both have the same $\gamma_{\text{actual}}$ family-wise error probability. Note that for $p = 15$ and $n = 40$, we discovered $\gamma_{\text{actual}} = 0.105$ where Cerioli 2010 reported 0.084.

Consider two example comparisons for $n = 60$ and $p = 5$. Letting $\gamma_{\text{nominal}} = 0.01$, the actual family-wise error rate used in the FSRMCD-MAC method was $\gamma_{\text{actual}} = 0.017$. Notationally, in each configuration $(1 - \tau)n$ observations come from $N(0_{5 \times 1}, I_{5 \times 5})$ and $\tau n$ from $N(\lambda e_{5 \times 1}, I_{5 \times 5})$. Figure 1(a) illustrates the percent of outliers detected by the FSRMCD-MAC and IRMCD methods for $\tau = 0.05$ and selected $\lambda$’s. The swamping rates never exceeded 1.5% for FSRMCD-MAC and 1% for IRMCD. Figure 1(b) illustrates the percent of outliers detected by the FSRMCD-MAC and IRMCD methods for $\tau = 0.20$ and selected $\lambda$’s. The swamping rates never exceeded 2.6% for FSRMCD-MAC and remained below 1% for IRMCD. Hence, in both scenarios the power improvements due to implementation of the FSRMCD-MAC method were consistent and significant. Although the FSRMCD-MAC method produced swamping rates up to 2.6% compared to 1% for the IRMCD method, both methods had the same experiment-wise error rate. Since swamping rates only have meaning when at least one outlier has been signaled, a method with a somewhat larger swamping rate but consistently and significantly higher outlier detection probability may be preferred by analysts. We thoroughly address the logic of this choice later in Section 4.2.

Continuing, two example comparisons for $n = 200$ and $p = 5$ were also evaluated. As before, letting $\gamma_{\text{nominal}} = 0.01$ the family-wise error rate used in the FSRMCD-MAC method was $\gamma_{\text{actual}} = 0.011$. Figure 2(a) illustrates the percent of outliers detected by the FSRMCD-MAC and IRMCD methods for $\tau = 0.05$ and selected $\lambda$’s. The swamping rates never exceeded 1.2% for FSRMCD-MAC and 1% for IRMCD. Figure 2(b) illustrates the percent of outliers detected by the FSRMCD-MAC and IRMCD methods for $\tau = 0.20$ and selected $\lambda$’s. The swamping rates never exceeded 1.3% for FSRMCD-MAC and remained below 1% for IRMCD. Cerioli (2010) also evaluated IRMCD performance for outlier configurations where outliers were scattered from central to more extreme locations. In the following two examples, uncontaminated data were generated from $N(0_{5 \times 1}, I_{5 \times 5})$. For $n = 200$, $p = 5$, and $\gamma_{\text{nominal}} = 0.01$, the first example generated contaminated data from $N(0_{5 \times 1}, \psi I_{5 \times 5})$, $\psi = 2, 4, \ldots, 12$, and $\tau = 0.20$.

Figure 3(a) illustrates the outlier detection performance of the FSRMCD-MAC and IRMCD methods. The actual family-wise error rate used in the FSRMCD-MAC method was $\gamma_{\text{actual}} = 0.011$. The swamping rates for the FSRMCD-MAC method did not exceed 1.6% and that of the IRMCD method did not exceed 1%. Second, choosing $n = 200$, $p = 5$, $\gamma_{\text{nominal}} = 0.01$, and $\tau = 0.20$, we considered contaminated data coming from a $p = 5$ dimensional multivariate $t$-distribution with $\xi$ degrees of freedom, $\zeta = 1, 2, \ldots, 6$. Figure 3(b) illustrates the outlier detection performance of the FSRMCD-MAC and IRMCD methods for this second example. The actual family-wise error rate used in the FSRMCD-MAC method was $\gamma_{\text{actual}} = 0.011$. Swamping rates for the FSRMCD-MAC method did not exceed 1.3% and that of the IRMCD method did not exceed 1%.

### 4.2 Outlier Detection Performance Comparisons for Additional $(n, p)$ Combinations

In this section, outlier detection performances of IRMCD and FSRMCD-MAC methods for the complete set of 18 $(n, p)$ combinations outlined in Table 1 are compared and analyzed. For all $(n, p)$ combinations except for those where $p = 15$, our comparisons are based on the outlier detection performances over the 72 $(r, \lambda)$ combinations outlined earlier. Since for large $\lambda$’s we noticed both methods almost always produced a proportion of detected outliers equal to 1, we restricted our comparisons to those whose noncentrality parameter (ncp) did not exceed 15.

| $n$ | $\gamma_{\text{actual}}$ | $\gamma_{\text{actual}}$ | $\gamma_{\text{actual}}$ | $\gamma_{\text{actual}}$ |
|-----|-----------------|-----------------|-----------------|-----------------|
| 40  | 0.017           | 0.017           | 0.013           | 0.011           |
| 60  | 0.025           | 0.014           | 0.012           | 0.012           |
| 90  | 0.030           | 0.013           | 0.012           | 0.008           |
| 125 | 0.030           | 0.013           | 0.012           | 0.008           |
| 200 | 0.030           | 0.013           | 0.012           | 0.008           |
| 400 | 0.030           | 0.013           | 0.012           | 0.008           |

Table 1. Estimated size $\gamma_{\text{actual}}$ of the IRMCD rule corresponding to $\gamma_{\text{nominal}} = 0.01$.
Figure 2. Power of FSRMCD-MAC and IRMCD tests under a multivariate location-shift contamination model, for $p = 5$ and $n = 200$ where $\gamma_{actual} = 0.011$.

Figure 3. Power of FSRMCD-MAC and IRMCD tests under radial contamination (a) and multivariate $t$ contamination (b), for $p = 5$, $n = 200$, and $\tau = 0.20$ where $\gamma_{actual} = 0.011$.

Consider now one of the scenarios ($n = 40$, $p = 10$) where the FSRMCD-MAC method was not almost always uniformly better with regards to detection power. The outlier detection power was uniformly higher for FSRMCD-MAC when contamination was 25% and 30% ($r = 10$ or 12). If the contamination levels were 10% or 15%, outlier detection power of IRMCD was uniformly higher. When contamination levels were 5% or 20%, at some values of $\lambda$ the detection power values favored FSRMCD-MAC and for other $\lambda$’s the IRMCD method was superior. Swamping rates for the IRMCD method remained at or slightly lower than 1% for almost all $(r, \lambda)$ pairs. The swamping rates for FSRMCD-MAC approach remained at or below 2% for all $r$ and smaller $\lambda$’s. For $\lambda$ larger than 2.0 and all $r$, the swamping rates for FSRMCD-MAC were mostly from 2% to 8% with a few in excess of 10%. The other scenario ($n = 60$, $p = 15$) where the power detection levels from the FSRMCD-MAC approach were not uniformly larger than those produced by the IRMCD can be summarized as follows. For 5%, 25%, and 30% contamination levels and almost all $\lambda$, power detection levels were almost all larger from the FSRMCD-MAC method.
and sometimes much larger than those produced by the IRMCD approach. However, for 10%, 15%, and 20% contamination levels and most λ levels, detection power values from the IRMCD method were larger. Almost all swamping rates for the IRMCD method were less than 1.2%. FSRMCD-MAC swamping rates stayed below 3.9% when contamination levels were from 5% to 20%. Five FSRMCD-MAC swamping rates exceeded 15% and three were from 5% to 8% when contamination levels were at 25% or 30%. Other FSRMCD-MAC swamping rates were under 4%.

If only detection power determines the outlier detection method to use, FSRMCD-MAC is a reasonable choice for 16 of the 18 (n, p) pairs noted above. A natural question arises from this summary. Which outlier detection method should be chosen when both swamping and detection power are important? For n = 40, 60, 90, 125, and 200 and p = 5, a strong case can be made for choosing the FSRMCD-MAC approach because although its swamping rates were somewhat larger than those of IRMCD, detection power levels of the FSRMCD-MAC were much larger than those of IRMCD. See Figures 1 and 2. Of course, this is somewhat subjective. How one weights swamping versus detection power determines which method to choose.

We propose a model to guide which method to use for outlier detection that weighs swamping and detection power in a proportional fashion.

Let c equal the positive loss of identifying a good item as bad or an outlier. Let mult × c equal the positive gain or savings of identifying a bad or outlying item as being bad or an outlier and mult > 0. This is the notion of proportionality we choose to use in modeling the savings/loss aspect of outlier detection. Vardeman and Jobe (1999, p. 9) stated an "engineering rule of 10" that is often used in decision making related to repair versus detection power determines which method to choose.

For a selected (n, p) pair, based on 5000 simulations of the FSRMCD-MAC method, we let Po_i and So_i be the observed proportion of contaminated items correctly detected and non-contaminated items identified as contaminated for the ith (r, λ) pair, i = 1, 2, . . . , cnt. Similarly, based on 5000 simulations of the IRMCD method, we let PI_i be the observed proportion of contaminated items correctly detected and SI_i the proportion of non-contaminated items identified as contaminated. For all n, when p = 5 or 10, cnt = 72 and when p = 15, cnt = 48. Letting r denote the number of bad or contaminated or outlying items from a sample of n items and n − r the number of good or uncontaminated items, an estimated per item savings equation corresponding to application of the FSRMCD-MAC method for a given (n, p, r, λ) can be written as

\[ \text{Po}_i \times \frac{r}{n} \times c \times \text{mult} - \text{So}_i \times \frac{n-r}{n} \times c, \quad i = 1, \ldots, \text{cnt}. \]

The estimated change in per item savings \( \Delta \text{SVG}_i \), FSRMCD-MAC approach minus IRMCD method for the ith (r, λ) pair from a selected (n, p) pair becomes

\[ \Delta \text{SVG}_i = (\text{Po}_i - \text{PI}_i) \times \frac{r}{n} \times c \times \text{mult} \]

Since comparing \( \Delta \text{SVG} \) to 0 informs which method should be used for outlier detection, we let without loss of generality \( c = 1 \). If \( \Delta \text{SVG} \) exceeds 0, then for the ith (r, λ) combination FSRMCD-MAC is preferred over IRMCD. If \( \Delta \text{SVG} \) is less than zero, IRMCD is preferred for the ith (r, λ) combination. In practice, however, r and λ are unknown. To address this matter, for each (n, p) pair considered, we propose finding the estimated average based on the cnt \( \Delta \text{SVG}_i \)’s and a 90% two-sided confidence interval for the average \( \Delta \text{SVG} \). For each (n, p), the cnt \( \Delta \text{SVG}_i \)’s come from a preselected grid equally spaced over a selected (r, λ) space. We assume the preselected grid of cnt points comes from a population of (r, λ)’s that occur uniformly over a two-dimensional space \([1, 2, \ldots, \lceil 0.3n \rceil] \times [1.6, \sqrt{150}/\bar{p}]\). The sample average \( \Delta \text{SVG} \) becomes

\[ \Delta \text{SVG} = \frac{1}{\text{cnt}} \sum_{i=1}^{\text{cnt}} \Delta \text{SVG}_i. \]

Or, defining \( \bar{r} = (-1, \text{mult}) \), \( \bar{d}_r = (d_1, d_2) \), \( d_1 = (\text{So}_1 - \text{SI}_1) \times \frac{n-r}{n}, \quad d_2 = (\text{Po}_1 - \text{PI}_1) \times \frac{r}{n}, \quad i = 1, 2, \ldots, \text{cnt}, \)

\[ \Delta \text{SVG} = \bar{r} \hat{d}, \]

where \( \hat{d} \) is the estimated mean vector based on the cnt column vectors \( d_r \). Continuing, the 90% confidence interval for the true average \( \Delta \text{SVG} \) becomes

\[ \bar{r} \hat{d} \pm (\Phi^{-1}(0.95)/\sqrt{\text{cnt}}) \sqrt{\text{TI} \bar{d} \hat{d}} \]

where \( \Phi^{-1}(0.95) \approx 1.645 \) is the 95th percentile of the standard normal distribution and \( \text{S} \) is the estimated covariance matrix based on the cnt column vectors \( d_r \). For each of the 18 (n, p) pairs described above, we identified the range of mult values that make the lower bound exceed 0. These mult values inform the user to implement FSRMCD-MAC. Likewise, we identified the range of mult values that makes the upper bound less than 0. These mult values inform the user to implement IRMCD. Values of mult that guarantee a negative lower bound and positive upper bound were also found. These values inform the analyst using FSRMCD-MAC instead of IRMCD provides no advantage. Supplemental Table S4 displays our findings. For the pair (n = 60, p = 10), mult_{FSRMCD-MAC} = 1.1209 and mult_{IRMCD} = 0.8678. If the mult value appropriate for the application exceeds 1.1209, the FSRMCD-MAC method should be chosen. Likewise, for mult values below 0.8678, the IRMCD method should be chosen, and for mult values between 0.8678 and 1.1209, neither FSRMCD-MAC or IRMCD perform better on the average. For all (n, p) combinations considered, if the rule of 10 holds, the FSRMCD-MAC is clearly preferred. Practically, it seems mult is almost always much larger than 1 suggesting FSRMCD-MAC is to be used for most cases considered.

For smaller λ values, especially those less than 3, our proportional decision model shows FSRMCD-MAC method to frequently produce larger savings per item for individual applications than those from IRMCD for certain ranges of mult. Additional work is needed to explore and analyze the change in savings per item from any one application of FSRMCD-MAC compared to IRMCD.
5. DATA ANALYSIS

In this section, we compare outlier detection performance of the FSRMCD-MAC method to that of the IRMCD method based on application to the same data. The data we selected for this comparison were introduced by Flury and Riedwyl (1988). Six variables were measured on each of 200 Swiss banknotes, 100 being authentic and 100 forged.

We now turn to a contextually interesting question: How well does each outlier detection method perform when it comes to separating genuine from forged bills? We supposed the configuration of genuine to forged to be 100 to 100, that is, 200 genuine notes with 100 genuine and 100 forged. Since the number of bills belonging to the genuine group is 100 and \( h = [(n + p + 1)/2] = (200 + 6 + 1)/2 = 103 \), MCD estimators do breakdown. Another way to evaluate whether MCD estimators breakdown is to consider the breakdown point. In this example, the number of counterfeit bills is 100, which is more than the breakdown point given by \([n - p + 1)/2] = (200 - 6 + 1)/2 = 97\). Thus, assumptions regarding breakdown are violated and the MCD estimators used by both methods breakdown. Using the simulated \( \gamma_{\text{actual}} = 0.0118 \), the FSRMCD-MAC analysis is presented in Figure 4(a). Likewise, using \( \gamma_{\text{nominal}} = 0.01 \), the IRMCD analysis is presented in Figure 4(b). Note: in Figure 4, index = 1, \ldots, 100 corresponds to genuine notes and index = 101, \ldots, 200 corresponds to forged notes.

Both methods have the same swamping rate. FSRMCD-MAC incorrectly identifies five genuine notes as counterfeit and IRMCD incorrectly identifies five genuine notes as counterfeit. However, the detection power of the FSRMCD-MAC is 100 out of 100 whereas that of the IRMCD method is 15 out of 100. Even though both methods employ the reweighted MCD estimators, the FSRMCD-MAC method incorporates the reweighted MCD to build an estimated multivariate probability density function used by the cluster analysis to find a central bulk of data upon which to find a mean vector and covariance matrix to be used for calculating a robust test statistic. Equation (9) gives this statistic. It seems the FSRMCD-MAC method benefits from exploiting the reweighted MCD estimator as a bandwidth matrix to be used by the MEM and MAC algorithms to construct a cluster analysis helpful in filtering out unusual data values. In contrast, IRMCD uses the MCD estimator to directly construct a test statistic.

6. CONCLUSION

We have constructed and presented the FSRMCD-MAC algorithm for detecting multiple outliers. The FSRMCD-MAC algorithm was shown to detect with high probability multiple outliers that may occur among multivariate normal data. Similar to Cerioli (2010), FSRMCD-MAC uses two testing steps to identify outliers. The first testing step assures a selected family-wise error rate and the second testing step maintains a selected per-comparison error rate. For a family-wise error rate corresponding to \( \gamma_{\text{nominal}} = 0.01 \), comparisons of the IRMCD and FSRMCD-MAC outlier detection ability based on extensive simulations show the FSRMCD-MAC method regularly outperforms the power of IRMCD, the leading multivariate outlier detection method proposed by Cerioli (2010), for a wide range of \((n, p, r, \lambda)\) combinations. Further, for many of the \((n, p, r, \lambda)\) combinations the FSRMCD-MAC method produces swamping rates that are quite competitive with those produced by IRMCD.

We proposed a model to guide which method to use for outlier detection that weights swamping and detection power in a proportional fashion. Supplemental Table S4 summarizes our decision rule for choosing FSRMCD-MAC or IRMCD. This rule uses a multiplicative model that is a function of the savings and loss that occur as a result of outlier detection decisions. Across the grid represented in Supplemental Table S4, as long as the savings from detecting an outlier is more than 4.5 times the loss of identifying a good item as an outlier, FSRMCD-MAC is preferred for the 18 scenarios considered. If the ratio is at least 2.02, FSRMCD-MAC is preferred for 15 of the 18 scenarios. The Swiss banknote example presented in Section 5 demonstrated the strength of the FSRMCD-MAC method over the IRMCD approach in a practical framework. Our simulation method for determining cut-off limits required in the FSRMCD-MAC two testing steps produced decision-rule limits that are accurate and not dependent on asymptotic assumptions. Further, our method overcomes dependencies not accounted for by the IRMCD method.
7. SUPPLEMENTARY MATERIALS

Contents of the supplementary files include a zip-archive of the Matlab files and a PDF with four additional tables. The zip-archive contains Matlab files implementing the FSRMCD-MAC algorithm as well as auxiliary routines. See each separate file for a detailed description. The tables referred to as Tables S1 through S4 in the text of the article are simply labeled as Tables 1 through 4 in the supplement.

[Received February 2013. Revised September 2014.]

REFERENCES

Cerioli, A. (2010), “Multivariate Outlier Detection With High-Breakdown Estimators,” Journal of the American Statistical Association, 105, 147–156. [1543,1544,1546,1547,1550]

Chacón, J. E., Duong, T., and Wand, M. P. (2011), “Asymptotics for General Multivariate Kernel Density Derivative Estimators,” Statistica Sinica, 21, 807–840. [1544]

Croux, C., and Haesbroeck, G. (1999), “Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator,” Journal of Multivariate Analysis, 71, 161–190. [1544]

Davies, P. L. (1987), “Asymptotic Behaviour of S-Estimates of Multivariate Location Parameter and Dispersion Matrices,” The Annals of Statistics, 15, 1269–1292. [1544]

Flury, B., and Riedwyl, H. (1988), Multivariate Statistics: A Practical Approach, London: Chapman & Hall. [1544,1550]

Hardin, J., and Rocke, D. M. (2005), “The Distribution of Robust Distances,” Journal of Computational and Graphical Statistics, 14, 910–927. [1544]

Härdele, W. K., Müller, M., Sperlich, S., and Werwatz, A. (2004), Nonparametric and Semiparametric Models, Berlin, Heidelberg: Springer. [1545]

Hawkins, D. M., and Olive, D. J. (1999), “Improved Feasible Solution Algorithms for High Breakdown Estimation,” Computational Statistics and Data Analysis, 30, 1–11. [1543]

Jensen, W. A., Birch, J. B., and Woodall, W. H. (2007), “High Breakdown Estimation Methods for Phase I Multivariate Control Charts,” Quality and Reliability Engineering International, 23, 615–629. [1546]

Jobe, J. M., and Pokojovy, M. (2009), “A Multistep, Cluster-Based Multivariate Chart for Retrospective Monitoring of Individuals,” Journal of Quality Technology, 41, 323–339. [1543,1545]

Li, J., Ray, S., and Lindsay, B. G. (2007), “A Nonparametric Statistical Approach to Clustering via Mode Identification,” Journal of Machine Learning Research, 8, 1687–1723. [1543,1545]

Lopuhai, H. P., and Rousseeuw, P. J. (1991), “Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices,” The Annals of Statistics, 19, 229–248. [1544]

Parzen, E. (1962), “On Estimation of Probability Density Function and Mode,” Annals of Mathematical Statistics, 33, 1065–1076. [1543,1544]

Pison, G., Van Aelst, S., and Willems, G. (2002), “Small Sample Corrections for LTS and MCD,” Metrika, 44, 111–123. [1544]

Rousseeuw, P. J., and Leroy, A. (1987), Robust Regression and Outlier Detection, New York: Wiley. [1544]

Rousseeuw, P. J., and van Driessen, K. (1999), “A Fast Algorithm for the Minimum Covariance Determinant Estimator,” Technometrics, 41, 212–223. [1543,1544]

Scott, D. W. (1992), Multivariate Density Estimation: Theory, Practice and Visualization, Hoboken, NJ: Wiley. [1544]

Tracy, N. D., Young, J. C., and Mason, R. L. (1992), “Multivariate Control Charts for Individual Observations,” Journal of Quality Technology, 24, 88–95. [1543]

Vardeman, S. B., and Jobe, J. M. (1999), Statistical Quality Assurance Methods for Engineers, New York: Wiley. [1549]

Wierda, S. J. (1994), “Multivariate Statistical Process Control—Recent Results and Directions for Future Research,” Statistica Neerlandica, 48, 147–168. [1543]