The Dissertation Committee for Paul Muse Ritter
certifies that this is the approved version of the following dissertation:

A Comparative Study of Correlational Outlier Detection Metrics

Committee:

Susan Natasha Beretvas, Supervisor

Edmund Emmer

Keenan Pituch

Tiffany A. Whittaker

Peter W.M. John

A Comparative Study of Correlational Outlier Detection Metrics

by

Paul Muse Ritter, B.A., B.S., M.S.STAT

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2008

To my sister, Lucy Ritter Todd, without whose help my comeback would have never been possible.

A Comparative Study of Correlational Outlier Detection Metrics

Publication No. _____

Paul Muse Ritter, Ph.D.
The University of Texas at Austin, 2008

Supervisor: Susan Natasha Beretvas

The present investigation was a Monte Carlo experiment designed to evaluate the performance of several metrics in spotting correlational outliers. Specifically, the metrics that were compared were the Mahalanobis D^2, Bacon MLD, Carrig D, MCD, Robust PCLOW and Robust PCHIGH. This was the first comparative simulation study to include robust PCLOW and robust PCHIGH. The Mahalanobis D^2, MCD, Robust PCLOW and Robust PCHIGH were each applied using an approximate statistical criterion. The Carrig D and Bacon MLD were applied using a "natural drop "approach that separated scores on the metric into two groups: outlying and non-outlying. The "natural drop "utilizes a k-means algorithm from cluster analysis to separate the scores into the two groups.

Both majority and contaminant observations were generated from multivariate normal distributions based on factor-analytic models. Experimental factors included majority versus contaminant communality level, majority-contaminant factor models scenario, number of variables, sample size and fraction of outliers.

Results indicated that the "natural drop "method of application for the Carrig D and Bacon MLD leads to intolerably high false-alarm rates. Overall, PCLOW clearly outperformed PCHIGH. Suprisingly, PCLOW did not distinguish itself from MCD in terms of performance as expected in certain experimental conditions.

The conditions in this study were limited. Future comparative studies of the metrics could include conditions of non-normality and hybrid types of outliers (i.e. outliers that are both mean shift and correlational). Despite its poor performance in this study, I theorize that robust PCHIGH could have an advantage over MCD in spotting certain kinds of mean-shift outliers. Also, research into the distributional properties of the Carrig D is warranted.

Contents

Chapter 1

Introduction

The subject of outliers is troublesome and ill-defined in univariate analyses. It is even more troublesome and ill-defined when one moves to multivariate analyses. Gnanadesikan and Kettenring highlight one of the additional complexities in the multivariate situation: "A single univariate outlier may be typically thought of as 'the one that sticks out on the end 'but no such simple idea suffices in higher dimensions. "(1972) There are an infinity of "ends "in the multivariate case, and they cannot be visualized in the case where the number of variables is greater than three. The most common method of overcoming this dimensionality problem is to devise a function that takes each multivariate observation in a sample and maps it to a scalar that presumably indicates the degree to which the observation is outlying. Obviously, information will be lost when this method is employed. Also researchers have pointed out that an observation can be outlying in different ways. For example, an observation can be outlying with respect to the multivariate mean without being outlying with respect to the correlational structure, or it can be outlying with respect to the multivariate mean and the correlational structure. Thus, there are different types of multivariate outliers. Will a function that spots one type of outlier spot another type?

It should be mentioned that in some situations one can avoid the problem of detecting multivariate outliers altogether. The analysis can "accommodate "the presence of outliers. In other words, the analysis will be resistant to the inordinate influence that the outliers would normally have. This study will not touch on the topics of influence and robust methods of analysis. It will concentrate on the problem of identification only. One may ask what is the point in detecting outliers if robust analyses can be performed? The answer is that there are reasons other than the concern over influence that motivate one to detect outliers. In their landmark *Technometrics* article, Beckman and Cook provided five reasons for the interest in detecting outliers: special interest, detection of specific alternate phenomena, diagnostic indication, influence and accommodation (1983). Thus, acccommodation does not eliminate the interest in detecting outliers.

The complexity of multivariate outliers notwithstanding, detection methods have been developed. With the exception of the practice of looking at all possible bivariate and trivariate plots, detection methods can be divided into those that are purely quantitative and those that are both quantitative and graphical in nature. All purely quantitative methods involve the mapping of each observation to a scalar and the determination of whether this value exceeds some pre-determined cut-off value. Purely quantitative methods include bona-fide hypothesis tests known as formal tests of discordancy. The alternative hypotheis in these tests of hypotheses is that a single observation or a group of observations came from a distribution that is different from the single distribution for the rest of the observations (Barnett & Lewis, 1984; Hawkins, 1980). Formal tests of discordancy are rarely used because the sampling distributions of the statistics involved are highly complex. Methods that I classify as informal and purely quantitative are not bona-fide tests of hypotheses; observed values on the metric associated with each observation are compared to a cut-off value or are plotted and a subjective determination of which ones are outlying

is made. Methods that are both quantitative and graphical in nature are classified as informal as well. This study will focus on informal detection methods exclusively.

There are fairly well-known examples of informal quantitative detection metrics in univariate analyses. Specifically, there is the practice of flagging observations in an univariate sample that are more than 1.5 interquartile ranges beyond the first or third quartile. Another univariate example is the guideline that standardized scores exceeding a certain cut-off value (usually 3.00) should be flagged as potential outliers (this practice assumes that the data came from a normally distributed population). Neither one of these practices can be considered a formal test of discordancy. Consider the second practice where the standardized scores are compared against a cut-off of 3.00. Granted, under the assumption of normality, $P(X > 3.00) \doteq 0.0013$, but this does not mean that the test assesses whether a single observation is a contaminant at the $\alpha = 0.0013$ level even though this is what some practitioners believe. This is not the case because what we are dealing with are actually order statistics. This will be explained more fully, later.

The Mahalanobis D^2 is by far the best known metric that is employed informally to detect multivariate outliers. Other, fairly well-known metrics employed over the years include the diagonal elements of the Hat Matrix and Mardia's multivariate kurtosis measure. However, both of these metrics have been shown to be monotonic functions of the Mahalanobis D^2. Thus, they contribute little over and above the Mahalanobis D^2 in terms of use as an informal quantitative outlier detection metric (Bacon, 1995).

Gnanadesikan and Kettenring expressed two concerns regarding the Mahalanobis D^2: it is susceptible to masking and swamping effects and that it is only useful in detecting mean-shift outliers (1972). Two separate and distinct lines of research grew out of these two concerns. Let us consider the second concern first.

A constant theme in the literature is that different detection metrics should

be tailored to detect different kinds of outliers (Gnanadesikan & Kettenring, 1972). Among the types of outliers identified are mean-shift outliers and correlational outliers. Mean-shift outliers come from a distribution that has a different population mean vector but has the same population covariance matrix in comparison to the legitimate population. A purely correlational outlier comes from a population whose mean vector is the same as that for the legitimate population but has a different population correlation matrix. A specific example of the second concern expressed by Gnanadesikan and Kettenring is that the Mahalanobis D^2 does not spot correlational outliers (Comrey, 1985).

Three methods other than the Mahalanobis D^2 have been specifically developed to detect these correlational outliers: Comrey's D (Comrey, 1985), Bacon's MLD (Bacon, 1995) and Carrig's D (Carrig, 2005). Simulation studies comparing the performance of the Mahalanobis D^2 and Comrey's D in detecting correlational outliers suggest that the Comrey's D does not have an advantage over the Mahalanobis D^2 in the detection of correlational outliers (Rassmussen, 1988; Bacon, 1995). However, simulation results do suggest that Bacon's MLD does offer something over and above the Mahalanobis D^2 in the detection of correlational outliers (Bacon, 1995; Carrig, 2005). Now we shall consider "robust "distance metrics that are designed to be resistant to the masking and swamping effects.

Several robust distance metrics that are based on robust estimates of the mean vector and covariance matrix have been developed: the ellipsoidal multivariate trimming procedure (Gnanadesikan & Kettenring, 1972, MVT), minimum volume ellipsoid (Rousseeuw, 1985, MVE), minimum covariance determinant (Rousseeuw, 1985, MCD), smallest half volume (Egan & Morgan, 1998, SHV) and resampling by half means (Egan & Morgan, 1998, RHM). Simulation studies have shown that all of these robust distance metrics perform substantially better than the Mahalanobis D^2 when the fraction of outliers in the sample is relatively large (Gnanadesikan

& Kettenring, 1972; Rousseeuw & Zomeren, 1990; Egan & Morgan, 1998; Carrig, 2005).

A note should be made at this point regarding the use of the term "outlier. "To many practicing statisticians the term "outlier "implies "a few "discordant observations and not the large number that are present when testing the robust distance metrics. When we have a large number of discordant observations in a data set, the subject at hand may be more properly characterized as involving "mixture distribution "than as an "outlier "problem. In the case where there are a large number of discordant observations, I shall use the term "contaminant "rather than "outlier. "

Of note is the fact that the research into metrics that detect correlational outliers and the research into robust distance metrics were never connected until Carrig's Ph.D. thesis in 2005. She proposed a metric, Carrig's D which combined research from the hitherto disparate lines. Additionally, her study was the first and only one to compare metrics designed to detect correlational outliers head to head with some of the robust distance metrics. The experimental conditions in her study included samples where the outliers were of the mean-shift variety, samples in which they were correlational and samples where they were a mixture of the two. No single detection metric nor combination of metrics was best across the experimental conditions. A major theme in the multivariate outlier detection research that there is no truly omnibus detection method seemed to be confirmed (Carrig, 2005).

However, there is a thread in the research that has been left dangling. That thread is the use of principal components to detect multivariate outliers. Rao suggested that multivariate outliers would be revealed by looking at projections of observations on the last few principal components (1964). Later, Hawkins added the refinement that the projections should be weighted by the corresponding variances of the principal components (1974). In his 1986 book, *Principal Components*

Analysis, Jolliffe asserted that the metric suggested by Hawkins would be useful in detecting correlational outliers. This metric has never been given a name—I shall call it PCLOW—nor has it been tested against other metrics in a simulation study. Contrastingly, there is the suggestion by Gnanadesikan and Kettenring that observations should be plotted against the first principal components to spot outliers. They also contend that observations that are outlying in this respect will tend to distort correlation estimates. A quantitative detection method exactly like Hawkin's except that it is computed from the first few rather than the last few principal components can be inferred from this graphical suggestion. I shall call this metric PCHIGH. Thus, there are two seemingly contradictory suggestions as to how to spot correlational outliers. Which metric, PCLOW or PCHIGH, is more useful in spotting correlational outliers? Does PCLOW offer anything over and above PCLOW? Are some metrics more robust than others to changes in the difference between communalities for the legitimate and contaminant population? Are some metrics more robust than others to changes in the n/p ratio? Also, will the results of Carrig's comparative study be repeated under the conditions of my study?

Chapter 2

Review of the Literature

2.1 What is an Outlier?

The concept of an outlier, whether in the univariate or the multivariate case, has always been somewhat nebulous. Beckman and Cook state: "Although much has been written, the notion of an outlier seems as a vague today as it was 200 years ago "(1983, p. 120). What follows are definitions supplied by present-day authorities on the subject of outliers. Barnett and Lewis: "We shall define an outlier in a set of data to be an observation (or subset of observations) which appear(s) to be inconsistent with the remainder of that set of data "(1984, p. 4). Hawkins states: "The intuitive definition of an outlier would be 'an observation which deviates so much from the other observations as to arouse suspicions that it was generated by different mechanisms '"(1980, p. 1). Both of these definitions betray the vagueness to which Beckman and Cook alluded. Despite this vagueness of definition and the subjectivity involved in labeling an observation an outlier, some univariate observations can stand out in a graphical display sufficiently, to compel virtually any observer to them as outliers. Consider Figure 2.1, a histogram of on-base percentages for the 2002 Major League Baseball season (Verzani, 2005). It is hard not to notice the

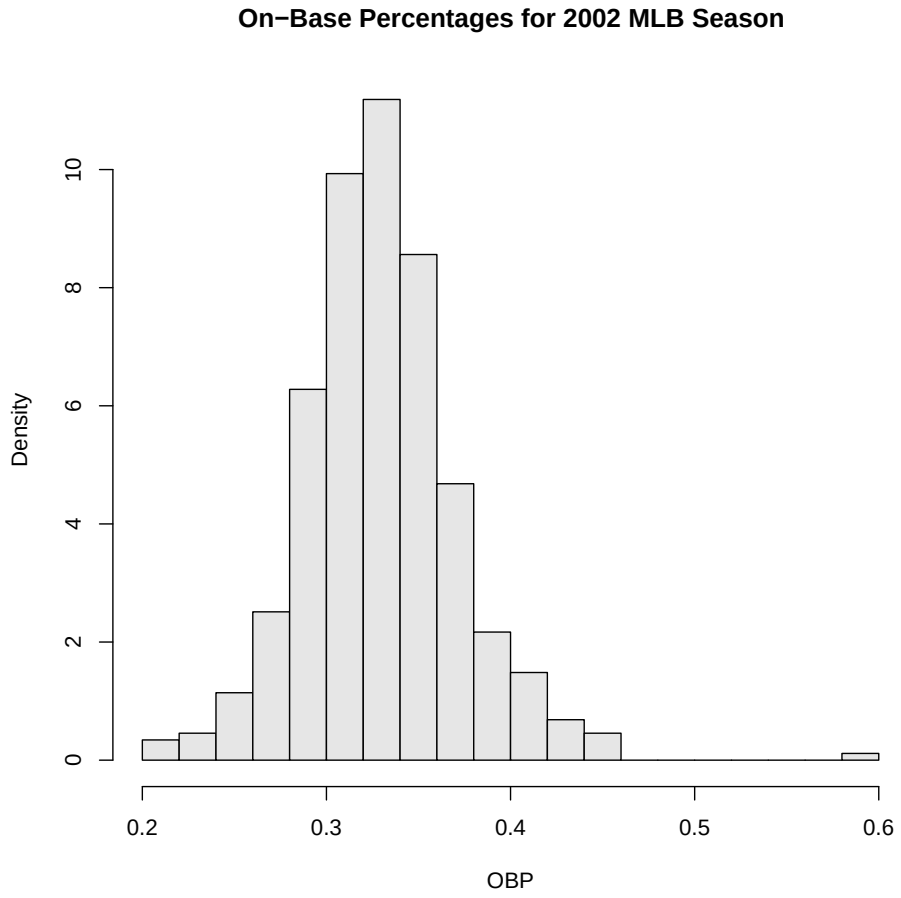

Figure 2.1: Example of an Univariate Outlier

single observation sticking out on the right end. This jibes with Gnanadesikan and Kettenring's conception of an univariate outlier as "one that sticks out on the end. "This observation is the on-base percentage for Barry Bonds, and its value is 0.582.

One may be interested in this observation because of the influence it exerts on the sample estimates of mean and variance; however, the influence may not be that great given the large sample size. It is more likely that one may interested

in this point for one of the other reasons mentioned in Chapter one. One may be motivated to spot outliers on batting statistics such as this one for the purpose of "special interest "; outliers in the positive represent truly exceptional performances by perhaps truly exceptional players. Based on the display, one could certainly make the claim that the one-base percentage achieved by Bonds in 2002 is quite remarkable; incidentally, the one-base percentage is an all-time Major League record. A little background into a current event in Major League Baseball reveals another reason for determining how outlying Bonds' performance was. There is a strong suspicion that Barry Bonds was using anabolic steroids during the 2002 season. That the Bonds on-base percentage is such an outlier could be considered a "diagnostic indication "(another reason discussed in Chapter One) that something other than normal human ability is responsible for this performance. This is in line with Hawkins' "intuitive definition "of an outlier being generated by a different mechanism than the one that was at work for the other observations. Now let us consider another definition of an outlier.

Beckman and Cook offer a less vague defintion of an outlier. They define an outlier as "... a collective to refer to either a contaminant or discordant observation "(1983, p. 121), and a contaminant is defined as "... any observation that is not a realization from the target distribution "(1983, p. 121). This study will concern itself exclusively with outliers as contaminants.

Once again, Bonds' on-base percentage observation is in line with this definition: It is possible that it is a not realization from the distribution of seasonal on-base percentages (target distribution) which applies to players who are not using steroids. At the very least the Bonds observation is certainly discordant and, thus, qualifies as an outlier using the Beckman and Cook definition.

The collective definition provided by Beckman and Cook is consistent with the two mechanisms Hawkins proposes to explain the "genesis of outliers. "(Hawkins,

1980, p. 2) Hawkins' first mechanism is the target distribution itself. An example of this mechanism in action would be if all of the data came from a Cauchy distribution. The Cauchy distribution is symmetric and bell shaped, but it has very heavy tails (in fact so heavy that the variance is infinite). Because of its heavy tails, the Cauchy distribution can produce observations that are discordant. In this case the outliers are not contaminants even though they are discordant. In regard to Hawkins' second mechanism, "The data arise from two distributions. One of these, the 'basic distribution ', generates 'good 'observations, while another, the 'contaminating distribution ', generates 'contaminants '"(1980, p. 2).

It should be pointed out that the use of the two purely quantitative methods for spotting a univariate outlier mentioned in Chapter one would have flagged the Bonds' on-base percentage as an outlier. The observation's standardized score is 5.99. Additionally,the observation is 4.56 interquartile ranges above the third quartile.

Keep in mind that this example was of a univariate outlier. The multivariate situation where you may have , say, seven measurements on each observation is more complicated. It is unclear where the end is in seven-dimensional space.

2.2 Formal Tests of Discordancy

Let us now discuss early multivariate outlier detection methods. The earliest research into the development of methods for multivariate outlier detection is nicely organized in the two books: *Outliers in Statistical Data* by Barnett and Lewis (1984) and *Identification of Outliers* by Hawkins (1980). Barnett and Lewis refer to the early methods of outlier detection in both univariate and multivariate analyses as "tests of discordancy ". I have chosen to add the adjective "formal "to emphasize the fact that tests of hypotheses are involved in these methods. The use of the methodology of hypothesis testing distinguishes the earliest effort from the current

research into the topic. As I will discuss later, I feel that the failure to make this distinction has resulted in some confusion regarding the properties of the detection metrics employed. What follows is a synthesis of presentations on formal tests of discordancy in the aforementioned books.

Consider a random sample $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ of p component vectors that are realizations of the random variables $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}$. The null hypothesis for every one of the tests that will be discussed shall be the same:

$$H_0: \quad \mathbf{X}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad i = 1, \ldots, n \tag{2.1}$$

First let us consider the test for just a single contaminant. In this case, most of the formal tests of discordancy that have been developed have as their alternate hypothesis one of the following.

$$\begin{aligned} H_1: \quad & \mathbf{X}_j \sim N(\boldsymbol{\mu} + \mathbf{a}, \boldsymbol{\Sigma}) \\ & \mathbf{X}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \forall i \neq j \end{aligned} \tag{2.2}$$

or

$$\begin{aligned} H_2: \quad & \mathbf{X}_j \sim N(\boldsymbol{\mu}, a\boldsymbol{\Sigma}) \\ & \mathbf{X}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \forall i \neq j \end{aligned} \tag{2.3}$$

or

$$\begin{aligned} H_3: \quad & \mathbf{X}_j \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_j) \\ & \mathbf{X}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \forall i \neq j \end{aligned} \tag{2.4}$$

where $\mathbf{X}_j$ represents the single observation that is an outlier candidate. Assuming that both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown, there exists, for the first two alternative hypotheses,

a test statistic which is optimal for the class of tests that is invariant under aribitrary, full-rank linear transformations. This statistic is the Mahalanobis D^2:

$$D_j^2 = (\mathbf{x}_j - \overline{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_j - \overline{\mathbf{x}}) \tag{2.5}$$

where $\mathbf{x}_j$ is the sample observation with the largest value of D^2. The Mahalanobis D^2 shall be discussed in greater depth in section (2.3). There are a couple of test statistics that will yield equivalent results in terms of the result of the hypothesis test. These test statistics are degenerate cases of Wilk's Lambda and Hotelling's T^2. Recall that Wilk's Lambda is used in the multivariate analysis of variance and Hotelling's T^2 is used in making inferences about a pair of population mean vectors. When one is testing the equivalency of two population mean vectors the statistics are equivalent just like t^2 and F in the analagous univariate situation. I have chosen to refer to the forms of these statistics in the context of testing whether a single observation is an outlier as degenerate because the observation x_j plays the role of a sample mean in these two statistics. However, x_j is a mean based on one observation. It should be noted that the value of $\overline{\mathbf{x}}$ used in the computation of all three statistics is to be computed from the subsample that does not include the single value $\mathbf{x}_j$. Of the three statistics, the Mahalanobis D^2 is the most commonly used form of the test statistic (Barnett & Lewis, 1984; Hawkins, 1980).

A point of fundamental importance concerns the sampling distribution of the three test statistics mentioned in the preceding paragraph. They are *order statistics* (see Appendix B. Specifically, they are the n^{th} largest order statistics. In this case, the distribution of the order statistic is very different from the distribution for a specific observation, $\mathbf{x}_i$. Siotani was the first one to investigate the distribution of the n^{th} order statistic of the Mahalanobis D^2. He did so within the context of spotting multivariate outliers. The distribution of the Mahalanobis D^2 order statistic is very complicated. Siotani computed approximate 5% and 1% critical values of the order

statistic for bivariate normal samples. The table was reprinted by Barnett and Lewis (Siotani, 1959; Barnett & Lewis, 1984). Because of the complicated distributions of the test statistics, research into formal tests of discordancy for multivariate normal samples ended with the publication of the books by Hawkins and the tandem of Barnett and Lewis. For the same reason, formal tests of discordancy are rarely used in applied work. Even in this day and age of awesome computing power and Markov Chain Monte Carlo simulation, the problem of computing p-values for the order statistic is still not practical. One last thing to note is that Hawkins has shown that there is no optimal test for the situation in which the alternative hypothesis is (2.4). Recall from Chapter one that this is for the situation in which contaminants are purely correlational outliers. This does not bode well for success in spotting correlational outliers.

2.3 The Mahalanobis D squared and Related Metrics

In the preceding section it was noted that the most frequently used statistic in a formal test of discordancy is the Mahalanbobis D^2. As we shall see later, this statistic is commonly used in an informal manner to detect multivariate outliers. Its use has been so widespread that it merits further discussion here.

The terms Mahalanobis Distance, Mahalanobis Generalized Distance, Mahalanobis Squared Distance and Mahalanobis D^2 have each been applied to several different statistics. These different statistics are similar in that they are each of the form $\mathbf{d'V^{-1}d}$ where $\mathbf{d}$ is a deviation or difference vector of some kind and $\mathbf{V}$ is a sample or population variance-covariance matrix. The statistics are different, however, and some clarification is in order. Some historical background will aid in this clarification. There are several different statistics with the Mahalanobis appellation. Let us be clear about which one is used in multivariate outlier detection.

Prasanta Chandra Mahalanobis, "...father of the statistical movement in

India "(Rudra, 1996, p. vii), "was very much interested in the divergence between various Indian races, especially hill tribes, and had devised for this purpose a statistical measure, now known as the Mahalanobis Distance or D^2 "(Rudra, 1996, p. 419). Specifically, another scientist "...had taken anthropological measurements such as stature, head length, head breadth, nasal length, upper face length, etc. of 300 Anglo-Indians in Calcutta "(Mahalanobis, 1983, p. 33). Using this data and similar data on other communities, castes or tribes, Mahalanobis was interested in whether the Anglo-Indians as a group had an affinity toward any of the other groups. He also wanted to be able to understand in some way the extent of affinities, as well.

2.3.1 The Mahalanobis Generalized Distance

In pursuit of his research questions, Mahalanobis analyzed the p-dimensional mean vectors for the different groups where the p dimensions correspond to the p different anthropometric measurements taken on each group. His work toward quantifying the separation between two groups in p-space led to what is now known as the Mahalanobis Generalized Distance:

$$G = (\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_j)'\mathbf{S}^{-1}(\overline{\mathbf{x}}_i - \overline{\mathbf{x}}_j) \tag{2.6}$$

where $\overline{\mathbf{x}}_i$ and $\overline{\mathbf{x}}_j$ are the means for two distinct groups, and $\mathbf{S}$ is the sample variance-covariance matrix pooled over the two groups (Rudra, 1996; Dillon & Goldstein, 1984).

A multiple, Z, of G can be used to test the hypothesis $H_0 : \boldsymbol{\mu}_i = \boldsymbol{\mu}_j$:

$$Z = \frac{n_i n_j}{n_i + n_j} \frac{n_i + n_j - p - 1}{(n_i + n_j - 2)p} D^2. \tag{2.7}$$

Assuming that the null is true,

$$Z \sim F_{(p,n_i+n_j-p-1)}. \tag{2.8}$$

(Dillon & Goldstein, 1984).

2.3.2 Form Used to Detect Multivariate Outliers

The form used to detect multivariate outliers, equation 2.5, is a special case of the Mahalanobis Generalized Distance, and it shall henceforth be referred to as the Mahalanobis D^2. If $\mathbf{x}_i$ is excluded from the computation of $\overline{\mathbf{x}}$ and $\mathbf{S}$, then D_i^2 is a special instance of the Mahalanobis Generalized Distance in that $\mathbf{x}_i$ is a sample mean, albeit of one observation, and $\mathbf{S}$ can still be considered to be the pooled sample variance-covariance matrix (it has the same number of degrees of freedom). In order to expedite the calculations that are involved, most software packages use all of the observations in the computations of $\overline{\mathbf{x}}$ and $\mathbf{S}$, and this form is not a special case of the Mahalanobis generalized distance. This is done in spite of the fact that the exclusion of the observation is preferable in terms of alleviating a masking effect

The Mahalanobis D^2 has an appealing geometric interpretation. Let us consider some examples of bivariate normally distributed data and the idea of "statistical distance "(Johnson & Wichern, 1992). Figure 2.3.2 is a graphical depiction of a random sample ($n = 100$) of two uncorrelated, normally distributed random variables, in addition to two observations indicated by the green hash mark and the red triangle. The values of x and y are in deviation form. The random variable represented by the abcissa is standard normal (mean = 0, standard deviation = 1). The normal random variable represented by the ordinate has mean of zero and a standard deviation of 0.5. The coordinates of the red triangle and green cross are (2, 0) and (0, 2), respectively. Note that the Euclidean distances from the centroid or origin, $d_i = \sqrt{x^2 + y^2}$, for the two points are the same: $d = 2$. However, let us

consider the question of which point is farther from the pattern established by the sample of 100 points. Clearly, the point represented by the green hash mark is more outlying. Thus, the Euclidean distance is not a valid measure of how outlying an observation is from the sample distribution. How can we assign a distance to each point that indicates how far that observation is from the sample distribution? A logical choice would be to weight the square of the deviation by its variance. In our example, this would yield the squared distnce: $d^2 = \frac{x^2}{1} + \frac{y^2}{0.25}$. Observe that for a constant d^2 this is an equation for an ellipse with the centroid as its center. Thus, points with the same distance lie on an ellipse (Johnson & Wichern, 1992).

Now consider the case of two random normal variables that are correlated. Figure 2.3.2 depicts two variables, x and y, that are highly correlated. Once again, the Euclidean distance does not convey a "statistical distance "or how probable or improbable a point is nor does the sum of the deviation scores weighted by their variances that was used in the case of uncorrelated normal random variables. Obviously, the point represented by the red triangle should be assigned a greater "statistical distance "than the observation represented by the green cross (which is in the upper-right corner on top of a point represented by a circle). The problem can be solved if we change our perspective so that we have the same situation that was depicted in Figure 2.3.2, namely two uncorrelated variables with differing variances. Imagine rotating the x and y axes until the projections of the points onto the rotated axes become uncorrelated. This is depicted in Figure 2.3.2 (It is a rigid rotation. The rotated axes should be perpendicular even though they do not appear to be.) The rotated axes are the major and minor axes of the ellipses that define a constant statistical distance just as they were in the previous example. The rotated axes define a new coordinate system. The projection of an observation onto the major axis is its new x coordinate, x', and the the projection of the observation onto the minor axis is the new y coordinate, y'. Remember that a projection can be thought

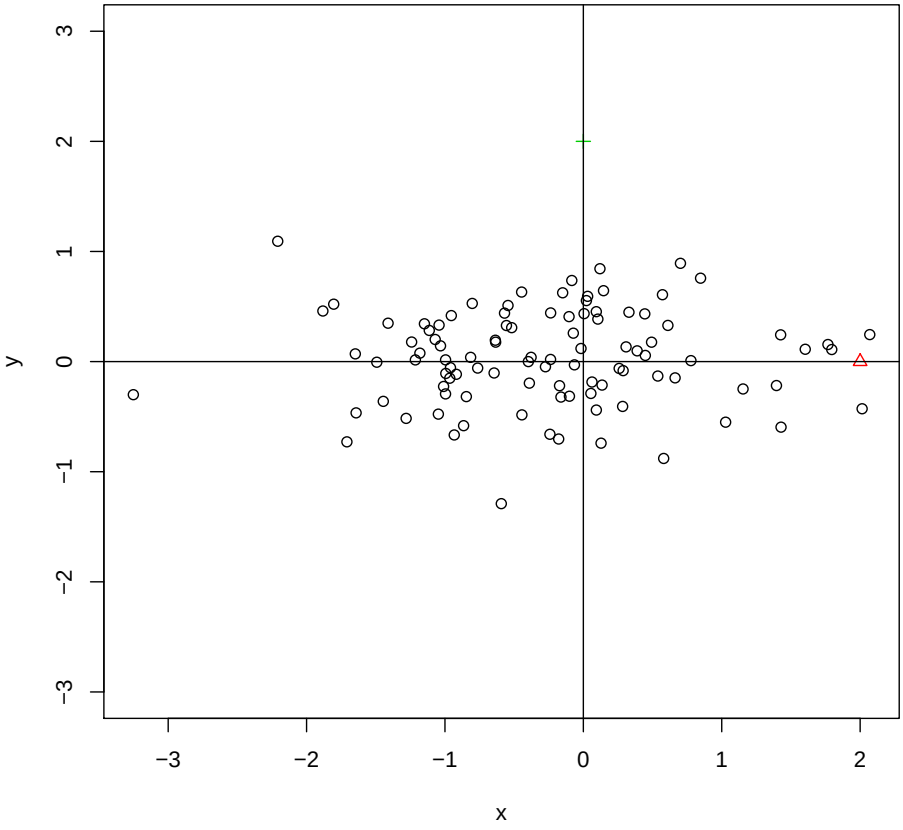

Figure 2.2: Uncorrelated Bivariate Normal Data with Differing Variances

of as a inner product of two vectors. One vector starts at the origin of the coordinate system with its terminus at the point representing the observation. The other vector is of unit length starting at the origin and running down the x' axis. The projection of the observation onto the x' axis is the inner product of these two vectors and that projection is the x' coordinate. The y' coordinate can be found in the same way. The new equation for the statistical distance is

$$d^2 = \frac{(x')^2}{s^2_{x'}} + \frac{(y')^2}{s^2_{y'}} \tag{2.9}$$

where $s^2_{x'}$ and $s^2_{y'}$ are the variances x' and y', respectively.

This idea of statistical distance can be generalized to points in p space. Rigidly rotate the p orthogonal axes until they line up with the axes of the elliptical pattern of the data. It should be emphasized again that we are assuming that the data come from a multivariate normal population.

A Geometric Interpretation of the Mahalanobis D squared

Now consider the Mahalanobis D^2,(2.5). Since it is a squared distance, it makes sense that it must always be positive. In the world of matrix algebra, this is expressed by the statement: $\mathbf{S}^{-1}$ is positive definite. There is a theorem known as the *Spectral Decomposition Theorem* which states that

$$\mathbf{S}^{-1} = \frac{\mathbf{e}_1\mathbf{e}_1'}{\lambda_1} + \frac{\mathbf{e}_2\mathbf{e}_2'}{\lambda_2} + \cdots + \frac{\mathbf{e}_p\mathbf{e}_p'}{\lambda_p}. \tag{2.10}$$

The $\mathbf{e}_i$'s are the p eigenvectors of both $\mathbf{S}$ and $\mathbf{S}^{-1}$, and the λ_i's are the p eigenvalues of $\mathbf{S}$ (Note that the eigenvalues of $\mathbf{S}^{-1}$ are easily shown to be the reciprocals of the eigenvalues of $\mathbf{S}$). The eigenvectors and eigenvalues of a sample variance-covariance matrix are the sample principal components and their corresponding variances (Johnson & Wichern, 1992). A brief review of principal components is be

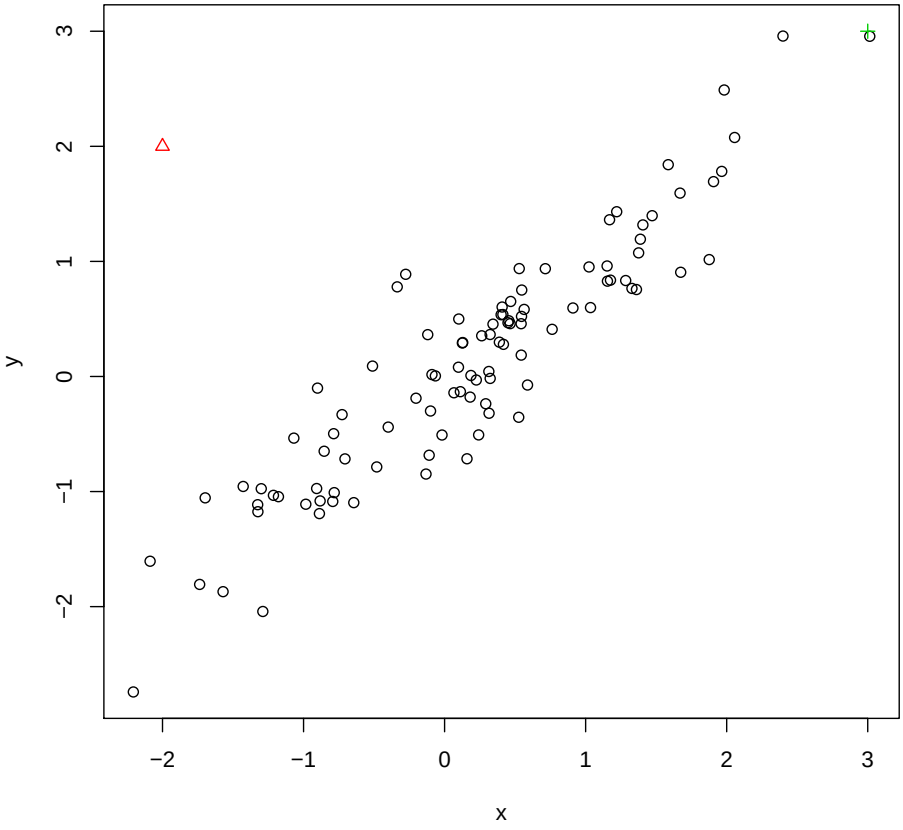

Figure 2.3: Bivariate Normal Sample from Population with a High Correlation

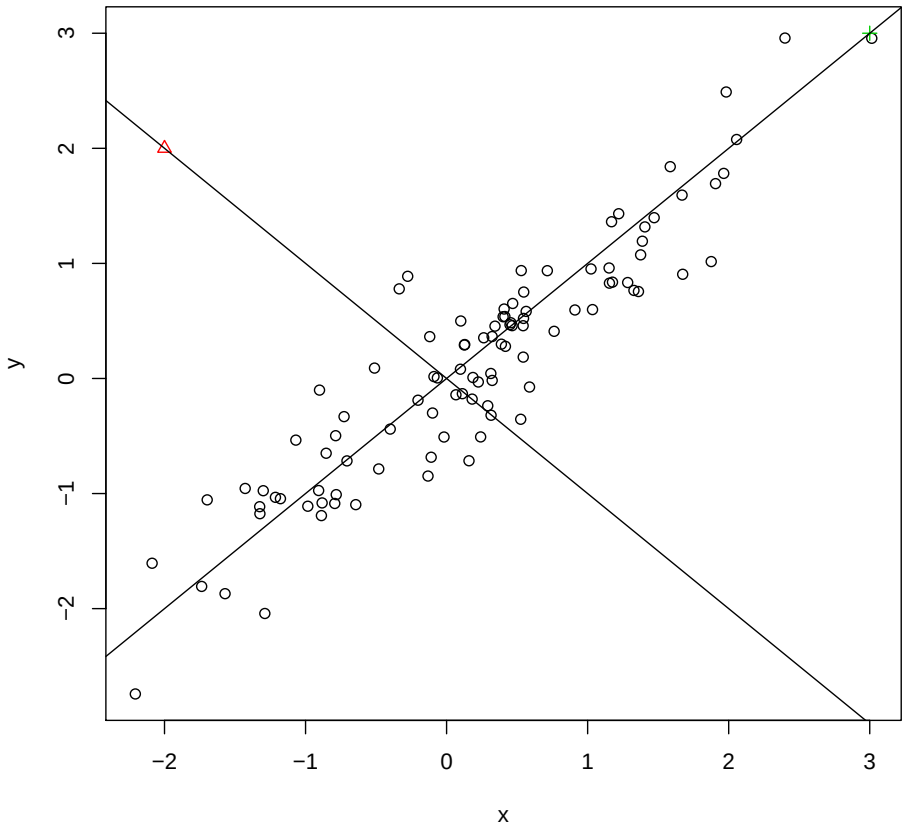

Figure 2.4: Depiction of Axes Rotated in the Direction of Principal Components

provided next.

There are basically two interpretations of the principal component analysis of a sample variance-covariance matrix. Assume that we have a multivariate normal sample on p variables. Consider the situation where we reduce each multivariate observation down to a univariate value by taking a linear combination of the p variables. We know that multivariate normality implies the marginal normality of each of the p variables, and a linear combination of normally distributed random variables is distributed normally. Obviously, there are an infinite number of possible linear combinations of the p variables. A linear combination with a maximum variance would be of interest. There is a slight problem in looking for such a maximum-variance linear combination. We can always multiply a linear combination by a constant, and the variance will be increased by the square of this constant. So, we had the restriction that the sum of the squares of the coefficients in the linear combination is one. (In other words, the length of the vector corresponding to the linear combination is one). It can be shown with a variant of the Cauchy-Schwartz inequality that the linear combination with the maximum variance has coefficients that correspond to the elements of the first eigenvector, and the variance of this linear combination is the first eigenvalue. Next consider the set of linear combinations (with the restriction that the corresponding vector is of length one) whose corresponding vector is perpendicular to the one for the maximum-variance linear combination. Which linear combination meeting these restrictions has the maximum variance? It can be shown that it is the linear combination corresponding to the second normalized eigenvector, and its variance is the corresponding eigenvalue. Now consider an arbitrary i^{th} linear combination with the restrictions that its corresponding vector is of length one and is perpendicular to the preceding $i-1$ linear combinations. Of all such linear combinations which one has the maximum variance? It is the i^{th} principal component, and its variance is the corresponding

i^{th} eigenvalue (Johnson & Wichern, 1992).

The second interpretation of principal components was first articulated by Karl Pearson over one hundred years ago (Johnson & Wichern, 1992). Consider the plot of each sample observation in p space. What is the best-fitting k dimensional hyperplane($k < p$)? An observation's fitted value lying in the plane is its orthogonal projection into the plane. Fit is assessed by the sum of the squares of the lengths of the residual vectors that are each perpendicular to the fitted plane. It can be shown that the answer is the hyperplane that is the span of the first k eigenvectors and passes through the centroid of the data. The residual vector for each point is its projection on the remaining $p - k$ eigenvectors (Johnson & Wichern, 1992; Stapleton, 1998). We shall revisit the second interpretation of principal components later when we discuss a line of research in multivariate outlier detection that began with Rao. For a very readable derivation of both interpretations of principal components analysis refer to Johnson and Wichern's book, *Applied Multivariate Statistical Analysis* (1992). Now let us return to the discussion of the Spectral Decomposition Theorem and the Mahalanobis distance.

Substituting the spectral decomposition of $\mathbf{S}^{-1}$ into (2.5), we obtain:

$$D_i^2 = (\mathbf{x}_i - \overline{\mathbf{x}})' \left(\frac{\mathbf{e}_1 \mathbf{e}_1'}{\lambda_1} + \cdots + \frac{\mathbf{e}_p \mathbf{e}_p'}{\lambda_p} \right) (\mathbf{x}_i - \overline{\mathbf{x}}). \tag{2.11}$$

Now using the distributive properties for vector and matrix multiplication we have:

$$D_i^2 = \frac{(\mathbf{x}_i - \overline{\mathbf{x}})' \mathbf{e}_1 \mathbf{e}_1' (\mathbf{x}_i - \overline{\mathbf{x}})}{\lambda_1} + \cdots + \frac{(\mathbf{x}_i - \overline{\mathbf{x}})' \mathbf{e}_p \mathbf{e}_p' (\mathbf{x}_i - \overline{\mathbf{x}})}{\lambda p}. \tag{2.12}$$

If you look carefully, you can see that the numerator in each term is the square of the projection of the observation, in deviation form, onto the corresponding eigenvector (or principal component), and the denominator is the variance of the principal component.

Thinking back to our bivariate example and remembering that principal components are linear combinations with maximum variances subject to the fact that they are orthogonal to each other, you can see that the major axis of the ellipitical pattern in Figure 2.3.2 is the first principal component. (It is the direction in which there is maximum variability) and that the minor axis is the second principal component. This geometrical interpretation generalizes to higher dimensions. In this case the principal components are the axes of a hyperellipse. The main points are that the Mahalanobis D^2 can be expressed as the sum of the weighted projections onto the p principal components, the principal components are in the directions of the axes of the elliptical pattern; thus, the Mahalanobis D^2 is the "statistical distance" 2.9 we developed logically earlier. Thus, the Mahalanobis D^2 is not only elegant (more will be said about this elegance, shortly) in terms of matrix algebra but it also logical from a geometric point of view.

Algebraic Interpretation

Now more about the Mahalanobis D^2's algebraic appeal. It is the multivariate analog of the square of the sample standard score. This is demonstrated by the following algebraic relationship:

$$\begin{aligned} z_i^2 &= \left(\frac{x_i - \overline{x}}{s}\right)^2 \\ &= (x_i - \overline{x})(s^2)^{-1}(x_i - \overline{x}) \end{aligned} \tag{2.13}$$

Equation (2.13) is in the univariate form of the Mahalanobis D^2 (2.5), and it is in fact the Mahalanobis D^2 for one dimensional data (a degenerate case).

2.3.3 Informal Use of the Mahalanobis Squared Distance

The history of the use of the Mahalanobis D^2 closely parallels that of the univariate sample standard score, $z_i = (x_i - \overline{x})/s$. We have already seen that order statistics

of the Mahalanobis D^2 were used in formal tests of discordancy. Even earlier order statistics of the sample standard score were used in univariate tests of discordancy. However, like the distribution of order statistics for the Mahalanobis D^2, the distribution of order statistics of the sample standard score are highly complicated (Nair, 1948, 1952). The end result of this was that formal tests of univariate discordancy were not feasible as the corresponding critical values were unobtainable most of the time. There was no clear declaration of this in the literature, but somewhere along the way practitioners began using informal cut-off values. Another informal method that appeared without a formal declaration was the plotting of standard scores. The practitioner looked at the plot and subjectively determined which observations were potential outliers.

The logic used in determining cut-off values often involves the approximate distribution for an individual observation and not the distribution of the order statistics for that distribution. Under the assumption of normality and a relatively large sample size,

$$d_i = \frac{x_i - \overline{x}}{s} \doteq \frac{x_i - \mu}{\sigma} \sim N(0,1). \tag{2.14}$$

Thus the individual values of the sample standard scores are approximately distributed as a standard normal random variable.

In their book, *Statistical Methods in Education and Psychology*, Glass and Hopkins employ such reasoning to use a cut-off value of $|4.72|$ for what they call an "extreme outlier. "(Glass & Hopkins, 1996, p. 27) Using the cumulative distribution function for a standard random normal variable, they state, "... an extreme outlier would be expected in only about one score in a million "(Glass & Hopkins, 1996, p. 27). Note that this statement of probability is incorrect. The probabilities should be based on the distribution of order statistics for the sample size used. This a subtle point but one of importance nonetheless.

An almost identical evolution in the use of the Mahalanobis D^2 occurred.

Informal cut-offs can be informed by an approximating distribution for the an individual observation:

$$D^2 = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}) \doteq (\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \sim \chi^2_p. \tag{2.15}$$

Of course we are assuming that data are coming from a multivariate normal distribution and that we have a relatively large sample size. The degrees of freedom, p, for the χ^2_p random variable is the number of variates. The fact that Equation 2.15 with the population parameters is distributed as χ^2_p random variable can be seen by applying rules for the expectation and variance of a linear combination of a random vector. $\mathrm{E}(\mathbf{e}_i'\mathbf{X}) = \mathbf{e}_i'\mathrm{E}(\mathbf{X}) = \mathbf{e}_i'\mathbf{0} = 0$. Remember $\mathbf{X}$ is in deviation form around the population mean vector. The covariance between $\mathbf{e}_i'\mathbf{X}$ and $\mathbf{e}_j'\mathbf{X}$ for i not equal to j is $\mathbf{e}_i'\boldsymbol{\Sigma}\mathbf{e}_j$ and since $\mathbf{e}_j$ is an eigenvector of $\boldsymbol{\Sigma}\mathbf{e}_j = \lambda_j\mathbf{e}_j$ Substituting this last expression into the first one for the covariance we have a scalar, λ_j times the inner product of $\mathbf{e}_i$ and $\mathbf{e}_j$ which is zero because of the orthogonality of the eigenvectors (Johnson & Wichern, 1992). As in the case of of univariate sample standard score, the beginning of this informal usage of the Mahalanobis D^2 was not heralded by any formal research. A popular and contemporary applied multivariate statistics book advocates using the $99.9th$ percentile value of the χ^2_p as a cut-off value (Tabachnick & Fidell, 1996).

2.3.4 Informal Metrics Related to the Mahalanobis Squared Distance

Several other metrics that have been shown to be closely related to the Mahalanobis D^2 have been developed and used over the years to detect multivariate outliers. Probably the most widely used are the diagonal elements of the Hat Matrix.

The Diagonal Elements of the Hat Matrix

Consider the the regression model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$. $\mathbf{Y}$ is the n x 1 vector of observations on the dependent variable. $\mathbf{X}$ is the n x (p + 1) design matrix where p is the number of independent variables. The design matrix has p + 1 columns because one column is the vector comprised of all ones that corresponds to the intercept term in the model. It is not a part of the design matrix which becomes a n x p matrix when the model passes through the origin. $\boldsymbol{\beta}$ is the $(p + 1) \times 1$ vector of coefficients and $\boldsymbol{\epsilon}$ is the n x 1 vector of normal random errors. $\boldsymbol{\epsilon} \sim N(\mathbf{O}, \sigma^2\mathbf{I}_{nxn})$. In other words the individual ϵ_i's are identically and independently distributed normal random variables. We know that sample estimates (based on the least squares criterion or the maximum likelihood criterion) of the conditional means, $E(Y_i|\mathbf{x}_i)$ are the elements of the vector, $\hat{\mathbf{Y}}$, and this vector is obtained from the matrix product $\hat{\mathbf{Y}} = \mathbf{HY} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ where $\mathbf{H}$ is the Hat Matrix. Looking at the equation, you can see that $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'$. Thus, the column vector of residuals, $\mathbf{Y} - \hat{\mathbf{Y}}$ equals $(\mathbf{I} - \mathbf{H})\mathbf{Y}$. Therefore the variance-covariance matrix of the residuals is $(\mathbf{I} - \mathbf{H})\boldsymbol{\Sigma}_Y(\mathbf{I} - \mathbf{H})' = (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I})(\mathbf{I} - \mathbf{H})'$. Distributing the transpose operator, we have $\sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I}' - \mathbf{H}')$. Then performing the matrix multiplication using one of the distributive properties of matrix multiplication and combining like terms we have $(\mathbf{I} - \mathbf{H})\sigma^2$. Thus the variances of the residuals are the diagonal elements of this matrix, $(1 - h_{ii})\sigma^2$ (Hoaglin & Welsch, 1978; Draper & Smith, 1981; Johnson & Wichern, 1992; Stapleton, 1998).

It can be shown that the diagonal elements of the Hat Matrix, h_{ii} are between 0 and 1 inclusive; therefore, a h_{ii} with a value near one will result in a very small variance for the corresponding residual (Belsley, Kuh, & Welsch, 1980). This means that the estimated regression line or (hyper)plane will be pulled toward this point whatever the value of the Y_i. Thus, a point with a corresponding h_{ii} near one will exert a lot of influence on the estimated parameters of the fitted line. In the case

of the simple linear regression of one response variable onto one predictor, it can be demonstrated that it is the observations that are outlying with respect to the single predictor variable that are influential (Hoaglin & Welsch, 1978).

Consider the data depicted in Figure 2.5. The data represents IQs for twins separated at birth. For each set of twins, one twin was raised by his or her biological parents and the other was raised by foster parents (Verzani, 2005). I have added a spurious data point represented by a red triangle. Notice how the regression line is pulled toward the spurious point. Figure 2.6 , a boxplot, shows that the value on the predictor variable for the point is an outlier with respect to the distribution of IQs of twins raised by their biological parents. We can reason that since it is points that are outlying with respect to the factor space that are influential, the value h_{ii} indicates the degree to which an observation is outlying with respect to the factor space.

Bollen suggested that the diagonal elements of the Hat Matrix formed from a data matrix of standardized scores can be used to determine whether the corresponding observations are possible outliers. He made this suggestion within the context of confirmatory factor analysis. Further he suggested spotting outlying values on h_{ii} via a graphical display (Bollen, 1987).

Cut-off values applied to determine whether an observation in a regression model is influential can also be applied to determine whether the corresponding observation is an outlier. It can be shown that the sum of the diagonal elements of the hat matrix is p, the number of columns in the data matrix (Belsley et al., 1980; Hoaglin & Welsch, 1978) Therefore, the average value of h_{ii} will always be p/n. Based on the idea that twice the mean is an extreme value, Hoaglin and Welsch have suggested that $2p/n$ be used a cut-off value in the determination of whether an observation is influential or outlying. This cut-off value seems to have become standard practice in the application of the diagonal elements of the Hat Matrix to

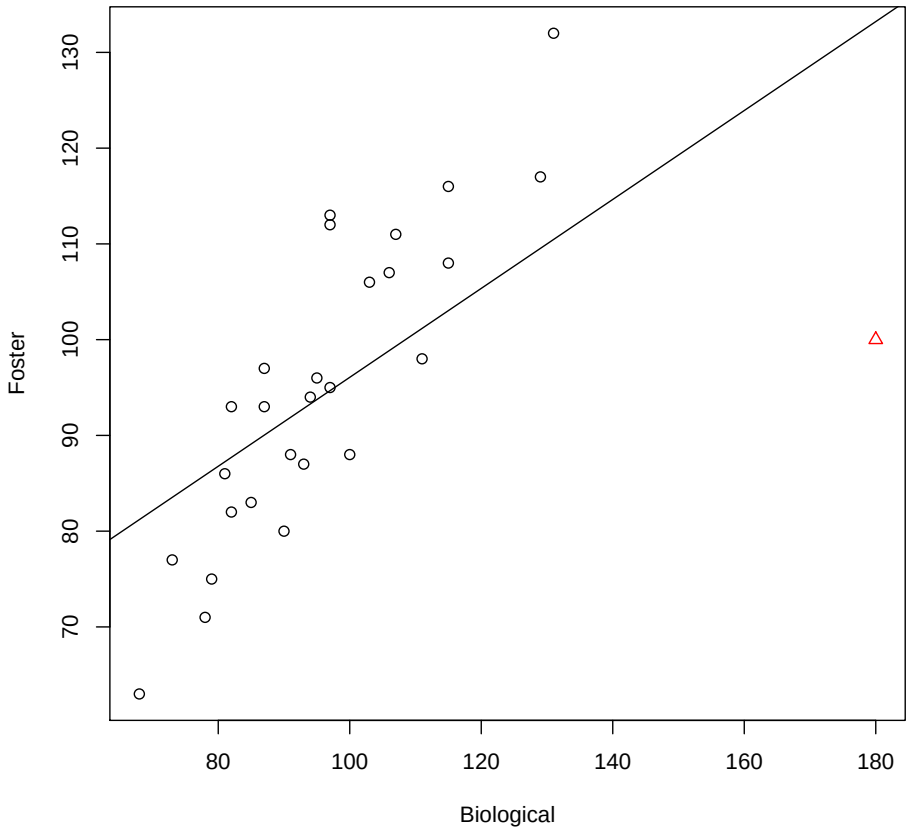

Figure 2.5: "Example of an Influential Point in a Simple Linear Regression"

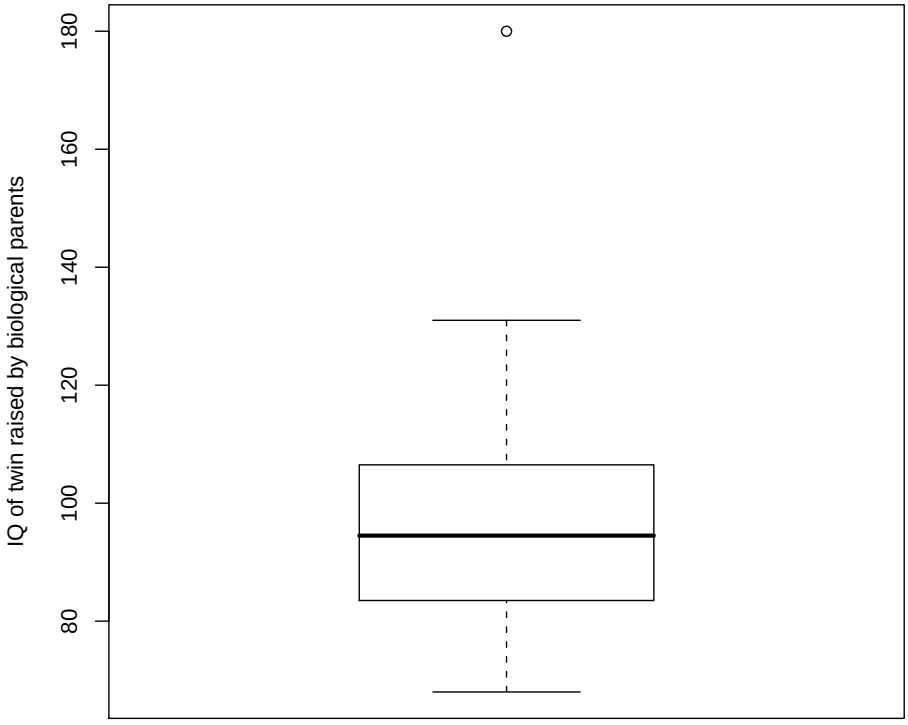

Figure 2.6: "Boxplot of Values on the Twin Data Predictor Variable"

outlier detection (Carrig, 2005). However, the statistical software package Minitab uses a more conservative value of $3p/n$. As part of its output for any analysis involving a linear model it automatically informs the user that a particular point is heavily influential if its corresponding h_{ii} exceeds $3p/n$ ("Minitab Reference Manual Release 11", 1999).

Rousseeuw and van Zomeren have shown that the diagonal elements of the Hat Matrix are a monotonic function of the Mahalanobis D^2:

$$h_{ii} = \frac{MD_i^2}{(n-1)} + \frac{1}{n} \tag{2.16}$$

(Rousseeuw & Zomeren, 1990) As will be apparent later, because of this monotonic relationship the diagonal elements of the Hat Matrix cannot offer anything over and above the Mahalanobis D^2 as far as this study is concerned. Their application would yield results identical to that of the Mahalanobis D^2.

Mardia's Multivariate Kurtosis Measure

The use of an observation's corresponding term in Mardia's multivariate kurtosis measure has been suggested for use as outlier detection by a number of researchers (Bacon, 1995). This measure is denoted by

$$b_{2,p} = \frac{1}{n}\sum_{i=1}^{n}\left[(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})\right]^2. \tag{2.17}$$

Outliers contribute heavily to the kurtosis of a distribution, so it stands to reason that a point that contributes heavily to the kurtosis is a possible outlier. An individual point's contribution to the measure of kurtosis is

$$k_i = \left[(\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \overline{\mathbf{x}})\right]^2. \tag{2.18}$$

This is clearly the square of the Mahalanobis D^2. Thus, it too is a monotonic function of the Mahalanobis D^2. And, as far as this study is concerned its application in the context of the study to be conducted later in this paper will yield results identical to the application of the Mahalanobis D^2.

2.4 The Line of Research Beginning with Pearson and then Rao

Recall in that in Section 2.3.2 two interpretations of principal components were outlined. There is a line of research in the area of multivariate outlier detection that began with Pearson's interpretation of principal components as best-fitting hyperplanes. Many years later in 1964, Rao used this interpretation to come up with a residual-based approach to spotting multivariate outliers (Rao, 1964). Hawkins refined Rao's idea and offered a couple of methods for spotting multivariate outliers (Hawkins, 1974; Hawkins & Fatti, 1984; Hawkins, 1980). Jolliffe justified and championed Hawkins ideas in his 1986 book, *Principal Components Analysis*. Surprisingly, this line of research ended with Jolliffe's book. There has never been a study in which the ideas of this line of research were tested against other metrics.

2.4.1 Rao's Suggestion

Some Background

Consider Pearson's interpretation of principal components analysis as the fitting of hyperplane models to the multivariate data. Consider data in which there are four measurements on each observation. The best fitting two-dimensional plane (defined by the criterion of minimizing the sum the square lengths of the perpendiculars from the points to the plane) is the one spanned by the first two eigenvectors (principal components) passing through the centroid. Note we are working under the assump-

tion that the vector value of an observation is its deviation from the centroid, so the centroid is the origin in the coordinate system. The fitted value of an observation is the orthogonal projection of the observation into the linear subspace spanned by the eigenvectors (principal components) (Johnson & Wichern, 1992; Stapleton, 1998). In the discussion that follows, the first k principal components are being extracted (based on for example the inspection of a Scree Plot or by applying some rule of thumb such as the Kaiser criterion of an eigenvalue greater than one). Using concepts gleaned from Stapleton, this projection can be expressed in matrix form (1998). For a $p \times n$ data matrix, $\mathbf{X}$:

$$\hat{\mathbf{X}} = \mathbf{PX} = \left(\mathbf{e}_1\mathbf{e}_1' + \mathbf{e}_2\mathbf{e}_2' + \cdots + \mathbf{e}_k\mathbf{e}_k'\right)\mathbf{X}. \tag{2.19}$$

I shall now outline a not-so-rigorous derivation of Equation 2.19. Recall that linear regression can be thought of as an orthogonal projection of the n dimensional vector of response values onto the p-dimensional subspace spanned by the columns of the design matrix. To avoid confusion with the data matrix, $\mathbf{X}$, in equation 2.19, let us name the design matrix for a linear regression $\mathbf{Z}$. We know that the $\hat{\mathbf{Y}}$ is obtained by multiplying $\mathbf{Y}$ by the Hat Matrix:

$$\hat{\mathbf{Y}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} \tag{2.20}$$

Because of this the Hat Matrix is also known as a projection matrix. Now, we have already stated that the projection of each multivariate observation into the best-fitting hyperplane is an orthogonal projection, so the projection of each observation is like a regression of sorts. Remember that in our principal components problem we are orthogonally projecting each observation into the hyperplane spanned by the first k eigenvectors (principal components). We also know that these k eigenvectors are mutually orthogonal. The fact that they are orthogonal is of key importance

in what follows. For the moment consider the problem of projecting the multivariate observations, $\mathbf{x}_i$ onto the subspace spanned by the first eigenvector (principal component). This can be thought of as a regression where our design matrix has the single column that is $\mathbf{e}_1$, so the projection matrix is $\mathbf{e}_1(\mathbf{e}_1'\mathbf{e}_1)^{-1}\mathbf{e}_1$. Because the eigenvectors are of unit length the middle term, $(\mathbf{e}_1'\mathbf{e}_1)^{-1}$, is just one, so the projection matrix is just $\mathbf{e}_1\mathbf{e}_1'$. Since all of the k eigenvectors are mutually orthogonal the subspaces spanned by them are mutually orthogonal. The subspace represented by the hyperplane that we are projecting each observation onto is a union of these mutually orthogonal subspaces. There is the somewhat intuitive result that when a vector is orthogonally projected into a subspace that is the union of mutuallly orthogonal subspaces the projection matrix is the sum of the projection matrices of the individual subspaces. Thus, we have Equation 2.19 (Stapleton, 1998).

Now, let us consider the residual vector:

$$\mathbf{x} - \hat{\mathbf{x}} = (\mathbf{I} - \mathbf{P})\mathbf{x}. \tag{2.21}$$

where $\hat{\mathbf{x}}$ is an estimate of the observation $\mathbf{x}$ obtained by projecting it into the best-fitting hyperplane. The matrix $\mathbf{I}$ is the $n \times n$ identity matrix. It can be shown that the matrix $\mathbf{I} - \mathbf{P}$ is itself a projection matrix. We also know that the Identity matrix is a projection matrix. The identity projection can be expressed as the sum of the projection matrices for each of the p eigenvectors. Once again this is due to mutual orthogonality of the individual eigenspaces. In other words, $\mathbf{I} = (\mathbf{e}_1\mathbf{e}_1' + \mathbf{e}_2\mathbf{e}_2' + \cdots + \mathbf{e}_p\mathbf{e}_p')$. Combining this result with Equation 2.19 yields:

$$(\mathbf{I} - \mathbf{P}) = (\mathbf{e}_{k+1}\mathbf{e}_{k+1}' + \cdots + \mathbf{e}_p\mathbf{e}_p'). \tag{2.22}$$

So, we have shown how we can obtain the residual vector corresponding to an observation's projection into the hyperplane spanned by the lesser principal components

of the covariance (correlation) matrix

Rao's Measure of Multivariate Discordancy

Rao's measure of multivariate discordancy is based on straightforward logic. As mentioned before, we are assuming that a determination of how many principal components to retain has been made and that the remaining principal components do not account for a substantial amount of the total sample variance. Essentially Rao observed that an observation whose corresponding point in p space is a large distance away from the best-fitting hyperplane is discordant; the observation falls outside the pattern established by the majority of data. The distance of a point to the best-fitting hyperplane is the length of the residual vector, so Rao suggested that we look at the squares of the lengths of the residual vectors (Rao, 1964). The square of the length of the residual vector can be easily computed using the last $p - k$ principal components. We know that the residual vector for an observtion, $\mathbf{x}_i$ is $(\mathbf{I} - \mathbf{P})\mathbf{x}_i$. The square of the length of this vector is the inner product with itelf: $\mathbf{x}_i'(\mathbf{I} - \mathbf{P})'(\mathbf{I} - \mathbf{P})\mathbf{x}_i$. Like all projection matrices, $(\mathbf{I} - \mathbf{P})$ is symmetric and idempotent, so the square of the length of the residual vector is $\mathbf{x}_i'(\mathbf{I} - \mathbf{P})\mathbf{x}_i$. Now combining this with Equation 2.22, we find that the square of the length, d_i^2, is

$$\begin{aligned} d_i^2 &= \mathbf{x}_i'(\mathbf{e}_{k+1}\mathbf{e}_{k+1}' + \cdots + \mathbf{e}_p\mathbf{e}_p')\mathbf{x}_i \\ &= \mathbf{x}_i'\mathbf{e}_{k+1}\mathbf{e}_{k+1}'\mathbf{x}_i + \cdots + \mathbf{x}_i'\mathbf{e}_p\mathbf{e}_p'\mathbf{x}_i \\ &= \sum_{j=k+1}^{p} \left[(\mathbf{e}_j'\mathbf{x}_i)^2\right] \end{aligned} \tag{2.23}$$

So, we see that the square of the length of the residual vector is the sum of the squared lengths of the projections of the observation on each of the lesser principal components. Recalling the result that was discussed toward the end of Section 2.4.1, we see that it is also the square of the length of the projection of the observation into

the hyperplane spanned by the lesser principal components. Rao did not entertain the question of how this metric is distributed. Thus, a practical way to apply this metric would be to compute it for each of the n observations and plot these values (e.g. stem-and-leaf plot) and look for observations that are outlying on the plot.

2.4.2 Hawkins' Refinement

Ten years after Rao's paper, Hawkins suggested a refinement to the Rao metric (2.23). In words, Hawkins' suggested refinement is that the projections of the observation on the lesser principal components be weighted by the corresponding variance of the principal component (eigenvalue). For reasons that will be clearer later in the paper, I shall call Hawkin's metric PCLOW (Neither Hawkins nor anyone else has ever given this metric a name). Its mathematical formulation follows.

$$\text{PCLOW} = \sum_{j=k+1}^{p} \left[(\mathbf{e}_j' \mathbf{x}_i)^2 / \lambda_j \right] \tag{2.24}$$

Before proceeding into Hawkins' reasoning for scaling the principal component scores, let us make an interesting observation. Recalling the discussion of statistical distance in Section 2.3.2, we see that PCLOW is the statistical distance of the point's projection into the residual subspace. Referring to Equation 2.12, we see that PCLOW is a partial sum of the Mahalanobis D^2. This seems to suggest that part of the sum that is the Mahalanobis Distance (Equation 2.12) is noise when it comes to detecting multivariate outliers.

Hawkins assumes that a multivariate outlier comes from the model: $\mathbf{x} = \mathbf{y} + \mathbf{e}$ where $\mathbf{y}$ can be regarded as the true value and $\mathbf{e}$ an error vector. Thus when we want to test whether observation $\mathbf{x}_i$ is a contaminant or outlier, we can equivalently test the null hypothesis: $\mathbf{e}_i = \mathbf{0}$ versus the alternative: $\mathbf{e}_i \neq \mathbf{0}$. Note that this is the same as the mean-slippage model discussed in 2.2. Without loss of generality, Hawkins assumes that each component of $\mathbf{Y}$ is a $N(0, 1)$ random

variable. Also, $\mathbf{Y}$ is a multivariate normal random vector. It follows from the previous assumption about the components that $\boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{Y}$ has covariance matrix $\boldsymbol{\Sigma}$. Now, $\boldsymbol{\Sigma}$ has eigenvalues $\lambda_1, \ldots, \lambda_p$ and corresponding eigenvectors $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p$. If the null hypothesis is true for the i^{th} observation, then for an arbitrary, j^{th} principal component, $\boldsymbol{\alpha}_j'\mathbf{x}_i$ is normally distributed with mean equal to zero and variance equal to λ_j. Thus, a test of the hypothesis could be done by testing whether the j^{th} principal component score came from a population with a mean equal to zero. Hawkins observes that in general larger magnitudes of the elements of $\boldsymbol{\alpha}_j$ will lead to a more powerful test of the hypothesis. However, the sums of the squares of the elements for any $\boldsymbol{\alpha}_j$ is the same, namely one. So, one would conclude in general that each of the principal component tests of the hypothesis are equally powerful.

Now, consider the weighted j^{th} principal component score: $\boldsymbol{\alpha}_j'\mathbf{x}/\sqrt{\lambda_j}$. The sum of the squares of the coefficients of the individual x variables is $1/\lambda_j$. So, because the coefficients in the linear combinations with small λ's will have a large sum of squares, Hawkins concludes that PCLOW (see Equation 2.24) is a powerful statistic for detecting mean-slippage outliers. And, we know that the distribution of the individual PCLOWs is an approximate chi-square random variable with $p-k$ degrees of freedom (recall that the variance of the j^{th} principal component score is λ_j and that these weighted squares of $N(0,1)$ random variables are independent by the same reasoning outlined in Section 2.3.3). So, a cut-off may be inferred from a χ^2 distribution with $p-k$ degrees of freedom (Hawkins, 1974).

I think that it is important to note that observations that have large PCLOW-values will also tend to have large values on Rao's metric when the Kaiser criterion is employed to determine how many principal components to retain. This is due to the fact that in PCLOW each term in the sum is square of the length of the projection onto the respective principal component divided by an eigenvalue that is less than one.

Because the principal components are extracted from the correlation matrix based on the entire sample, it stands to reason that PCLOW will be very susceptible to masking effects. I plan to introduce and investigate the performance of a robust version of PCLOW. I will discuss the computational details in Chapter 4.

2.4.3 Additional Research by Hawkins

Hawkins also discussed the special alternative hypothesis that just one of the components of the vector $\mathbf{e}_i$ is nonzero. Specifically, we have:

$$H_0: \quad e_{ij} \quad \text{for all } j$$

$$\begin{aligned} H_1: \quad & e_{ij} \neq 0, \quad \text{for some unknown } j \\ & e_{ik} = 0, \quad \text{for all } k \neq j. \end{aligned} \tag{2.25}$$

He asserts that in this case a powerful, or sensitive, statistic is $T_0 = \max(|x_{i1}|, |x_{i2}|, \ldots, |x_{ip}|)$. He mentions that Dixon showed that this statistic is more powerful than the Mahalanobis D^2 in testing H_0 against H_1 when the covariance matrix of $\mathbf{Y}$ is diagonal (Hawkins, 1974; Dixon, 1950).

Hawkins extends this result. Remember Hawkins' scaled squared j^{th} principal component score on observation i: $(\boldsymbol{\alpha}_j' \mathbf{x}_i)/\sqrt{\lambda_j}$. Let us call this value z_{ij}. We can denote the vector containing all of the scaled principal component scores as $\mathbf{Z}_i = \mathbf{B}\mathbf{x}_i$ where $\mathbf{B}$ is a $p \times p$ matrix with the j^{th} row equal to $\boldsymbol{\alpha}_j'/\sqrt{\lambda_j}$. Now, the value of Hawkins' statistic (2.24) which I have chosen to call PCLOW is invariant under orthogonal rotations. Let us consider an arbitrary orthogonal rotation matrix, $\mathbf{C}$. What if we choose $\mathbf{C}$ so that the $p-k$ rows corresponding to the small eigenvalues are rotated using the Varimax criterion commonly used in factor analysis. We now have $\mathbf{v}_i = \mathbf{C}\mathbf{Z}_i = \mathbf{C}\mathbf{B}\mathbf{x}_i$. According to Hawkins the simple structure imposed

by the Varimax rotation leads to an approximation of the conditions that Dixon discussed (namely that some of the components of $\mathbf{v}$ are approximately multiples of the components of $\mathbf{x}$ because of the rotation and they are also uncorrelated) (Dixon, 1950). Thus, a powerful test (more powerful than the Mahalanobis D^2) of the null against H_2 is $\max_{k+1\leq j\leq p}|v_{ij}|$, where the $p-k$ rotated eigenvectors that had the small eigenvalues are indexed by the numbers one through k (Hawkins, 1974).

2.4.4 Jolliffe

In his book, *Principal Components Analysis*, Jolliffe has a section entitled "Detection of Outliers Using Principal Components" (1986). In that section he states: "by the examining the values of the last few PCs, we may be able to detect observations which would violate the correlation structure imposed by the bulk of the data"(Jolliffe, 1986, p. 182) This is a significant statement for a couple of reasons. First, he is advocating the use of the last few PCs to spot "correlational "outliers. We shall see later in this paper that entire line of research has been dedicated to spotting "correlational" outliers, yet none of the researchers in that line seem to be aware of the work done by Jolliffe. Secondly, this statement is interesting because Hawkins came to the conclusion that the last few principal components are useful in spotting outliers of the mean-shift variety. Again, as we shall see later in this paper, a distinction will be made by researchers between mean-shift and correlational outliers, and it is suggested that different metrics be used to spot these different kinds of outliers. So, it is interesting that the last few principal components have been touted to spot both mean-shift and correlational outliers. Neither Jolliffe nor Hawkins seem to have been aware of this significant fact. Now let us explore the reasoning behind Jolliffe's assertion.

Jolliffe states that "A strong correlation structure between variables will imply that there are linear functions of the variables with small variances "(Jolliffe,

1986, p. 184). He considers this situation when there are just two variables. He uses height and weight as an example. Suppose we have a sample of observations on these two variables (observations in deviation form). In this case, there will certainly be a strong "correlation structure imposed by the bulk of the data "(Jolliffe, 1986, p. 184) He supposes that a regression of weight on height has been fitted. He notes that the form of the residual, which has a small variance, is similar to the form of the second principal component, which has the smallest variance for any linear combination. The residual and the second principal component both have the form $k_1x_2 - k_2x_1$ where k_1 and k_2 are both positive constants. If an observation has a large residual, we can say that it has deviated from the correlational structure of the data. Since both the residual and the second principal component both have small variances and a similar form, it stands to reason that if an observation is an outlier with respect to one, it is an outlier on the other as well. Therefore, observations that deviate from the correlational structure will be outliers on the second principal component. He states that the argument generalizes to the case when we have data with a dimensionality greater than two. It would seem that Jolliffe's reasoning would support the metric proposed by Rao (Equation 2.23), but Jolliffe points out some drawbacks.

Jolliffe points out that not enough weight is given to last few PCs. This especially true if the number of PCs used in Rao's metric is close to p. Remember that the very last few PCs have the smallest variances, so the principal components scores on the very last few will typically be small. This makes a case, albeit for different reasons than those espoused by Hawkins, for weighting the principal components scores by their corresponding variances (eigenvalues). Jolliffe goes on to add that it is the very last few principal components that are effective in spotting certain types of outliers. He does not elaborate on what these types are, but from his previous discussion, we can assume that correlational outliers are one of the types.

So, Jolliffe has produced an unique argument that Hawkins' PCLOW (2.24) is an effective outlier detection metric. As I mentioned before, this line of research comes to an end here. No one has ever done any kind of comparison of how Hawkin's PCLOW compares to other metrics that we shall discuss shortly. I plan to remedy this.

It should be noted that since all values are used in the estimation of the sample variance-covariance (or correlation) matrix it stands to reason that PCLOW will be susceptible to masking effects.

2.5 The Gnanadesikan and Kettenring Paper and its Twin Legacies

The 1972 *Biometrics* article by Gnanadesikan and Kettenring spawned two major lines of research. There is a quotation in the paper that had a profound influence on the work of applied statisticians in the behavioral sciences in the area of multivariate outlier detection:

> The complexity of of the multivariate case suggests that it would be fruitless to search for a truly omnibus outlier protection procedure. A more reasonable approach seems to be to tailor detection procedures to protect against different types of situations, e.g. correlation distortion, thus building up an arsenal of techniques with different sensitivities. This approach recognizes that an outlier for one purpose may not necessarily be one for another purpose! (1972, pp. 109–110)

This assertion that there are different kinds of multivariate outliers and that they are not all detectable by one metric spurred applied statisticians in the behavioral sciences to look for a metric to spot "correlational" outliers There is also an aside in the article which suggests a metric for detecting correlational outliers that no one

has tested.

Gnanadesikan and Kettenring also broached the topic of robust estimation of the mean vector and variance-covariance matrix for a sample of multivariate data. This spurred a line of research into robust distance metrics which rely on the robust estimation of center and dispersion. The reason for the interest in a robust distance metric is what is known as the masking effect (Gnanadesikan & Kettenring, 1972).

2.5.1 Masking Effect

Perhaps it easier to demonstrate the need for robust distance measures by first considering an univariate example. Consider Table 2.7. It is the all-time top-grossing movies (as of 2003) in descending order of total gross sales (Verzani, 2005). There is a problem with using the sample standard score as an outlier detection metric. It is susceptible to what is known as the masking effect. When there is a relatively large fraction of outliers and/or very very extreme observations, the estimates of $\overline{x}$ and s can be distorted to such an extent that outlying observations will fail to be identified by the sample standard score. Let us look at a boxplot of the data, Figure 2.7. The plot reveals five discordant observations. It should be noted that the five observations marked as outliers in the plot are legitimate observations. However, it is still instructive to know that these points are discordant. This information can indicate that the observations may merit special attention such as the application of a different label or category such as "super blockbuster." The boxplot uses the criterion that if an observation is more than 1.5 interquartile ranges beyond either the first or third quartile it is depicted as an outlier by plotting it as a circle beyond the 1.5 interquartile range "fence." However, if we apply the rule of flagging observations that are more than three standard scores above or below the mean, only *Titanic* is flagged as an outlier. The five outlying observations only represent a little over 6% of the data, so we see that the sample standard score is very

susceptible to the masking effect. The *breakdown point* is very low. Huber provides a layman's definition of the breakdown point of a statistic: "Roughly speaking, the breakdown point gives the limiting fraction of bad outliers the estimator can cope with ... The breakdown point of the α-trimmed mean is α (Huber, 1983, p. 81). The 1.5 interquartile range rule is said to be more robust than the standard score. Recall that the Mahalanobis D^2 is the the multivariate analog of the square of the sample standard score. Thus, it would stand to reason that it will be susceptible to masking effects as well, and in fact it is because it uses estimates of center and dispersion that are based on the entire sample. We shall consider metrics that are more robust that the Mahalanobis D^2 shortly, but first let us look at an example of the masking effect on the Mahalanobis D^2 using a famous data set.

Table 2.1: All-Time Top Grossing Movies

MOVIE	GROSS
Titanic	601
Star Wars	461
E.T.	435
Star Wars: The Phantom Menace	431
Spider-Man	404
Jurassic Park	357
The Two Towers	339
Forrest Gump	330
The Lion King	329
Harry Potter and the Socerer's Stone	318
The Fellowship of the Ring	313
Star Wars: Attack of the Clones	311

Continued on next page

MOVIE	GROSS
Return of the Jedi	309
Independence Day	306
The Sixth Sense	294
The Empire Strikes Back	290
Home Alone	286
Shrek	268
Harry Potter and the Chamber of Secrets	262
Jaws	260
How the Grinch Stole Christmas	260
The Matrix Reloaded	257
Monsters, Inc.	256
Batman	251
Men in Black	250
Toy Story 2	246
Raiders of the Lost Ark	242
Twister	242
My Big Fat Greek Wedding	241
Ghostbusters	239
Beverly Hills Cop	235
Cast Away	234
The Exorcist	233
The Lost World: Jurassic Park	229
Signs	228
Rush Hour 2	226

Continued on next page

MOVIE	GROSS
Mrs. Doubtfire	219
Ghost	218
Aladdin	217
Saving Private Ryan	216
Mission Impossible 2	215
Austin Powers in Goldmember	213
Back to the Future	212
X2: X-Men United	207
Austin Powers: The Spy Who Shagged Me	206
Terminator 2: Judgment Day	205
The Mummy Returns	202
Armageddon	202
Gone With the Wind	199
Pearl Harbor	199
Indiana Jones and the Last Crusade	197
Bruce Almighty	194
Toy Story	192
Finding Nemo	191
Men in Black II	190
Gladiator	188
Snow White and the Seven Dwarfs	185
Dances with Wolves	184
Batman Forever	184
The Fugitive	184

Continued on next page

MOVIE	GROSS
The Perfect Storm	183
What Women Want	183
Ocean's 11	183
Grease	181
Liar Liar	181
Mission Impossible	181
Jurassic Park III	181
Planet of the Apes	180
Indiana Jones and the Temple of Doom	180
Pretty Woman	178
Tootsie	177
Top Gun	177
There's Something About Mary	176
Ice Age	176
"Crocodile" Dundee	175
Apollo 13	174
Home Alone 2: Lost in New York	173
Air Force One	173
Rain Man	172

A Multivariate Example

Let us consider a subset of Fisher's Iris dataset 5.1. The original data set had four measurements: sepal length, sepal width, petal length and petal width on 150 flowers of three species: *setosa*, *versicolor* and *virginica*. There were fifty observations for each of three species. Let us consider the subsample displayed in Table 5.1. I have taken all of the *setosa* observations and added the first ten of the *versicolor*.

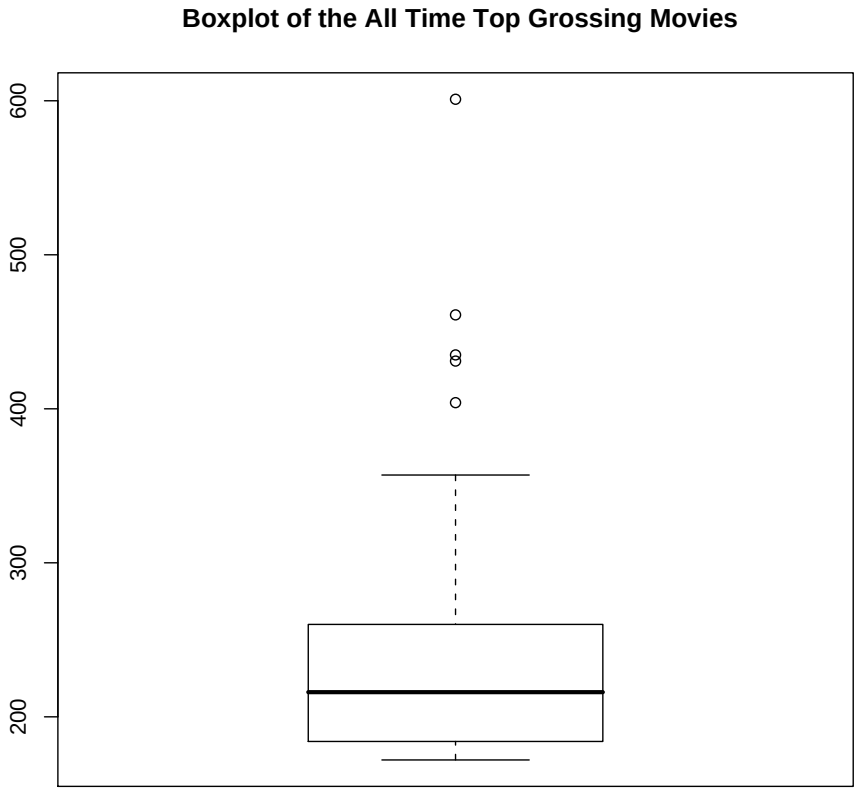

Figure 2.7: Boxplot of the 79 Top-Grossing Movies

Let us regard the *setosa* observations as legitimate and the *versicolor* observations contaminants (therefore outliers). It should be noted that a multivariate analysis of variance implies that these two species have differing mean vectors, so considering the *versicolor* observations as contaminants (outliers) is reasonable.

Let us apply the most commonly used outlier detection metric, the Mahalanobis D^2 to the data. Using the reasoning that the individual observations are approximate χ^2_4 random variables we could use the 99.9th percentile of the corresponding distribution. Recall that this is a suggested cut-off value found in a popular applied multivariate statistics book (Tabachnick & Fidell, 1996). Our cut-off value is $\chi^2_4(0.999) \doteq 18.4668$. None of the contaminants comes very close to this value. The maximum Mahalanobis D^2 for the contaminants is 12.3181, and the maximum value for the entire set is 12.6972 which came from observation 42, and this observation is legitimate. The contaminants (outliers) were masked by the fact that they constituted twenty percent of the sample and thus had an effect on the estimates of center and dispersion used in the computation of the Mahalanobis D^2.

Instead of using a cut-off value, a practioner may plot the Mahalanobis D^2's and determine potential outliers from the plot. So, let us look at a histogram of the values. This histogram appears in Figure 2.5.1. As we can see, no single observation or group of observations appears outlying.

2.5.2 Robust Distance Metrics

Robust distance metrics are Mahalanobis type distance metrics that are based on robust estimates of the mean vector and and/or variance-covariance matrix(Mahalanobis type in the sense that they are all of the form $(\mathbf{x} - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}})$.

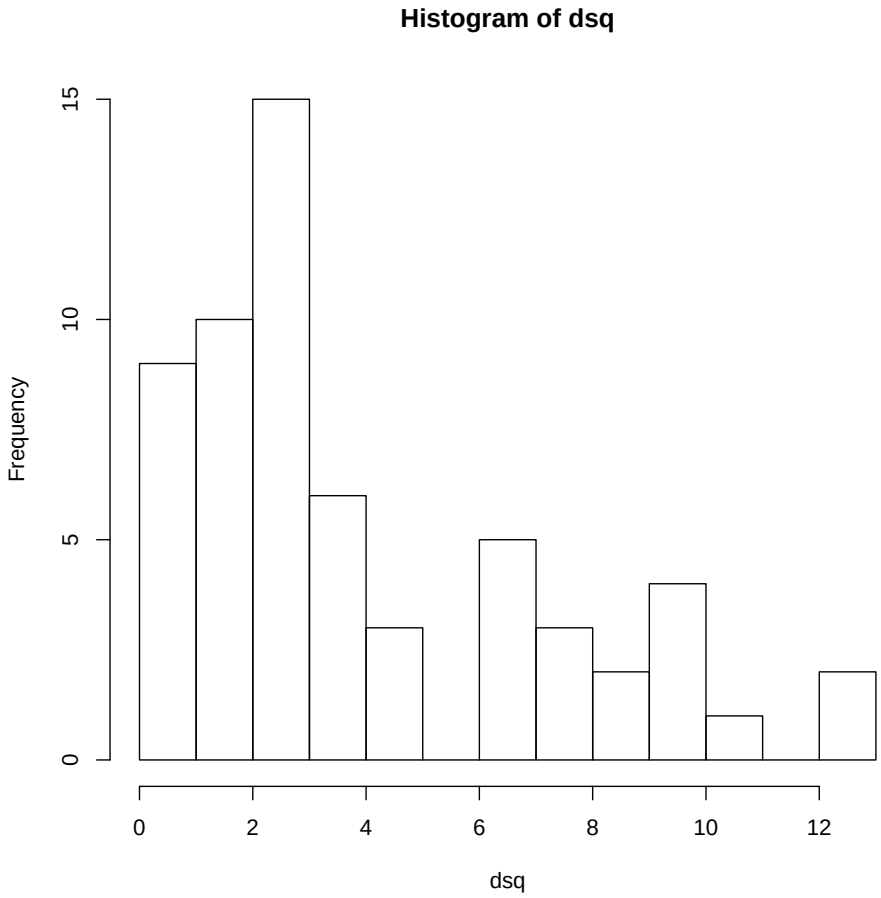

Figure 2.8: Histogram of Mahalanobis Distances for a Subset of the Iris Data

MVT

Probably the earliest method that is still used in practice to estimate multivariate variance-covariance is MVT (multivariate trimming). The amount of trim is represented by α. First, the Mahalanobis D^2 is computed for each observation. The observations with the smallest $100(1-\alpha)\%$ values on the Mahalanobis D^2 are retained. These observations are used to compute a location estimate, $\boldsymbol{\mu}^*$, and an estimate of the variance-covariance matrix, $\boldsymbol{\Sigma}^*$. On the second iteration, distances are computed for each observation using these new estimates. Then, the smallest $100(1-\alpha)\%$ are used to compute a new $\boldsymbol{\mu}*$ and $\boldsymbol{\Sigma}^*$. Iterations continue until the estimates of the elements of $\boldsymbol{\Sigma}^*$ stabilize according to a criterion such as Fisher z-transformed bivariate correlations changing by less than 10^{-3}. Outliers would be identified by computing a Mahalanobis D^2-like distance using the final values for $\mathbf{x}^*$ and $\boldsymbol{\Sigma}^*$. Since the distribution of these distances is highly complicated, it is difficult to establish cut-off values. However, graphical displays could be employed to spot discordant observations on the univariate distance metric. Egan and Morgan (1998) point out that by using Mahalanobis D^2 in the first step MVT is susceptible to the same masking effect that plagues the Mahalanobis D^2(Gnanadesikan & Kettenring, 1972; Carrig, 2005).

MCD and MVE

In 1985 Rousseeuw proposed two algorithms for finding robust estimates of multivariate variance-covariance structure, MVE (minimum volume ellipsoid) and MCD (minimum covariance determinant). The MVE algorithm seeks the smallest ellipsoid that covers at least $n/2+1$ data points from the sample (Rousseeuw & Zomeren, 1990; Carrig, 2005). The MCD algorithm seeks the subset of size $n/2$ whose covariance matrix has the smallest determinant. Woodruff and Rocke state: "The minimization required by MCD has the same objective function as the MVE, but

we constrain the covariance estimator, $\mathbf{C}$, by requiring the estimate be formed using half the points rather than it define an ellipsoid that has half "(Woodruff & Rocke, 1994, p.890). Specifically, MVE's objective is to find the pair $\mathbf{T}(\mathbf{X})$, a $p \times 1$ vector, and $C(\mathbf{X}$, a $p \times p$ positive semi-definite matrix where C has a minimum determinant with the constraint that

$$\#\{i; (\mathbf{x}_i - \mathbf{T})' \mathbf{C}^{-1} (\mathbf{x}_i - \mathbf{T}) \leq a^2\} \geq h \tag{2.26}$$

where h is the greatest integer less than or equal to $(n+p+1)/2$ Assuming only that the majority of data are good, a^2 is equal to $\chi^2_p(0.50)$. For this choice of a^2 and h, MVE has a breakdown point of nearly 50%. MCD's objective is to find the sample containing half the observations in which the determinant of $\mathbf{C}$ is a minimum. Since MVE has the additional restriction, algorithms for it require less computational time than thoose for MCD (Rousseeuw & Zomeren, 1990). One has to possess quite a bit of numerical analysis knowledge in order to comprehend the algorithms which find approximate solution.

SHV and RHM

In regard to MCD, MVE and other metrics, Egan and Morgan state: "While theoretically sound, the complexity of many proposed methods, in terms of comprehension and implementation, hampers their spread and use "(1998, p. 2373). Egan and Morgan offer two metrics "that are easy to understand, simple to implement and still handle the difficult problems that cause traditional methods to fail "(Egan & Morgan, 1998, p. 2374). These two methods are Mahalanobis D^2-like distances that are based on robust estimates of the mean vector and the variance-covariance matrix. They are RHM (resampling by half-means) and SHV (smallest half volume). RHM samples $n/2$ observations without replacement from the entire sample repeatedly. For each subsample, a mean vector, $\mathbf{m}_i$, is computed along with a vector

of standard deviations, $\mathbf{s}_i$. Next all of the sample values are scaled using components from $\mathbf{m}_i$ and $\mathbf{s}_i$. The lengths of these scaled vectors are computed. The n lengths corresponding to the i^{th} subsample form the i^{th} column of an $n \times s$ matrix where s is the number of subsamples taken. Each of the columns is sorted, and the observations appearing in the top 5% in terms of standard scores are tallied. At the completion of the subsampling process, for each observation, we have the number of times that observation appeared in the top 5% for a column. Potential outliers will stick out in the sense that they appeared in the top 5% a large number of times (Egan & Morgan, 1998). Note that this method is not sensitive to the correlational structure of the data.

In the application of SHV, each observation is represented as a vector in p space. The "distance "between two observations , $\mathbf{x}_i$ and $\mathbf{x}_j$ is the Euclidean length, or norm, of the difference vector, $\mathbf{x}_i - \mathbf{x}_j$. This length, or norm, is represented by $||\mathbf{x}_i - \mathbf{x}_j||$, where $(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j) = ||\mathbf{x}_i - \mathbf{x}_j||^2$. The distances between pairs of observations are placed in a distance matrix, $\mathbf{D}$, where $D_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||$. Obviously, the diagonal elements of the distance matrix equal zero. The columns of $\mathbf{D}$ are sorted in ascending order. The first $n/2$ sorted entries are summed for each column. The column with the smallest one of these sums is of interest. It represents the observation which is most similar to a subsample half the size of the original sample. If you think about it, the observation corresponding to this column represents a typical observation in a sense and is therefore an estimation of the center. And the subsample surrounding this observation represents the "smallest half volume. "Notice that this algorithm, MVE and MCD have analagous objectives where the volume in SHV is the sum of the Euclidean distances around a designated point. It is important to note that since Euclidean distances are used in this algorithm the correlational structure of the data is ignored—a disadvantage for this metric in the detection of purely correlational outliers.

It is readily apparent that the distributions of all of the robust distance metrics are highly complex; therefore, it is advisable to use graphical displays of the distance scores to spot potential outliers.

2.5.3 Metrics Designed to Detect Correlational Outliers

As mentioned previously, the two main concerns with using the Mahalanobis D^2 were susceptibility to masking effects and lack of sensitivity to all kinds of outliers. Gnanadesikan and Kettenring introduced the idea of a correlational outlier pointing out that there are some multivariate outliers that distort correlation estimates and some that do not (1972).

PCHIGH

In their 1972 *Biometrics* article, Gnanadesikan and Kettenring put forth a graphical technique that would supposedly expose correlational outliers. They asserted that observations that are outlying with respect to the first few principal components distort covariance or correlation estimates. I have created a plot (Figure 2.5.3), very similar to the one that appeared in Gnanadesikan and Kettenring's article, of some data that demonstrates their argument. The point marked by the red triangle is an obvious outlier. Notice that the horizontal and vertical axes define a coordinate system in the original variables: x_1 and x_2. Axes corresponding to the two principal components are inside of the plot but are not labeled. If we define a coordinate system using the axes corresponding to the two principal components, we see that the observation is outlying with respect to the first principal component but not necessarily so with respect to the second component. So, for the bivariate case, this is an example of the kind of outlier Gnanadesikan and Kettenring were referring to: "outlying with respect to the first few principal components" (there are only two principal components in this example, so outlying with respect to the first "few"

means outlying with respect to the first one). The outlier most certainly inflates variances and affects estimates of covariance or correlation. The sum of variances for the two variables without the outlier is 1.485; with it, the sum is 2.398. Without the outlier, the estimate of the correlation between the two variables is -0.040, and with the outlier the estimate is 0.392. This is one example, but is it always true that an observation which is outlying with respect to first few principal components distorts estimates of covariance or correlation? This question is something I shall address in my study. Back to Gnanadesikan and Kettenring. They suggest that, in the general case, points should be plotted with respect to the first few principal components, and they contend that points that are outlying in these plots are outliers that distort estimates of correlation (Gnanadesikan & Kettenring, 1972).

One can infer a quantitative method for spotting correlational outliers based on their suggestion. First, a determination is made of which principal components have substantial variances (again, it is assumed that this has been done via a SCREE plot or quantitative rule such as the Kaiser criterion). Next, imagine a hyperplane whose coordinate axes are the principal components (k of them) with substantial variances. The coordinates of the points in this sytem are the projections of the observation (in deviation form) onto the individual principal component axes. We can quantify how outlying a point is in this hyperplane by computing its statistical distance from the centroid (the origin cause the values are in deviation form). Because the principal components are orthogonal, the statistical distance takes the form:

$$d_i^2 = \frac{\mathbf{x}_i'\mathbf{e}_1\mathbf{e}_1'\mathbf{x}_i}{\lambda_1} + \cdots + \frac{\mathbf{x}_i'\mathbf{e}_k\mathbf{e}_k'\mathbf{x}_i}{\lambda_k} \tag{2.27}$$

where k is the number of principal components with substantial variances. Note that this has the same form of the Mahalanobis D^2 in the form where $\mathbf{S}^{-1}$ is expressed in the form of its spectral decomposition (see Equation 2.11). One can deduce using the same logic that was applied to the Mahalanobis D^2 and PCLOW that

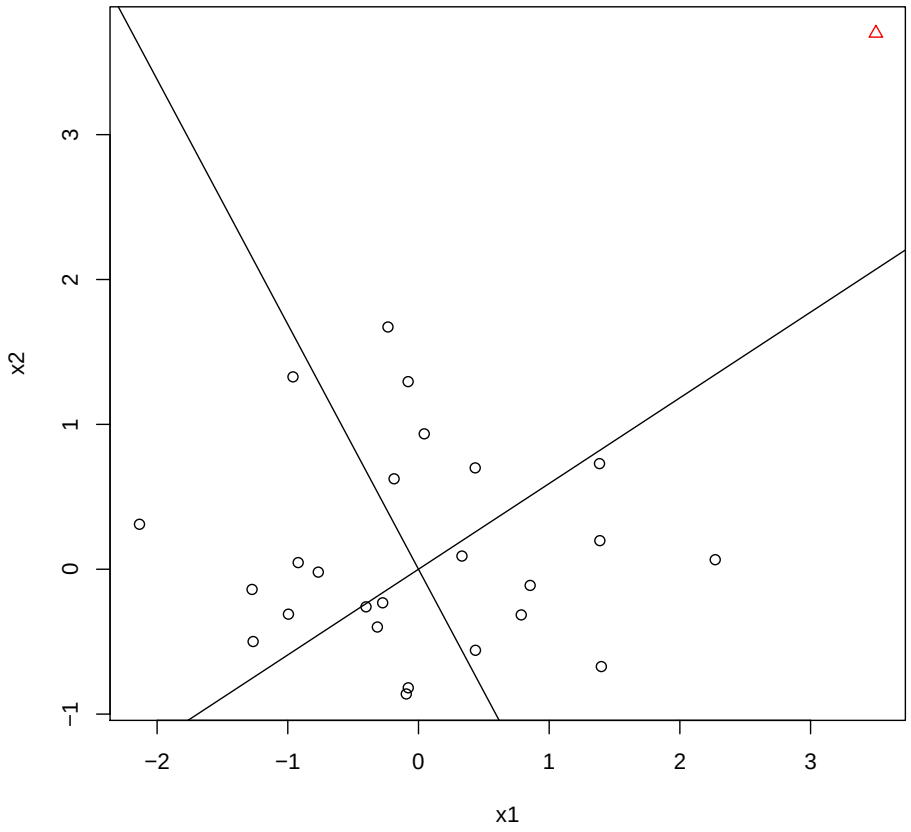

Figure 2.9: A Correlational Outlier Along with Principal Components Axes

the approximate distribution of the individual d_i^2's is χ_k^2 Henceforth, I shall refer to this metric as PCHIGH. Like PCLOW, it is a partial sum of the Mahalanobis D^2. Note that this metric is in direct opposition to Hawkins' PCLOW which was also touted by Jolliffe as a means of spotting correlational outliers. Which one of these metrics will spot correlational outliers? This is a question that has yet to be answered. The study that I will propose should help to resolve this matter. There is an interesting way to look at both PCHIGH and PCLOW. They both can be viewed as the Mahalanobis D^2 minus some noise that is not useful in the detection of multivariate outliers, but they are in direct opposition concerning the question of which partial sum of the Mahalanobis D^2 is noise and which is useful for detecting correlational outliers.

Again, as with PCLOW, one would suspect that PCHIGH would since succumb to masking effects. Because of this I shall use a robust version of PCHIGH in this study. The computation of PCHIGH shall be discussed in Chapter 4.

Comrey's Metric

Comrey, a psychologist, was concerned about the influence of correlational outliers on the results of factor analyses. In 1985 he developed a metric that purportedly measured an observation's total degree of correlational discordancy. It marked a unique turn in the literature. The form of the metric in no way resembles the Mahalanobis D^2. Before we look at the actual formula for the metric, let us explore Comrey's motivations.

> One problem with the Mahalanobis D^2 approach is that it may not identify those outliers that are distorting the correlation coefficients and hence affecting factor analytic results. Extreme values with large Mahalanobis distances may still fall close to the bivariate regression lines, thereby fitting in consistently with the correlations that would be ob-

tained without their presence (1985, p. 275)

Note that Comrey makes a case for the Mahalanobis D^2 generating false alarms when it comes to correlational outliers, but he fails to make a case against the Mahalanobis D^2 in regard to systematic generation of false negatives. Now let us consider a verbal description of the Comrey's D before looking at the formula.

The metric called Comrey's D that he proposed is "a method of detecting outliers that assesses directly the impact of possible outliers on factor analytic results by determining how deviant these cases are with respect to the underlying correlation matrix for the data "(Comrey, 1985, p. 275).

Mathematically, Comrey's D for the k^{th} observation is

$$D_k = \frac{2}{p(p-1)} \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} (r_{ij} - Z_{ki} Z_{kj})^2 \tag{2.28}$$

The rationale behind this formulation is straightforward. Because the Pearson product-moment correlation coefficient can be thought of as an average of cross-products of sample standard scores across two variables, the squared deviation of an observation's standard score product from the sample correlation coefficient, r, is a measure of an observation's correlational deviancy on the two variables in question. Looking at the formula we see that these squared deviations are summed over all of the $\frac{p(p-1)}{2}$ possible pairings of the p variables. Thus, the metric can be thought of as an average squared correlational deviation over all possible pairings of variables. D_k is a measure of correlational deviancy, but how does one determine what is an unusually large value on this metric?

Comrey erroneously assumed that the individual D_ks are normally distributed and this informed his decision about a cut-off value or what constitutes an unusually large value on this metric. He also gives no indication that he understands that when deciding whether an individual observation or group of observations is

outlying that he is dealing with order statistics. This is reflected by the fact that he suggest using the 95th percentile of a normal random variable as his cut-off value (Comrey, 1985).

Bacon pointed out some drawbacks of the Comrey's D. He demonstrated that it will systematically fail to detect some observations that are indeed correlational outliers. Consider the bivariate case where $\rho = 0.9$ and we have a standardized observation of $x_1 = 1/3$ and $x_2 = 3$. This observation would clearly show up as an outlier in a bivariate plot of a sample generated from a bivariate standard normal distribution with the aforementioned value for the parameter ρ. Assuming the sample correlation coefficient is close to the population value the Comrey's D for this observation would be 0.01, a very small value. Thus the Comrey's D would not detect a correlational outlier of this form (i.e., close to a coordinate axis) (Bacon, 1995).

Bacon also demonstrated that the Comrey's D will systematically produce false alarms under certain conditions. Again, consider a bivariate standard normal distribution with $\rho = 0.9$ that generated a sample containing as one of its observations: $x_1 = 3$ and $x_2 = 3$. More than likely, this point would not appear discordant in a plot of a sample from the aforementioned distribution, yet it would yield a Comrey's D of 65.61 (Bacon, 1995).

Carrig points out another drawback of using the Comrey D: Z_i, Z_j, and r are all computed using full sample estimates $\overline{x}$ and s. Thus, it will be susceptible to masking effects just like the Mahalanobis D^2.

Due to the fact that the distribution of even the individual observations on Comrey's D, much less its order statistics, spotting outliers on this metric via graphical displays would seem to be advisable.

Bacon's MLD

In 1995, Bacon proposed an alternative to Comrey's D that supposedly would not have the two drawbacks he exposed. The metric he proposed is called Bacon's Maximum Likelihood Distance or Bacon's MLD. The mathematical formula for the metric on the i^{th} observation is

$$MLD_i = \frac{2}{p(p-1)} \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} \frac{\left| \overline{MLCE}_{.jk} - MLCE_{ijk} \right|}{U_{ijk}}. \tag{2.29}$$

$MLCE_{ijk}$ is the maximum likelihood estimate of the population correlation coefficient, ρ, using only observation i (see Appendix A for a discussion and sample computation). $\overline{MLCE}_{.jk}$ is the individual maximum likelihood correlation estimates for variables j and k averaged over all observations. U_{ijk} is a weight known as the uncertainty and will be discussed below. Note that the form of the metric is similar to Comrey's D. It is an average deviation that has been summed over $\frac{p(p-1)}{2}$ possible pairings of variables. U_{ijk} is the "uncertainty" associated with the maximum likelihood estimate of the correlation between variables j and k using observation i only. Specifically, the mathematical formula for this uncertainty is

$$U_{ijk} = \sqrt{\frac{\sum_{q=1}^{m} L(p_{qjk})(p_q - MLCE_{ijk})^2}{\sum_{q=1}^{m} L(p_{qjk})}} \tag{2.30}$$

p_{qjk} takes on 250 equally spaced intermidiate values between -1 and 1. Thus, m equals 250.

Carrig points out that all observations are used in the computation of $\overline{MLCE}_{.jk}$. Therefore, Bacon's MLD is susceptible to masking effects (Carrig, 2005). Carrig proposed a metric of her own that would be purportedly sensitive to correlational outliers and not succomb to the masking effect.

Again, because of distributional difficulties graphical methods of identifica-

tion of outliers using this metric are advisable.

Carrig's D

Carrig's D is based on the following alternative formula for the sample correlation coefficient:

$$r_{jk} = 1 - \frac{1}{2}\left[\frac{\sum_{i=1}^{n}(z_{ij} - z_{ik})^2}{n}\right] \tag{2.31}$$

She neglects to mention that this is equivalent to the sum of the cross-products divided by n not $n-1$. This can be shown with some rather tedious algebra. Using this alternative formula as a basis, Carrig proposes the following discordancy metric:

$$\text{Carrig}D_i = \sum_{j=1}^{p-1}\sum_{k=j+1}^{p}\left|2(1 - r_{jk}) - (z_{ij} - z_{ik})^2\right|. \tag{2.32}$$

If we solve 2.31 for $\sum_{i=1}^{n}(z_{ij} - z_{ik})^2/n$, we get $2(1 - r_{jk})$. So $2(1 - r_{jk})$ is equal to an average of the squared differences of the sample standard scores averaged over all observations. Thus a measure of correlational deviancy on variables j and k for observation i would be $\left|2(1 - r_{jk}) - (z_{ij} - z_{ik})^2\right|$, and like Comrey's D and Bacon MLD, an observation's overall correlational discordancy is the sum of these individual discordancies over all possible pairings of variables. This metric should be free from the problem that Comrey's D has in spotting correlational outliers that lie near the axes of the variables as well as the problem of wrongly flagging observations that are not correlational outliers i.e. false alarms. Carrig's D should be free from these problems because it does not contain the product of Z scores which is responsible for the two problems with the Comrey's D. To avoid masking effects, Carrig added the refinement of using robust estimates of location, variance and correlation. The robust estimates will come from the application of the MCD algorithm (Carrig, 2005). Like the other metrics the best prescription would be to graph the observations' values on the metric and look for outliers in the plot.

2.6 Comparative Studies of Detection Metrics

Comparative studies have included comparisons of Comrey's D with the Mahalanobis D^2, the Mahalanobis D^2 with both Comrey's D and the Bacon MLD, and comparisons of the robust distance metrics with the Mahalanobis D^2. As mentioned earlier the work on robust distance measures and metrics designed to detect correlational outliers have been completely separate paths in the literature. A real breakthrough occurred with Carrig's dissertation in which the some of correlational detection metrics were directly compared with the robust distance measures.

2.6.1 Comrey

In the same article in which he put forth the Comrey's D as a metric for spotting correlational outliers, Comrey compared the performance of his metric with the Mahalanobis D^2. He used real world data gleaned from the administration of the Comrey Personality Scale to two samples. The two metrics were compared in each of the samples.

The first data set consisted of 185 observations on forty personality variables ($n = 185$, $p = 40$). The second data set consisted of 135 observations on the same forty variables. In deciding on cut-off values for the Mahalanobis D^2, he erroneously assumed that the Mahalanobis D^2 was distributed as a statistic that is a multiple of the Mahalanobis D^2 (see 2.8 in Section 2.3.1). He used the ninety-ninth percentile of this F distribution as his cut-off value. As was mentioned earlier, in Section 2.5.3, Comrey erroneously assumed that the distribution of individual observations on his metric were normally distributed. He used this false assumption to inform his choice of a cut-off value on his metric. He standardized observations on his metric and used the ninety-ninth percentile of the standard normal distribution as his cut-off value.

For the first data set, Comrey reported the Mahalanobis D^2 identified seventeen of the 185 observations as potential outliers, and the Comrey's D identified

twenty-one observations. The overlap between the two sets was fifty-three percent. Using his knowledge of the instrument and his knowledge of psychology, Comrey analyzed in more detail the observations that were flagged by either or both of the metrics. In his opinion, eight of these observations were "faked "(the personality scale was administered in conjunction with application for a driver license in Israel; therefore, there is some motivation for faked responses). Of these eight observations, six were identified by Comrey's D only. Only two of the eight were identified by the Mahalanobis D^2.

In the second set of data, the Mahalanobis D^2 identified fifteen of the 135 observations as potential outliers while the Comrey's D flagged twenty-one. Comrey reported an overlap of forty percent. The second data set resulted from an administration given to volunteer students at the University of California, Los Angeles. Comrey reasoned that there was not the same level of motivation for these subjects to fake their responses; therefore, the flagged observations "typically represented extreme profiles showing adjustment difficulties "(1985, p. 280).

In summary, Comrey's study demonstrates that his metric and the Mahalanobis D^2 are not largely redundant and that Comrey's D has some value as an informal quantitative outlier detection metric; at least with these two particular sets of data. The next study comparing these two metrics was a true Monte Carlo study.

2.6.2 Rassmussen

In 1988, Rassmussen conducted a Monte Carlo simulation study in which Comrey's D and the Mahalanobis D^2 were directly compared. The experimental design included four factors: outliers per sample, sample size, number of variables and the population bivariate correlation between each pair of variables (OS, N, M, ρ). OS had two levels: one and three. There were three levels of N: thirty, sixty and ninety. M, the number of variables, had four levels: two, three, four and five. The pop-

ulation correlation coefficient had four levels: 0.2, 0.4, 0.6 and 0.8. Two of the dependent variables were hit rate and false-alarm rate. The majority observations were generated from a multivariate normal distribution in which the mean was fifty and the standard deviation was five for each of the variables. The outliers also came from a multivariate normal population, but the population mean vector, $\boldsymbol{\mu}$, was different than the one for the legitimate observations. The individual elements of the mean vector for the outliers each differed from fifty by at least five and as much as fifteen. Each of the variates in the outlier population had a standard deviation of one. Rasmussen does not say anything in regard to the correlation matrix for the outlying observations. From what he did say, it was clear that the outliers could be classified as both mean-shift and variance deflation, but it is unclear whether they were correlational (Rassmussen, 1988).

As in Comrey's study, the Comrey's D was erroneously assumed to be normally distributed. The cut-off value applied to the Comrey's D scores corresponded to the ninety-fifth percentile of a standard normal random variable. Also like the Comrey study, the cut-off values for the Mahalanobis D^2 were ninety-fifth percentiles of a F distribution.

Overlap, computed as "the proportion of those cases identified as outliers which were labeled as such by both statistics "(Rassmussen, 1988, p. 193), ranged from 19.1% to 55.0% over all experimental conditions. Thus, there was pretty clear evidence that the two metrics were not largely redundant.

Very clear results were obtained in regard to hit rate. "The results indicated that the Mahalanobis D^2 detected a greater proportion of outliers than Comrey's D under all of the conditions simulated "(1988, p. 193). The overall difference in hit rate was 0.26. Predictably, hit rates increased with increasing sample size and increasing number of variables (Rassmussen, 1988). I conjecture that the former was due to a decreased masking effect and the latter to the increased degrees of freedom

with its concomitant increase in power. Overall, false-alarm rates were small, 3.3% for the Mahalanobis D^2 and 2.6% for the Comrey's D. Rassmussen also reported that these false-alarm rates were stable across the experimental conditions.

In summary, Rassmussen's study demonstrated that the two methods were not redundant to a large degree and that in the conditions simulated by the study the Mahalanobis D^2 was clearly superior to Comrey's D in terms of hit rate.

2.6.3 Bacon

In the same article in which he introduced the MLD, Bacon reported results from a Monte Carlo simulation study that compared the Mahalanobis D^2, Comrey's D and his Bacon MLD. Unlike Rassmussen's Monte Carlo study, the outliers that were generated were of the purely correlational variety. Recall that purely correlational outliers come from a distribution with the same mean vector as the legitimate observations but whose population correlation matrix is different that the one for the legitimate population. However, like Rasmussen's study, both contaminants and legitimate observations came from a population whose correlation matrix exhibited compound symmetry (i.e. all of the bivariate correlations are identical). Both legitimate and contaminant observations came from multivariate normal distributions (Bacon, 1995). Since he did not know how the individual values of his MLD metric were distributed, Bacon did not employ cut-off values in the application of any of the three metrics. Instead he used a trimming approach: he flagged observations corresponding to the top ten percent on the metrics. This is unrealistic especially given the fact that the proportion of outliers was a constant ten percent across all of his experimental conditions.

In his study, Bacon employed a 2 x 2 x 4 factorial design. The first factor was overall level of correlation with two levels: high and low. The second factor had two levels, lower and higher, which indicated whether the correlation among

legitimate variates was lower or higher than the correlation for the contaminant variates. The third factor was the number of variables. It had four levels: two, five, ten and twenty. Each replication within an experimental cell consisted of a sample of 100 observations, ninety of which came from the legitimate population and ten from the contaminant. There were 200 replications within each experimental condition. The high overall level of correlation and higher correlational level for the legitimate population conditions had a legitimate population correlation coefficientof 0.9 and a contaminant population correlation of 0.2. The low overall correlational level and higher correlational level for the legitimate population conditions had legitimate population correlation coefficients of 0.7 and contaminant correlations of zero. For the two conditions corresponding to legitimate correlations lower and overall correlation high and low, just switch the aformentioned correlations.

The Mahalanobis D^2 performed overwhelmingly better than Comrey's D when the correlations for the legitimate population were larger that those for the contaminants. The Comrey's D performed more poorly that would be expected given purely random selection of observations as outliers in some of the experimental conditions where the legitimate correlations were higher. Within this group of experimental conditions, the Mahalanobis D^2 performed increasingly better as the number of variables increased. I assert that this is due to increasing power. The performance of Bacon's MLD was in between the Comrey's D and the Mahalanobis D^2 in this class of experimental conditions where the legitimate correlations were higher than the contaminant correlations.

In the group of experimental conditions where the legitimate correlations were lower than the contaminant correlations, Bacon's MLD performed best followed by Comrey's D and the Mahalanobis D^2. In all of the experimental conditions of this type, the Mahalanobis D^2 performed at the level of chance or worse. While performing better than the Mahalanobis D^2 in these conditions, Bacon's MLD and

Comrey's D did not perform substantially better. Across all relevant experimental conditions in the group we are discussing, the Comrey's D had an overall hit rate of a little over sixteen percent. The average hit rate for the Bacon MLD in these conditions was only about twenty-two percent. In sum, none of the metrics performed well in the conditions where the legitimate correlation was lower than the contaminant correlation.

2.6.4 Egan and Morgan

Egan and Morgan conducted a Monte Carlo simulation that compared the performance of several robust distance metrics. Specifically, the authors compared MVT, MCD to the two metrics that they proposed on the basis of being less costly computationally and easier for the practitioner to understand: RHM and SHV. Hit rates were averaged over eight experimental conditions in which the number of variables ranged from five to fifty, sample size from twenty to 200 and fraction of outliers from 0.1 to 0.5. "Distance of outliers from the edge "and "spherical spread of outliers "were varied over the eight cells. No definition nor quantity was provided for these two different types of distances. Thus, it is unclear if the outliers were strictly mean-shift, purely correlational or some combination of the two. No indication was given of what if any cut-off values were used for each metric in flagging observations as potential outliers. It seems likely a trimming approach was applied. There were three measures of performance: percent of samples where there was perfect identification, percent where some outliers were missed and percent where there was total failure (i.e. not one single contaminant was identified). MVT performed very poorly. MVT with fifteen percent trimming perfectly identified the contaminants in only 3.0% of the samples. It had a total failure rate of 59.6%. MVT with 25% trimming perfectly identified the contaminants in 3.6% of the samples. Its total failure rate was 65.6%. MCD perfectly identified all contaminants in only 18% of all

samples, but it had a respectable total failure rate of 26.8%. RHM perfectly identified all of the contaminants in 37.4% of the samples and had a total failure rate close to MCD's. SHV perfect identification rate was 58.6% and had a total failure rate close to RHM and MCD. Egan and Morgan reported that the false-alarm rate was nearly the same for all of the metrics, but they did not report what this rate was (Egan & Morgan, 1998).

2.6.5 Novotny

Novotny's Monte Carlo study was distinctive in that she used factor analytic models as a starting point in the generation of both the legitimate and outlying data. Thus her study has bearing on the application of outlier detection metrics to psychometric data. The outliers that were generated were purely correlational in nature; I will elaborate on this shortly. The metrics that she compared that are of interest in this study are the Mahalanobis D^2, Comrey's D, individual log-likelihoods and factor scores. Each of the models used to generate legitimate and contaminant observations can be expressed as

$$\mathbf{X} - \boldsymbol{\mu} = \boldsymbol{\Lambda}\mathbf{F} + \boldsymbol{\epsilon} \tag{2.33}$$

where $\mathbf{X}$ and $\boldsymbol{\mu}$ are $p \times 1$ vectors. $\boldsymbol{\Lambda}$ is the $p \times m$ matrix of factor loadings; thus, there are m factors. And, $\boldsymbol{F}$ is the $m \times 1$ vector of values on the latent factors. $\mathbf{F}$ and $\boldsymbol{\epsilon}$ are uncorrelated multivariate normal vectors.

Novotny generated data to fit a one-factor model with six variables for the legitimate observations in each of the experimental conditions. All of the loadings in this model were 0.6. Since this is a one factor model the covariance matrix derived from Equation 2.33 is

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}. \tag{2.34}$$

where $\boldsymbol{\Psi}$ is the $p \times p$ diagonal matrix with specific variances as entries. Thus,

Novotny's one-factor model led to a covariance matrix in which all of the diagonal elements were the same, 1.16 The off-diagonal elements were all 0.36. The experimental conditions differed only in the factor model that was used to generate the contaminants. In each of the five experimental conditions, the sample size was $n = 200$, the proportion of outliers was 10% and the number of replicates was twenty. In the first experimental condition the outlying factor-analytic model was such that all of the off-diagonal elements in the resulting covariance matrix were equal and were a bit larger than those in the legitimate population. This experimental condition was known as Model A. The second experimental condition, Model B, produced essentially the same covariance matrix as Model A. Model C produced a covariance matrix with some covariances larger than those in the legitimate population and some smaller than those in the legitimate population. Model D led to a covariance matrix with equal off-diagonal elements which were smaller than those in the legitimate population. Model E represented the case where there was no covariance whatsoever between the six variables, i.e. the covariance matrix was the $p \times p$ identity matrix.

Cut-off values informed by the distributions of individual observations were used in the application of the Mahalanobis D^2 and the Comrey's D. She erroneously assumed that individual Comrey D's were normally distributed, and she also erroneously assumed that the individual Mahalanobis D^2's were distributed as a F random variable. However, in her application of the Mahalanobis D^2, she did realize that she was dealing with order statistics; she used a Bonferroni correction to her cut-off value. In her application of the Comrey's D, she used $\alpha = 0.05$ and $\alpha = 0.01$ critical values for a normally distributed random variable. In her application of the log-likelihoods and factor scores, she used a 10% trim.

The results for the Mahalanobis D^2 and Comrey's D with cut-off values informed by statistical distributions were abysmal. False-alarm rates for the Comrey's

D with $\alpha = 0.05$ and $\alpha = 0.01$ were 4% and 2.5% respectively. This rate was consistent across all experimental conditions. The hit rate in both applications was worse than it would be for chance identification of observations as potential outliers. These rates for the Comrey's D were consistent across all of the experimental conditions. Novotny did not publish the exact hit and false-alarm rates for the Mahalanobis D^2; she just noted that they were equally as bad. Based on these results, she decided to employ 10% trim approach with the Comrey's D and Mahalanobis D^2. The resulting hit and false-alarm rates from the appliction of the trimming approach for all of the metrics are discussed next.

Overall, the hit rates for all of the metrics were low. The highest rate occurring in an experimental condition was 23% for the log-likelihood in the condition where the contaminants were generated using model E. Averaging across all experimental conditions, the Mahalanobis D^2 and the individual log-likelihood had almost identical hit rates at 14% and 15%, respectively. The Comrey's D came in next at roughly 10% and then the factor scores at roughly 7%.

Like Bacon's study, the Mahalanobis D^2 performed best when variables on the legitimate observations had a higher correlation than those of the contaminants. However, the Comrey's D did not necessarily perform better when the contaminant correlations were higher than the legitimate ones, unlike Bacon's finding. It is interesting to note that all of the metrics showed their best performance when the contaminants were generated using Model E. The lone exception was the factor scores; they registered their poorest performance in this condition.

Novotny concluded that the application of any of the metrics will result in the identification of more legitimate observations as potential outliers than the contaminants. She states: "Overall, these results imply a grim prognosis for correlational outlier detection ". Perhaps these poor results were due to masking and swamping effects even though 10% seems to be a relatively low rate of contamination.

2.6.6 Carrig

Carrig performed a comparative study of the Mahalanobis D^2, MCD, Bacon's MLD, Carrig's D and SHV. The independent variables included proportion of outliers, with levels of 0.02 and 0.08; outlier separation, with levels "less "and "more "; covariance matrix shape separation; outlier type, with four levels: different shape with no mean slippage (a difference in the population mean vectors), different shape with close mean slippage and different shape with far mean slippage; number of variables (four and sixteen) and sample size (fifty and 200). Unfortunately, she did not provide quantitative measures of "less "and "more " shape separation nor did she quantitatively define "close "and "far "mean slippage. Her measures of performance included hit rate, percentage of samples with perfect identification and percentage of samples in which there was a total failure to correctly identify any of the outlying observations.

Three decision rules were used in the application of each metric. There was a natural drop off criterion which used a clustering algorithm, specifically, SAS's PROC FASTCLUS. This decision rule has the advantage of simulating the identification of outliers by graphical means. The second method employed a 5% trim rule. Finally cut values were used that were informed by previous simulations that investigated the distribution of the metric under the condition of no outliers (i.e. the null diistributions of the metrics). These simulations were a part of a pilot study conducted by Carrig. Another pilot study conducted by Carrig revealed that the natural drop-off criterion using PROC FASTCLUS was the best criterion: it had maximum hit rate and the minimum false-alarm rate. I shall now provide a brief description of SAS' PROC FASTCLUS.

PROC FASTCLUS uses a variant of a nonhierarchical clustering method known as the K-means method. The number of groups, K, is specified by the user. Essentially it assigns an observation to the group with the centroid closest to that

observation. In our case Euclidean distance will be used. In our application we have one-dimensional data, the one dimension being the score on the metric. The square of the difference between an observation's metric score and the mean of a group will be used as the measure of distance. In our case, K equals two because we want to separate the data into two groups: legitimate and outlying. There are three steps in a K means algorithm. I will outline the steps for one particular algorithm. The first step is the initial assignment of the observations to K groups such that each observation belongs to one and only one group. This step can be carried out in several different ways. In all cases, K "seeds "are specified. These seeds are the initial values for the centroids. They can be obtained by randomly sampling K observations that are a specified distance from one another. Step two involves taking each observation one at a time and assigning it to the group defined by having the centroid that is closest to the observation. If the observation has been moved from one group to another, the centroids of the two groups involved in this exchange are immediately updated to reflect the change. Step two is repeated until no observations are reassigned to different groups. (Khattree & Naik, 2000; Johnson & Wichern, 1992).

Carrig hypothesized that the proportion of outliers would reduce the performance of all the metrics but less so for the robust ones: MCD, SHV and the Carrig D. The hypothesis was true for the Mahalanobis D^2, MCD and the Carrig D. Additionally, Carrig reported that the MCD was less robust than expected. Also, the hypothesis did not hold for Bacon's MLD across all three outlier types (no apparent susceptibility to masking effects). Perhaps these last two results were due to the fact there was not a great enough range in her proportion of outliers condition (0.02 and 0.08).

Carrig hypothesized the conditions in which particular metrics would perform well and those in which they would perform poorly. In the conditions where the

fraction of outliers was high, she predicted that the robust statistics would have fewer false negatives. This will be due to lesser susceptibility to masking effects. She reported that the results provide only weak evidence for this. Again, perhaps there was not a sufficiently broad range in the fraction of outliers investigated. Carrig predicted that neither the classical nor the robust distance metrics (Mahalanobis D^2, MCD and SHV) would perform well in the conditions where there was no mean slippage and different covariance matrix shapes. This was confirmed for the Mahalanobis D^2. It had an overall hit rate of 48.2% in outlier condition one, 55.22% in outlier condition two and 69.94% in condition three. This was also true for the MCD: 50.52%, 61.56% and 77.00%, respectively. (Note that these hit rates are not substantially better for the MCD. Again a low upper level of 0.08 for the fraction of outliers factors may be responsible.) SHV did indeed perform poorly in outlier condition one (no mean-shift just differently shaped covariance matrices): 35.24% compared to its performance in the other two in which there was mean slippage: 51.10% and 72.45%. Note that the robust distance metrics did not perform substantially better than the Mahalanobis D^2. Again, this could be due to the fact that the highest fraction of outliers was 0.08. It was hypothesized that the Bacon MLD would perform better than the classical and robust distance metrics when the shapes of the covariance matrices diverge. The way to investigate this hypothesis is to look at the overall hit rates under outlier condition one (no mean slippage but different covariance matrix shapes). The Bacon MLD did have the highest hit rate under condition one: 53.27%. However, the hit rates for the Mahalanobis D^2 and MCD under this condition were 48.20% and 50.52%, respectively. Thus one could question whether the result is practically significant. Finally, Carrig hyptohesized that the Bacon MLD would be susceptible to masking effects. Carrig reported that fraction of outliers alone did not lead to decreased hit rates. Again, the reason could lie in the fact that the fraction of outliers factor had such a limited range. An

interesting result is the fact that contrary to her expectations SHV did not perform significantly worse when the sample data clouds were more elongated. This was expected because the procedure for computing SHV robust distances uses Euclidean distances. The next hypothesis has important implications for the general practice of detecting multivariate outliers.

Carrig hypothesized that the study would support the contention put forth by Gnanadesikan and Kettenring namely that "The complexity of the multivariate case suggests that it would be fruitless to search for a truly omnibus outlier protection procedure "(1972, p. 109). The results of the study supported this in that no single metric or combination of metrics emerged as being best across all experimental conditions. Next I shall discuss results pertaining to the metrics specifically designed to detect correlational outliers.

An interesting result emerged in regard to Bacon's MLD. Only under outlier condition one did increasing outlier separation lead to an increase in hit rate and a decrease in false alarm rate. Similarly, for the Carrig D, an increase in separation failed to predict an increase in hit rate. Now let us look at performance measures averaged across experimental conditions.

In terms of overall hit rate, the MCD robust distance was the best using the natural drop off criterion. MCD using natural drop off was also the best in terms of proportion of samples in which perfect identification was achieved. Overall, it achieved perfect identification in 30.48% of the samples. This was nearly three times the grand mean percentage for all metrics: 10.18%. MCD also had the lowest proportion of samples in which it failed to identify a single outlier: 11.73% compared to overall average of 37.63% for the rest of the metrics. It also had the best overall hit rate (65.05%). This hit rate was substantially better than the overall average for the rest of the metrics (29.79%). I think the overall performance is especially interesting in that all three of the outlier conditions involved some form of differing

correlational structure between the legitimate and outlying populations. However, it should be mentioned again that there was no single metric or combination of metrics which was found to perform best across all experimental conditions.

Chapter 3

Unanswered Questions

Throughout this chapter, I shall make references to the MCD robust distance and robust PCLOW and robust PCHIGH. Since PCHIGH and PCLOW are very likely very susceptible to masking effects, I shall examine robust version of PCLOW and PCHIGH instead of the regular PCLOW and PCHIGH. I shall discuss the computation of robust PCHIGH and robust PCLOW in Chapter 4

The scope of the enquiry in this study will be limited to the detection of purely correlational outliers. One reason is that there has been a whole line of research in quantitative psychology devoted to finding metrics to detect correlational outliers. Another reason is that Carrig's study revealed that currently we are least successful in spotting purely correlational outliers, so this is an area that begs for improvement.

Carrig's dissertation was ground breaking in that it compared some of the robust distance metrics against metrics specifically designed to spot correlational outliers. However, a drawback of her study was that both legitimate and outlying observations were generated from principal component models rather than factor analytic ones. Data arising from factor analytic models are far more typical in psychology. Novotny generated the data in her dissertation from factor analytic

models, but she did not include robust metrics nor did she include Bacon's MLD. So, we are left with the question of how the metrics shown to perform well in Carrig's study as well as PCHIGH and PCLOW perform under factor analytic conditions.

One of the most salient questions that arises from a review of the literature is whether Gnanadesikan and Kettenring were right in suggesting that the practitioner look at observations that are outlying with respect to the first few principal components if he is interested in spotting correlational outliers. Or is Jollife correct in suggesting that one look at observations that are outlying with respect to the last few principal components, or is it the case that neither is right? This is an interesting question given the fact that the two assertions are diammetrically opposed (Recall that if we substitute the spectral decomposition of $\mathbf{S}^{-1}$, Equation 2.10, into the formula for the Mahalanobis D^2 (Equation 2.5) we obtain Equation 2.12. Comparing the formulas for PCHIGH, Equation 2.27, and PCLOW, Equation 2.24, we see that touting PCHIGH implies that the part of the Mahalanobis D^2 corresponding to insubstantial principal components is noise. Contrastingly, the suggested use of PCLOW implies that the part of Mahalanobis D^2 corresponding to the first few, or substantial, principal components is noise). It should be noted here that in this investigation into the effectiveness of PCHIGH and PCLOW robust forms of these metrics will be used given that as each is a partial sum of the Mahalanobis D^2 they will most certainly be susceptible to masking effects.

There is also an interesting question related to the results of Bacon's study. Remember that the Mahalanobis D^2 did not fare well in experimental conditions where the population bivariate correlations for the legitimate observations were smaller than those for the contaminants. How will robust PCHIGH and/or robust PCLOW fare in this scenario given that they are each statistical distances like the Mahalanobis D^2?

I have a hypothesis as to which metric, rbust PCLOW or robust PCHIGH,

will spot correlational outliers. I agree with Jolliffe's assertion that PCLOW will be more effective in spotting correlational outliers, and I add a qualification conditions under which PCLOW will do an effective job. PCLOW will perform best when the substantial principal components for the legitimate population are orthogonal to the substantial principal components for the outlying population. I am also assuming in the illustration that follows that there is a large difference between the variances for the "substantial "principal components and the "insubstantial "principal components. PCLOW will perform increasingly poorly as the principal components of each population coincide. This is readily apparent with bivariate normal data. If the bivariate correlation coefficient for the legitimate population is $\rho_{\text{legit}} = 0.9$, and $\rho_{\text{out}} = -0.09$ for the outlying population, then is there is one substantial principal component (it accounts for 95% of the total variance) for the legitimate population lying along the 45 degree line, and there is one substantial principal component for the outlying population (also accounting for 95% of the variance) lying along the 135 degree line. Thus, the minor, or low, prinicipal component for the legitimate population is identical to the substantial, or high, principal component for the outlying population. Assuming we have a robust method of estimating the principal components, the outliers are going to lie close to and have large projections onto the second principal component. Perhaps this logic extends to the case of higher dimensions, and if so we would expect that PCLOW performs less well when the legitimate and outlying constant-density hyper-ellipsoids are not orthogonally oriented. The interesting question is whether Bacon's MLD or Carrig's D will outperform robust PCLOW in this situation. As I shall discuss in the Method Chapter 4, a step will be taken to answer this question.

It can be shown that in the bivariate case the principal components of the correlation matrix will always be the 45 degree and 135 degree lines. And, if one of the populations has its higher principal component accounting for considerably

less variance than the higher principal component for the other, then we have the situation that was modeled in Bacon's study. I am referring to the condition of the outlying population having higher or lower correlation than the legitimate population, and recall Bacon had very definite results for these two conditions.

If either PCHIGH or PCLOW is effective in spotting correlational outliers, it will be interesting to see how it compares across various experimental conditions to non-distance metrics, Carrig's D and Bacon's MLD, that were specifically designed to detect correlational outliers.

Another question involves an anomalous result in Carrig's study: Bacon's MLD performed better than Carrig's D in the experimental conditions corresponding to a high proportion of outliers. Recall that Carrig had two levels for the factor proportion of outliers: 0.02 and 0.08. I will investigate if this anomaly persists in the presence of higher proportions of outliers such as 0.16 and 0.32. A related question is whether the effect associated with increasing the proportion of outliers (again using levels higher than Carrig) will be the same for the differing metrics. The next question involves conjecture on my part.

We have demonstrated that the Mahalanobis D^2 rests on solid ground geometrically in that it is equivalent to the idea of "statistical distance ". PCLOW and PCHIGH are also statistical distances of a point's projection into the hyperplanes spanned by the insubstantial and substantial principal components, respectively. It is my belief that due to its elegance and simplicity the idea of statistical distance is superior to any other basis for defining discordancy. So, I am led to make the rather bold assertion that use of the combination of robust versions of PCLOW, PCHIGH and the MCD robust distance (if an observation is outlying on at least one of them, it is flagged as a potential outlier) will be the best across all of the experimental conditions in my study. I also extend this conjecture to all types of outliers: mean-shift, variance-inflation, correlational and all combinations thereof. Obviously, I cannot

test this grand conjecture completely in this study given that only purely correlational outliers will be simulated, but I plan to test it in later studies. This study is one step towards testing this conjecture. Note that this conjecture is antithetical to one of the tenets of multivariate outlier detection put forth by Gnanadesikan and Kettenring in their highly influential article: "it would be fruitless to search for a truly omnibus outlier detection procedure. A more reasonable approach seems to be to tailor detection procedures to protect against specific types of situations "(1972, pp. 109–110).

Given the limited range of the fraction of outliers in Carrig's study, the general question of how the metrics will compare in experimental conditions with a high fraction of outliers (e.g. 0.32) will be at least partially answered. I say partially because I am simulating correlational outliers only.

In scenarios where the number of substantial principal components, k, is small, there is the question of whether PCLOW will offer anything over and above a Mahalanobis-like distance. PCLOW is the Mahalanobis D^2 minus the terms corresponding to the k substantial principal components, so when k is small, their values will be very similar. And if so, is there a ratio of k to p at which PCLOW starts to offer something over and above the Mahalanobis D^2? I shall investigate the very similar question of whether robust PCLOW offers anything over and above the robust MCD distance. I plan to make a preliminary investigation into these questions by having experimental conditions with differing values of k with p held constant.

Finally, an interesting thing to investigate is the performance of non-Mahalanobis-like metrics, Bacon's MLD and Carrig's D, against the Mahalanobis D^2 and the robust distance metrics, MCD; robust PCLOW and the MCD robust distance, at a low n to p ratio. Rousseuw and van Zomeren recommend applying MVE only when $n/p > 5$ (Rousseeuw & Zomeren, 1990). Since MVE and MCD have such similar goals it stands to reason that this rule of thumb would apply to the use of the MCD

robust distance. It should apply to robust PCLOW and robust PCHIGH as well given that the principal components used are extracted from the corerlation matrix output by MCD. I would like to make the aforementioned comparison at $n/p = 5$ to see if the non-Mahalanobis-like distances are more or less robust to this condition.

I will use a Monte Carlo study with relevant factors, Legitimate and Outlier factor models; level of communality; sample size; number of variables; and fraction of outliers, to shed some light on the aforementioned questions. Six metrics, the Mahalanobis D^2, MCD, Bacon's MLD, Carrig's D, PCLOW and PCHIGH, will be compared in terms of hit rate and false-alarm rate across the experimental conditions.

Chapter 4

Method

4.1 Outlier Detection Methods

At the most fundamental level, my study entails a split-plot design with the six different outlier detection metrics being the repeated factor. Two measures of performance (the dependent variables), hit rate and false-alarm rate, were investigated, so there are really two split-plot studies. Levels of this repeated factor will be the Mahalanobis D^2; MCD robust, Mahalanobis-like distance; Bacon's MLD; Carrig's D; robust PCHIGH and robust PCLOW. The computation of the Mahalanobis D^2 will exactly follow equation 2.6. It is important to note that the observation $\mathbf{x}_i$ will be excluded from the computation of the multivariate mean, $\overline{\mathbf{x}}_j$, as this form requires. I have decided to exclude SHV from the the design due to the increasing availability of the MCD algorithm as a part of statistical packages such as SAS, R and S-Plus and the poor performance of SHV compared to MCD in Carrig's study. Its inclusion in these statistical packages seems to indicate that the computational expensiveness of forms of the algorithm is no longer an issue. The MCD robust,

Mahalanobis-like distance was computed according to the following equation:

$$D^2_{\text{rob}} = (\mathbf{x}_i - \overline{\mathbf{x}}^*)'(\mathbf{S}^*)^{-1}(\mathbf{x}_i - \overline{\mathbf{x}}^*), \tag{4.1}$$

where $\overline{\boldsymbol{x}}^*$ and $\mathbf{S}^*$ are the robust estimates of the population mean vector and variance-covariance matrix output by the application of the MCD algorithm through R's `cov.rob()` function. Bacon's MLD was computed in exactly the manner laid out by equations 2.29 and 2.30. Carrig's D was computed using Equation 2.32, and robust estimates of correlation and standard scores based on the output of the MCD algorithm were used. Despite the fact that Bacon's MLD outperformed Carrig's D in Carrig's study. The current study uses some considerably higher fraction of outliers conditions. Comrey's D was not included due to its poor performance in Rassmussen's, Bacon's and Novotny's studies. The robust versions of PCLOW and PCHIGH were computed using the principal components extracted from the correlation matrix output by the application of the MCD algorithm. There is precedence for the extraction of principal components from robust variance-covariance matrices. Chemometricians Walczak and Massart extracted principal components from robust variance-covariance matrices output by the MVT algorithm as a part of their study involving robust principal components regression (Walczak & Massart, 1995). Egan and Morgan demonstrated that MVT is susceptible to masking effects while MCD is not nearly as susceptible (Egan & Morgan, 1998). Given these two facts it seems reasonable to assume that it would be fruitful to extract principal components from a correlation matrix output by the MCD algorithm. There is a recently developed algorithm for the robust estimation of principal components but it is based on a complex algorithm involving MOP (marker object pursuit)(Hove, Liang, & Kvalheim, 1995). I will now discuss how the different metrics were applied.

Given a relatively large sample size, it seems reasonable that cut-off values for the MCD Distance, robust PCHIGH and robust PCLOW in addition to the

Mahalanobis D^2 can be inferred from chi-square distributions. There is a precedence for doing this. Rousseeu and van Zomeren used χ^2_p cut-off values in their paper which illustrated the use of an MVE algorithm. What they called the robust distance was a Mahalanobis distance using the esitmates of location and dispersion output by the algorithm. It would stand to reason that the same cut-off values that are used in the non-robust versions of these metrics can be used with the robust versions. In this study the 99.9th percentile (the recommendation put forth by Tabachnick and Fidell (1996)) of the corresponding χ^2 distributions was used as a cut-off value for flagging an observation as a potential outlier. I had to make a choice between doing this or using a K-means algorithm like Carrig did in her dissertation for separating the sample of metric values into two groups. I took the pessimistic view that more applied researchers would prefer the application of a cut-off to plotting the values on the metric or the application of both.

I have chosen to use a K-means algorithm (the one in the R statistical package) to separate univariate scores on the Bacon MLD and Carrig's D to separate the scores into two groups. The group with the larger values will be the collection of points that are flagged as being potential outliers. My decision was based on the fact that this method worked so well in Carrig's study. Also, I feel that this mimics the action of plotting values on the particular metric and deciding which ones should be flagged as potential outliers. The user of Bacon's MLD and/or Carrig's D has no choice but to use graphical methods given the fact that the distributions of even individual values on metrics are intractable.

4.2 Simulation Study Conditions

There are six between-subjects factors that are completely crossed. At the forefront of my experimental deign is the fact that both the legitimate observations and outlying observations will be generated using a factor analytic model. There was

a design factor called legitimate-outlier scenario. It will have two levels. Legitimate and outlying observations were generated according to the model in Equation 2.33. In the first level of this factor, the legitimate observations came from a model in which all variables load on one factor. The outlying observations at this level came from a model in which there were two uncorrelated latent factors with half loading on one factor and half on the other. The second level of legitimate-outlier scenario had the legitimate observations coming from a model in which there are four uncorrelated latent factors with a quarter of the variables loading on each of the factors. The outlying model at this level was the same as the one in the first level (two uncorrelated factors). I selected these two scenarios so that in level one the outlying model has twice as many factors as the legitimate and in level two the outlying model has half as many.

The second design factor is p, the number of variables. It will have two levels: eight and sixteen. When p is crossed with the legitimate-outlier scenario this enables us to see if the robust PCLOW is redundant in regard to the MCD robust distance for various combinations of p and k.

The next factor is n, sample size, with the levels: 80, 160, 320. The crossing of this with p will shed some light on whether the non-Mahalanobis-like distance metrics are more or less robust compared to the distance metrics at low levels of n/p. The fourth factor is very similar to the factor that was found to be so important in Bacon's study, whether the legitimate population had higher or lower intercorrelations than the outlying. Since we are generating the data from factor models a factor very similar to this is whether the communalities of the legitimate population are the same, higher or lower than those for the outlying model. The communalities for each of the variables will be the same. This factor enables us to see how robust the metrics are to the condition of outlying variables having higher population correlations than the legitimate. Remember the metrics that Bacon

Design	
Between Samples	
Factor	Levels
communality	higher, high-same, lower, low-same
scenario	I and II
number of variables	8 and 16
sample size	80, 160, 320
fraction of outliers	0.08, 0.16, 0.32
samples within cells	1 to 100
Within Samples	
Factor	Levels
metric	Mahal D^2, MCD, PCHIGH, PCLOW, Bacon MLD, Carrig D

Table 4.1: Experimental Design

compared performed very poorly in this condition.

The fifth condition will be proportion of outliers. This enables us to see if the anomalous result in Carrig's study, Bacon's MLD being more robust to a higher fraction of outliers than Carrig's D, holds for higher proportions. There will be four levels of this factor: 0.0, 0.08, 0.16 and 0.32. Finally, the sixth between-samples factor is samples within the communality—scenario—number-of-variables—sample-size—fraction-of-outliers cells.

The crossing of legitimate-outlier scenario with number of variables and communality difference produces sixteen different combinations of correlation matrices for the legitimate observations and the outliers. I will now enumerate these sixteen combinations.

Legitimate-Outlier Scenario I, p = 8, Communality Higher

$$\mathbf{\Lambda}'_{\mathrm{L}} = \begin{bmatrix} 0.8 & 0.8 & 0.8 & 0.8 & 0.8 & 0.8 & 0.8 & 0.8 \end{bmatrix} \tag{4.2}$$

$$\boldsymbol{\rho}_{\mathrm{L}} = \begin{bmatrix} 1 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 \\ 0.64 & 1 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 \\ 0.64 & 0.64 & 1 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 \\ 0.64 & 0.64 & 0.64 & 1 & 0.64 & 0.64 & 0.64 & 0.64 \\ 0.64 & 0.64 & 0.64 & 0.64 & 1 & 0.64 & 0.64 & 0.64 \\ 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 1 & 0.64 & 0.64 \\ 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 1 & 0.64 \\ 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 0.64 & 1 \end{bmatrix} \tag{4.3}$$

$$\boldsymbol{\Lambda}'_{\mathrm{O}} = \begin{bmatrix} 0.4 & 0.4 & 0.4 & 0.4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.4 & 0.4 & 0.4 & 0.4 \end{bmatrix} \tag{4.4}$$

$$\boldsymbol{\rho}_{\mathrm{O}} = \begin{bmatrix} 1 & 0.16 & 0.16 & 0.16 & 0 & 0 & 0 & 0 \\ 0.16 & 1 & 0.16 & 0.16 & 0 & 0 & 0 & 0 \\ 0.16 & 0.16 & 1 & 0.16 & 0 & 0 & 0 & 0 \\ 0.16 & 0.16 & 0.16 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0.16 & 0.16 & 0.16 \\ 0 & 0 & 0 & 0 & 0.16 & 1 & 0.16 & 0.16 \\ 0 & 0 & 0 & 0 & 0.16 & 0.16 & 1 & 0.16 \\ 0 & 0 & 0 & 0 & 0.16 & 0.16 & 0.16 & 1 \end{bmatrix} \tag{4.5}$$

Legitimate-Outlier Scenario I, p = 16, Communality Higher

$$\boldsymbol{\Lambda}_{\mathrm{L}} = (0.8)\mathbf{1}_{16\times 1} \tag{4.6}$$

$$\boldsymbol{\rho}_{\mathrm{L}} = 16 \text{ x } 16 \text{ matrix where } \rho_{ii} = 1 \text{ and } \rho_{ij} = 0.64 \; i \neq j \tag{4.7}$$

$$\boldsymbol{\Lambda}_{\mathrm{O}} = \begin{bmatrix} (0.4)\mathbf{1}_{8\times 1} & \mathbf{0}_{8\times 1} \\ \mathbf{0}_{8\times 1} & (0.4)\mathbf{1}_{8\times 1} \end{bmatrix} \tag{4.8}$$

$$\boldsymbol{\rho}_{\mathrm{O}} = \begin{bmatrix} \mathbf{A} & \mathbf{0}_{8\times 8} \\ \mathbf{0}_{8\times 8} & \mathbf{A} \end{bmatrix} \tag{4.9}$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & 0.16 & 0.16 & 0.16 & 0.16 & 0.16 & 0.16 & 0.16 \\ 0.16 & 1 & 0.16 & 0.16 & 0.16 & 0.16 & 0.16 & 0.16 \\ 0.16 & 0.16 & 1 & 0.16 & 0.16 & 0.16 & 0.16 & 0.16 \\ 0.16 & 0.16 & 0.16 & 1 & 0.16 & 0.16 & 0.16 & 0.16 \\ 0.16 & 0.16 & 0.16 & 0.16 & 1 & 0.16 & 0.16 & 0.16 \\ 0.16 & 0.16 & 0.16 & 0.16 & 0.16 & 1 & 0.16 & 0.16 \\ 0.16 & 0.16 & 0.16 & 0.16 & 0.16 & 0.16 & 1 & 0.16 \\ 0.16 & 0.16 & 0.16 & 0.16 & 0.16 & 0.16 & 0.16 & 1 \end{bmatrix} \tag{4.10}$$

Legitimate-Outlier Scenario II, p = 8, Communality Higher

$$\boldsymbol{\Lambda}_{\mathrm{L}} = \begin{bmatrix} 0.8 & 0 & 0 & 0 \\ 0.8 & 0 & 0 & 0 \\ 0 & 0.8 & 0 & 0 \\ 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 0.8 \\ 0 & 0 & 0 & 0.8 \end{bmatrix} \tag{4.11}$$

$$\boldsymbol{\rho}_{\mathrm{L}} = \begin{bmatrix} 1 & 0.64 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.64 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.64 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.64 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0.64 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.64 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0.64 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.64 & 1 \end{bmatrix} \tag{4.12}$$

$$\boldsymbol{\Lambda}_{\mathrm{O}} = \text{same as Equation (4.4)} \tag{4.13}$$

$$\boldsymbol{\rho}_{\mathrm{O}} = \text{same as Equation (4.5)} \tag{4.14}$$

Legitimate-Outlier Scenario II, p = 16, Communality Higher

$$\boldsymbol{\Lambda}_{\mathrm{L}} = \begin{bmatrix} (0.8)\mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (0.8)\mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & (0.8)\mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & (0.8)\mathbf{1} \end{bmatrix} \tag{4.15}$$

where $\mathbf{1}$ and $\mathbf{0}$ are 4×1 vectors.

$$\boldsymbol{\rho}_{\mathrm{L}} = \begin{bmatrix} \mathbf{A} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A} \end{bmatrix} \tag{4.16}$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & 0.64 & 0.64 & 0.64 \\ 0.64 & 1 & 0.64 & 0.64 \\ 0.64 & 0.64 & 1 & 0.64 \\ 0.64 & 0.64 & 0.64 & 1 \end{bmatrix} \tag{4.17}$$

and $\mathbf{0}$ is 4×4 matrix of all zeros.

$$\boldsymbol{\Lambda}_{\mathrm{O}} = \text{same as equation (4.8)} \tag{4.18}$$

$$\boldsymbol{\rho}_{\mathrm{O}} = \text{same as equation (4.9)} \tag{4.19}$$

Legitimate-Outlier Scenario I, p = 8, Communality Lower

$$\boldsymbol{\Lambda}_{\mathrm{L}} = (0.4)\mathbf{1}_{8\times 1} \tag{4.20}$$

$\boldsymbol{\rho}_{\mathrm{L}} = 8 \times 8$ correlation matrix with all off-diagonal elements equal to 0.16 (4.21)

$$\boldsymbol{\Lambda}_{\mathrm{O}} = \begin{bmatrix} (0.8)\mathbf{1} & \mathbf{0} \\ \mathbf{0} & (0.8)\mathbf{1} \end{bmatrix} \tag{4.22}$$

where both $\mathbf{1}$ and $\mathbf{0}$ are both 4×1 vectors.

$$\boldsymbol{\rho}_{\mathrm{O}} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} \end{bmatrix} \tag{4.23}$$

where

$$\mathbf{A} = \begin{bmatrix} 1 & 0.64 & 0.64 & 0.64 \\ 0.64 & 1 & 0.64 & 0.64 \\ 0.64 & 0.64 & 1 & 0.64 \\ 0.64 & 0.64 & 0.64 & 1 \end{bmatrix} \tag{4.24}$$

Legitimate-Outlier Scenario I, p = 16, Communality Lower

$$\boldsymbol{\Lambda}_{\mathrm{L}} = (0.4)\mathbf{1}_{16\times 1} \tag{4.25}$$

$\boldsymbol{\rho}_{\mathrm{L}} = 16 \times 16$correlation matrix with all off-diagonal elements equal to 0.16 (4.26)

$$\boldsymbol{\Lambda}_{\mathrm{O}} = \begin{bmatrix} (0.8)\mathbf{1}_{8\times 1} & \mathbf{0}_{8\times 1} \\ \mathbf{0}_{8\times 1} & (0.8)\mathbf{1}_{8\times 1} \end{bmatrix} \tag{4.27}$$

$$\boldsymbol{\rho}_{\mathrm{O}} = \begin{bmatrix} \mathbf{A} & \mathbf{0}_{8\times 8} \\ \mathbf{0}_{8\times 8} & \mathbf{A} \end{bmatrix} \tag{4.28}$$

where $\mathbf{A}$ is an 8×8 correlation matrix with all of the off-diagonal elements equal to 0.64

Legitimate-Outlier Scenario II, p = 8, Communality lower

$$\boldsymbol{\Lambda}_{\mathrm{L}} = \begin{bmatrix} 0.4 & 0 & 0 & 0 \\ 0.4 & 0 & 0 & 0 \\ 0 & 0.4 & 0 & 0 \\ 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0.4 \\ 0 & 0 & 0 & 0.4 \end{bmatrix} \tag{4.29}$$

$$\boldsymbol{\rho}_{\mathrm{L}} = \begin{bmatrix} \mathbf{A} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A} \end{bmatrix} \tag{4.30}$$

$$\boldsymbol{\Lambda}_{\mathrm{O}} = \text{same as equation 4.22} \tag{4.31}$$

$$\boldsymbol{\rho}_{\mathrm{O}} = \text{same as equation 4.23} \tag{4.32}$$

Legitimate-Outlier Scenario II, p = 16, Communality Lower

$$\boldsymbol{\Lambda}_{\mathrm{L}} = \begin{bmatrix} (0.4)\mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (0.4)\mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & (0.4)\mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & (0.4)\mathbf{1} \end{bmatrix} \tag{4.33}$$

where **1** and **0** are 4 x 1 vectors.

$$\boldsymbol{\rho}_{\mathrm{L}} = \begin{bmatrix} \mathbf{A} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{A} \end{bmatrix} \tag{4.34}$$

$$\boldsymbol{\Lambda}_{\mathrm{O}} = \text{same as equation 4.28} \tag{4.35}$$

$$\boldsymbol{\rho}_{\mathrm{O}} = \text{same as equation 4.28} \tag{4.36}$$

Legitimate-Outlier Scenario I, p = 8, Communalities High and Equal

$$\boldsymbol{\Lambda}_{\mathrm{L}} = \text{same as equation 4.2} \tag{4.37}$$

$$\boldsymbol{\rho}_{\mathrm{L}} = \text{same as equation 4.3} \tag{4.38}$$

$$\boldsymbol{\Lambda}_{\mathrm{O}} = \text{same as equation 4.22} \tag{4.39}$$

$$\boldsymbol{\rho}_{\mathrm{O}} = \text{same as equation 4.23} \tag{4.40}$$

Legitimate-Outlier Scenario I, p = 16, Communalities High and Equal

$$\boldsymbol{\Lambda}_{\mathrm{L}} = \text{same as equation 4.6} \tag{4.41}$$

$$\boldsymbol{\rho}_{\mathrm{L}} = \text{same as equation 4.7} \tag{4.42}$$

$$\boldsymbol{\Lambda}_{\mathrm{O}} = \text{same as equation 4.27} \tag{4.43}$$

$$\boldsymbol{\rho}_{\mathrm{O}} = \text{same as equation 4.28} \tag{4.44}$$

Legitimate-Outlier Scenario II, p = 8, Communalities High and Equal

$$\boldsymbol{\Lambda}_{\mathrm{L}} = \text{same as equation 4.11} \tag{4.45}$$

$$\boldsymbol{\rho}_{\mathrm{L}} = \text{same as equation 4.12} \tag{4.46}$$

$$\boldsymbol{\Lambda}_{\mathrm{O}} = \text{same as equation 4.22} \tag{4.47}$$

$$\boldsymbol{\rho}_{\mathrm{O}} = \text{same as equation 4.23} \tag{4.48}$$

Legitimate-Outlier Scenario II, p = 16, Communalities High and Equal

$$\boldsymbol{\Lambda}_{\mathrm{L}} = \text{same as equation 4.15} \tag{4.49}$$

$$\boldsymbol{\rho}_{\mathrm{L}} = \text{same as equation 4.16} \tag{4.50}$$

$$\boldsymbol{\Lambda}_{\mathrm{O}} = \text{same as equation 4.27} \tag{4.51}$$

$$\boldsymbol{\rho}_{\mathrm{O}} = \text{same as equation 4.28} \tag{4.52}$$

Legitimate-Outlier Scenario I, p = 8, Communalities Low and Equal

$$\boldsymbol{\Lambda}_{\mathrm{L}} = \text{same as equation 4.20} \tag{4.53}$$

$$\boldsymbol{\rho}_{\mathrm{L}} = \text{same as equation 4.21} \tag{4.54}$$

$$\boldsymbol{\Lambda}_{\mathrm{O}} = \text{same as equation 4.4} \tag{4.55}$$

$$\boldsymbol{\rho}_{\mathrm{O}} = \text{same as equation 4.5} \tag{4.56}$$

Legitimate-Outlier Scenario I, p = 16, Communalities Low and Equal

$$\boldsymbol{\Lambda}_{\mathrm{L}} = \text{same as equation 4.25} \tag{4.57}$$

$$\boldsymbol{\rho}_{\mathrm{L}} = \text{same as equation 4.26} \tag{4.58}$$

$$\boldsymbol{\Lambda}_{\mathrm{O}} = \text{same as equation 4.8} \tag{4.59}$$

$$\boldsymbol{\rho}_{\mathrm{O}} = \text{same as equation 4.9} \tag{4.60}$$

Legitimate-Outlier Scenario II, p = 8, Communalities Low and Equal

$$\boldsymbol{\Lambda}_{\mathrm{L}} = \text{same as equation 4.29} \tag{4.61}$$

$$\boldsymbol{\rho}_{\mathrm{L}} = \text{same as equation 4.30} \tag{4.62}$$

$$\boldsymbol{\Lambda}_{\mathrm{O}} = \text{same as equation 4.4} \tag{4.63}$$

$$\boldsymbol{\rho}_{\mathrm{O}} = \text{same as equation 4.5} \tag{4.64}$$

Legitimate-Outlier Scenario II, p = 16, Communalities Low and Equal

$$\boldsymbol{\Lambda}_{\mathrm{L}} = \text{same as equation 4.33} \tag{4.65}$$

$$\boldsymbol{\rho}_{\mathrm{L}} = \text{same as equation 4.34} \tag{4.66}$$

$$\boldsymbol{\Lambda}_{\mathrm{O}} = \text{same as equation 4.8} \tag{4.67}$$

$$\boldsymbol{\rho}_{\mathrm{O}} = \text{same as equation 4.9} \tag{4.68}$$

Chapter 5

Examples

5.1 Fisher Iris Data

Let us see how some of the metrics perform on a real set of data. I shall use a subset of the famous Fisher Iris Data Set taken from Johnson and Wichern's multivariate textbook (1992). The data set contained fifty observations from each of three species of iris: *Iris setosa, Iris versicolor* and *Iris virginica*. Each observation was comprised of four measurements: sepal length, sepal width, petal length and petal width. The subset that will be used for an illustration of the performance of the six metrics studied consists of all fifty of the *Iris setosa* observations and ten of the *Iris versicolor* observations which will be considered contaminants. Thus, the proportion of contaminants is approximately 0.17.

Table 5.1: Subset of Fisher Iris Data

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	*setosa*
4.9	3.0	1.4	0.2	*setosa*

Continued on next page

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
4.7	3.2	1.3	0.2	*setosa*
4.6	3.1	1.5	0.2	*setosa*
5.0	3.6	1.4	0.2	*setosa*
5.4	3.9	1.7	0.4	*setosa*
4.6	3.4	1.4	0.3	*setosa*
5.0	3.4	1.5	0.2	*setosa*
4.4	2.9	1.4	0.2	*setosa*
4.9	3.1	1.5	0.1	*setosa*
5.4	3.7	1.5	0.2	*setosa*
4.8	3.4	1.6	0.2	*setosa*
4.8	3.0	1.4	0.1	*setosa*
4.3	3.0	1.1	0.1	*setosa*
5.8	4.0	1.2	0.2	*setosa*
5.7	4.4	1.5	0.4	*setosa*
5.4	3.9	1.3	0.4	*setosa*
5.1	3.5	1.4	0.3	*setosa*
5.7	3.8	1.7	0.3	*setosa*
5.1	3.8	1.5	0.3	*setosa*
5.4	3.4	1.7	0.2	*setosa*
5.1	3.7	1.5	0.4	*setosa*
4.6	3.6	1.0	0.2	*setosa*
5.1	3.3	1.7	0.5	*setosa*
4.8	3.4	1.9	0.2	*setosa*
5.0	3.0	1.6	0.2	*setosa*

Continued on next page

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.0	3.4	1.6	0.4	*setosa*
5.2	3.5	1.5	0.2	*setosa*
5.2	3.4	1.4	0.2	*setosa*
4.7	3.2	1.6	0.2	*setosa*
4.8	3.1	1.6	0.2	*setosa*
5.4	3.4	1.5	0.4	*setosa*
5.2	4.1	1.5	0.1	*setosa*
5.5	4.2	1.4	0.2	*setosa*
4.9	3.1	1.5	0.2	*setosa*
5.0	3.2	1.2	0.2	*setosa*
5.5	3.5	1.3	0.2	*setosa*
4.9	3.6	1.4	0.1	*setosa*
4.4	3.0	1.3	0.2	*setosa*
5.1	3.4	1.5	0.2	*setosa*
5.0	3.5	1.3	0.3	*setosa*
4.5	2.3	1.3	0.3	*setosa*
4.4	3.2	1.3	0.2	*setosa*
5.0	3.5	1.6	0.6	*setosa*
5.1	3.8	1.9	0.4	*setosa*
4.8	3.0	1.4	0.3	*setosa*
5.1	3.8	1.6	0.2	*setosa*
4.6	3.2	1.4	0.2	*setosa*
5.3	3.7	1.5	0.2	*setosa*
5.0	3.3	1.4	0.2	*setosa*

Continued on next page

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
7.0	3.2	4.7	1.4	*versicolor*
6.4	3.2	4.5	1.5	*versicolor*
6.9	3.1	4.9	1.5	*versicolor*
5.5	2.3	4.0	1.3	*versicolor*
6.5	2.8	4.6	1.5	*versicolor*
5.7	2.8	4.5	1.3	*versicolor*
6.3	3.3	4.7	1.6	*versicolor*
4.9	2.4	3.3	1.0	*versicolor*
6.6	2.9	4.6	1.3	*versicolor*
5.2	2.7	3.9	1.4	*versicolor*

Now let us look at how the metrics performed on the data. Each observation's values on the six metrics are presented in Table 5.5. The fact that an observation was flagged by a particular metric is indicated by displaying the observation's valule on that metric in red. We see that the Mahalanobis D^2 did not flag a single one of the contaminant observations. As was discussed previously, Section 2.5.1, this is due to the masking effect exerted by the relatively high proportion, 0.20, of outliers. Since the Mahalanobis D^2 did not flag a single majority observation either, it had a false-alarm rate of zero. The Carrig D, PCHIGH, PCLOW and MCD were resistant to the masking effect. Each one of these metrics perfectly identified the contaminants, hit rate of 1. The Carrig D and PCHIGH did not sound a single false alarm. MCD and PCLOW each wrongly identified one of the majority observations, observation forty-four, as an outlier. Thus, the false-alarm rate for these two metrics was 0.02, acceptable. The performance of MLD is interesting. Its hit rate was 0.6, and its false-alarm rate was 0.38. The hit rate of 0.6 was surprising. There is no obvious reason to believe that the MLD is more resistant to the masking effect than the Mahalanobis D^2. The false-alarm rate of 0.38 is obviously unacceptable. I

decided to further investigate the MLD's resistance to masking effect by trimming the observations whose MLDs were in the top twenty percent of the set of observed values on the MLD. Of the twelve observations trimmed, five were contaminants. This means that the MLD applied with a twenty-percent trim rate yielded a hit rate of 0.50 and a still unacceptably high false-alarm rate of 0.14. The results of the trimming application indicate that the contaminants tended to have relatively high scores on the metric. However, a trimming application of any metric is not acceptable in practice: if there are no outliers, you wind up throwing out good data, and, furthermore, how do you determine what percentage to trim? Out of curiosity, I looked at a stem-and-leaf plot of the MLDs to see if any of the values stuck out on the end of plot. Table 5.2 contains the stem-and-leaf plot. None of the observations stick out with respect to this plot. Thus, a plot of the observed MLDs is of no utility in spotting the contaminants in this example. The stem-and-leaf plot also provides insight into why the way in which we used the MLD values led to such a high false-alarm rate. The plot shows that there are not two distinguishable groups of observations. Nevertheless, the k-means algorithm still divided the observations into two groups. The algorithm artificially defined a higher group with twenty-five observations, almost half of the total number of observations. Perhaps it is the case that when there are not two distinguishable groups the k-means algorithm creates two groups with a nearly equal number of observations in each.

I employed this same twenty-percent trimming application method with regard to the observed Mahalanobis D^2 values. None of the contaminants fell in the top 20% on the Mahalanobis D^2. This example suggests that the MLD is more resistant to the masking effect than the Mahalanobis D^2. I can think of no obvious reason for this. Recall that the Mahalanobis D^2 uses non robust estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and that MLD deviates an observation's maximum likelihood correlation estimate about the mean of all such estimates. I think it is a reasonable conjecture

3	|	9
4	|	
5	|	9
6	|	00111223445788
7	|	0023334456668999
8	|	0012222233445689
9	|	123349
10	|	045
11	|	89
12	|	
13	|	5

Table 5.2: MLDs for Fisher Iris Example

that the mean of all the maximum likelihood estimates would be strongly impacted by a relativel large proportion of contamiants in the sample.

Another interesting result is that both PCLOW and PCHIGH perfectly identified the contaminants. Recall that PCHIGH and PCLOW sum to the Mahalanobis D^2. Likewise, Robust PCLOW and Robust PCHIGH sum to the MCD. Also recall the point made that touting either PCLOW or PCHIGH implies that the other constitutes noise that affects the ability of the Mahalanobis D^2 or MCD to detect outliers. According to this example, neither PCLOW nor PCHIGH is noise. In attempting to explain this result, I first investigated the nature of the outliers. Were they likely purely mean-shift, purely correlational or both mean-shift and correlational? My general scheme for investigating this was to make inferences about the popuulation correlation matrices, $\boldsymbol{\rho}_{\text{setosa}}$ and $\boldsymbol{\rho}_{\text{versicolor}}$, and the population mean vectors, $\boldsymbol{\mu}_{\text{setosa}}$ and $\boldsymbol{\mu}_{\text{versicolor}}$ The sample mean vectors and sample correlation matrices for the two species are presented in Table 5.3.

I investigated the question of whether the population correlation matrices for the *setosa* species and the *versicolor* species are equal by using the Box's M statistic which is a part of the MANOVA package in SPSS v.15. The p-value of the statistic

$$\overline{\mathbf{y}}_{\text{setosa}} = \begin{bmatrix} 5.0 & 3.4 & 1.5 & 0.2 \end{bmatrix}'$$

$$\overline{\mathbf{y}}_{\text{versicolor}} = \begin{bmatrix} 5.9 & 2.8 & 4.3 & 1.3 \end{bmatrix}'$$

$$\mathbf{R}_{\text{setosa}} = \begin{bmatrix} 1 & 0.74 & 0.27 & 0.28 \\ 0.74 & 1 & 0.18 & 0.23 \\ 0.27 & 0.18 & 1 & 0.33 \\ 0.28 & 0.23 & 0.33 & 1 \end{bmatrix}$$

$$\mathbf{R}_{\text{versicolor}} = \begin{bmatrix} 1 & 0.53 & 0.75 & 0.55 \\ 0.53 & 1 & 0.56 & 0.66 \\ 0.75 & 0.56 & 1 & 0.79 \\ 0.55 & 0.66 & 0.79 & 1 \end{bmatrix}$$

Table 5.3: Sample Mean Vectors and Correlation Matrices for the Setosa and Versicolor Samples

was less than 0.0005, so I rejected the null hypothesis that the two population covariance matrices are equal [1] Thus, I inferred that the *versicolor* observations were correlational outliers at the very least. Next, I investigated the question of whether the *versicolor* observations are mean-shift outliers as well.

Even though it appears that the covariance matrices for the two species are not equal I performed a one-way MANOVA on the data to get an idea of whether the two population mean vectors are equal or not; MANOVA is somewhat robust to the violation of the assumption of equal covariance matrices when the data are balanced as they are in this case since I am using all fifty observations for each of the two species (Stevens, n.d.). The source table for the MANOVA is displayed in Table 5.4. The observed value of Wilks' Lambda is also displayed at the bottom of

[1] It should be mentioned that Box's M is sensitive to departures from multivariate normality. I inspected marginal probability plots of the four variables for each species, and there appeared to be some departures from marginal normality.

the table. That value led me to reject the null hypothesis that the two population mean vectors were equal. Combining this inference with one above concerning the population correlation matrices, I concluded that the ten *versicolor* contaminants in this example were both mean-shift and correlational outliers. The next step I took in seeking an explanation for the performance of PCLOW and PCHIGH in this example was to investigate the distance (arising from the mean shift) of the contaminant population from the majority population as well as the principal component structure of the majority population.

Source of variation	Matrix of sum of squares and cross-products	Degrees of freedom
Species	$\begin{bmatrix} 21.62 & -15.30 & 65.05 & 25.11 \\ -15.30 & 10.82 & -46.03 & -17.77 \\ 65.05 & -46.03 & 195.72 & 75.55 \\ 25.11 & -17.77 & 75.55 & 29.16 \end{bmatrix}$	1
Residual	$\begin{bmatrix} 19.14 & 9.04 & 9.76 & 3.24 \\ 9.04 & 11.87 & 4.62 & 2.47 \\ 9.76 & 4.62 & 12.30 & 3.88 \\ 3.24 & 2.47 & 3.88 & 2.46 \end{bmatrix}$	98
Total (corrected)	$\begin{bmatrix} 40.76 & -6.26 & 74.81 & 28.35 \\ -6.26 & 22.69 & -41.41 & -15.30 \\ 74.81 & -41.41 & 208.02 & 79.43 \\ 28.35 & -15.30 & 79.43 & 31.62 \end{bmatrix}$	99

df	Wilks' Lambda	approx F	df	Pr(> F)
1	0.03658	625.46	(4, 95)	< 0.05

Table 5.4: MANOVA source table and test statistic for subset of Fisher Iris data

In my investigation into the distance of the contaminant distribution from the majority distribution and the principal component structure of the majority

population, I discovered that standardized vectors (standardized with respect to the *setosa* sample) will have large projections into the plane spanned by the high principal components as well as large projections into the plane spanned by the low principal components (recall that the principal components are derived from a correlation matrix not a covariance matrix). Recall that it was pointed out that observations whose standardized vectors have large values on Rao's metric will tend to have large values on PCLOW. Also, recall that Rao's metric is simply the square of the length of the projection into the space spanned by the low principal components:

$$\text{Rao}'\text{s Metric} = \mathbf{x}_1'\mathbf{e}_{k+1}^*\mathbf{e}_{k+1}^{*\prime}\mathbf{x}_i + \ldots + \mathbf{x}_i'\mathbf{e}_p^*\mathbf{e}_p^{*\prime}\mathbf{x}_i. \tag{5.1}$$

and PCLOW is

$$\text{PCLOW} = \frac{\mathbf{x}_i'\mathbf{e}_{k+1}^*\mathbf{e}_{k+1}^{*\prime}\mathbf{x}_i}{\lambda_{k+1}^*} + \ldots + \frac{\mathbf{x}_i'\mathbf{e}_p^*\mathbf{e}_p^{*\prime}\mathbf{x}_i}{\lambda_p^*}. \tag{5.2}$$

where p is the number of variables and k is the number of high components. Again we see the similarity between the square of the length of the standardized vector's projection into the space spanned by the eigenvectors corresponding to the low components and PCLOW.

The same relationship exists between PCHIGH and the square of the length of the standardized vector's projection into the space spanned by the high components. Recall that

$$\text{High Proj} = \mathbf{x}_i'\mathbf{e}_1^*\mathbf{e}_1^{*\prime}\mathbf{x}_i + \ldots + \mathbf{x}_i'\mathbf{e}_k^*\mathbf{e}_k^{*\prime}\mathbf{x}_i \tag{5.3}$$

and

$$\text{PCHIGH} = \frac{\mathbf{x}_i'\mathbf{e}_1^*\mathbf{e}_1^{*\prime}\mathbf{x}_i}{\lambda_1^*} + \ldots + \frac{\mathbf{x}_i'\mathbf{e}_k^*\mathbf{e}_k^{*\prime}\mathbf{x}_i}{\lambda_k^*} \tag{5.4}$$

where k is the number of high components. Observations whose projections into the space spanned by the eigenvectors corresponding to the high components have large squared lengths so will tend to have large values on PCHIGH.

Remember that PCHIGH and PCLOW can be thought of as the statistical distances (centered about the projection of the mean vector) of the standardized observation's projection into the spaces spanned by the high components and low components, respectively.

Now let us go through the computations that demonstrate that the contaminants will have projections with large squared lengths in the both the plane spanned by the high components and the one spanned by the low components and thus large PCHIGH and PCLOW values. First, I wanted to get an idea of the distance between $\boldsymbol{\mu}_{\text{versicolor}}$ and $\boldsymbol{\mu}_{\text{setosa}}$. Since PCLOW and PCHIGH are statistical distances of a standardized observation's projection into the spaces spanned by components from a correlation matrix I compared the length of the standardized difference of the means vector to the standardized vectors for the *setosa* observations. The standardized difference of the sample means vector is

$$\mathbf{d}' = (\overline{\mathbf{x}}_{\text{versicolor}} - \overline{\mathbf{x}}_{\text{setosa}})'\mathbf{D}^{-1/2} \tag{5.5}$$

where

$$\mathbf{D}^{-1/2} = \begin{bmatrix} s_1^{-1} & 0 & 0 & 0 \\ 0 & s_2^{-1} & 0 & 0 \\ 0 & 0 & s_3^{-1} & 0 \\ 0 & 0 & 0 & s_4^{-1} \end{bmatrix} \tag{5.6}$$

and s_i is the standard deviation for variable i computed from the *setosa* sample. For the sake of computational ease, I will use the sample covariance and correlation matrices from the *setosa* samples rather than the ones output from the MCD algorithm. Since the proportion of contamination is nowhere near 0.50 the correlation

and covariance matrices output by the MCD algorithm and those from the *setosa* sample will be nearly equal. The length of the standardized difference of means vector was approximately 19.35. A histogram of this length along with the standardized lengths of the *setosa* observations follows in Figure 5.1. As you can see the length of the standardized difference of the mean vectors is much larger than the lengths of the standardized *setosa* observations. However, will the square of the length of the projection of the standardized difference of the means vector into the space spanned by the high principal components be large? We could easily answer this by computing the square of the length of the projection by using Equation 5.3 However, I wanted to understand why geometrically the projection is large or small. In order to answer this, I extracted the principal components from the *setosa* correlation matrix. Then I computed the angle between the standardized difference of the means vector and its projection into the space spanned by the high principal components. A relatively small angle will indicate that the length of the projection into the space will be almost as large as the vector while a relatively large angle will indicate that the length of the projection is a small fraction of the length of the original vector.

The *setosa* correlation matrix has two eigenvalues greater than one. They are 2.06 and 1.02. Summing these two and dividing by four, we see that the first two principal components account for approximately 77% of the variance of the standardized observations. The eigenvectors corresponding to these eigenvalues are $\mathbf{e}_1' = [0.604\ 0.576\ 0.375\ 0.403]$ and $\mathbf{e}_2' = [0.335\ 0.441\ -0.627\ -0.548]$. Thus, the space spanned by the high principal components is the plane spanned by the two vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ passing through the origin (the mean). The projection of the standardized difference of the means vector will be the sum of its projection onto $\mathbf{e}_1$ and $\mathbf{e}_2$. This is due to the orthogonality of $\mathbf{e}_1$ and $\mathbf{e}_2$. Mathematically, the

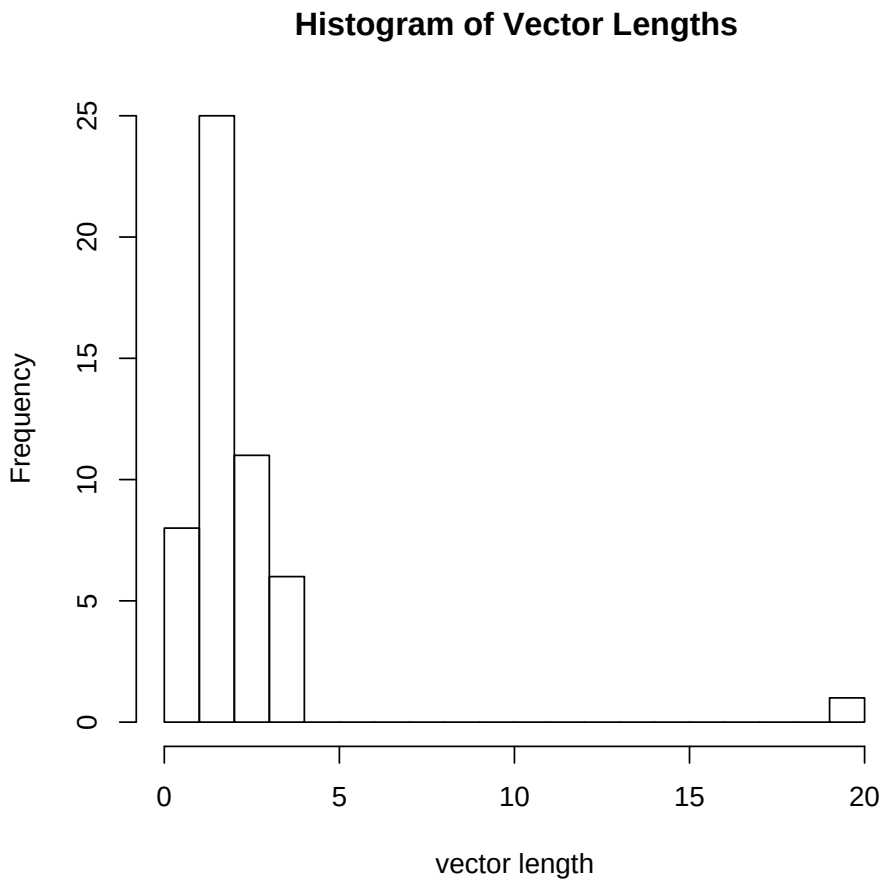

Figure 5.1: Standardized lengths of all of the *setosa* observations along with the standardized length of the difference of the *setosa* and *versicolor* mean vectors

projection is

$$\mathbf{p} = \mathbf{d}'\mathbf{e}_1\mathbf{e}_1 + \mathbf{d}'\mathbf{e}_2\mathbf{e}_2 \tag{5.7}$$

where $\mathbf{d}$ is the standardized difference of the means vector. I computed the projection, normalized it and took the inverse cosine of its inner product with the normalized standardized difference of the means vector. This computation yielded an angle of 11.61°. Thus, the length of the projection is $19.35\cos(11.61°) = 18.95$, very large in comparison with the lengths of the projections of the standardized *setosa* observations themselves. The angle between the vector $\mathbf{d}$ and the plane spanned by the high principal components is small. It indicates that the direction of the shift from the *setosa* to the *versicolor* mean was almost parallel to the plane spanned by the high components. I've established that the projection of the standardized difference of the means vector has a large projection into the plane spanned by the high components because the direction of the standardized mean shift was nearly coincident with the plane, but I have not established that individual *versicolor* observations will have long projections into the plane spanned by the high components.

As can be obviously seen from the results in Table 5.5, the *versicolor* contaminants had large PCHIGH values, so the squared length of the projection of their standardized vectors was large. Why? I suspected that the large distance of the mean shift from *setosa* to *versicolor* overshadowed the variation of the *versicolor* observations thus making the length and angle of the standardized vectors very close to those for the standardized difference of the means vector, $\mathbf{d}$. If this were indeed the case, this would establish why, geometrically, the *versicolor* contaminants had large PCHIGH values. The *versicolor* contaminants will have large PCHIGH for the very same reason that $\mathbf{d}$ had a large PCHIGH value: their displacement from the *setosa* mean is along a line that is almost coincident with the plane that is the span of the high principal components. I established that the lengths of standardized *versicolor* vectors were close to the length of $\mathbf{d}$ by comparing the length of the vector

$\overline{\mathbf{x}}_{\text{versicolor}} - \overline{\mathbf{x}}_{\text{setosa}}$ with the lengths of the fifty *versicolor* vectors that were deviated about the *versicolor* mean. Note that an individual *versicolor's* deviation vector about the *setosa* mean, $(\mathbf{x} - \overline{\mathbf{x}}_{\text{setosa}})$, is equal to the sum of $\overline{\mathbf{x}}_{\text{versicolor}} - \overline{\mathbf{x}}_{\text{setosa}}$ and $\mathbf{x} - \overline{\mathbf{x}}_{\text{versicolor}}$, so if the length of $\mathbf{x} - \overline{\mathbf{x}}_{\text{versicolor}}$ is small relative to $\overline{\mathbf{x}}_{\text{versicolor}} - \overline{\mathbf{x}}_{\text{setosa}}$, then the lengths of *versicolor* observations deviated about the *setosa* mean will be close to the length of $\overline{\mathbf{x}}_{\text{versicolor}} - \overline{\mathbf{x}}_{\text{setosa}}$. The orientations of the these vectors should also be close to that of the vector $\overline{\mathbf{x}}_{\text{versicolor}} - \overline{\mathbf{x}}_{\text{setosa}}$. These two facts imply that standardized difference vectors $(\mathbf{x} - \overline{\mathbf{x}}_{\text{setosa}})\mathbf{D}^{-1/2}$ will be very similar in length and orientation to $\mathbf{d}$. Thus, observations from the *versicolor* distribution will tend to have deviation vectors with close to the same angle with the plane spanned by the high components as the difference vector of the means, and they will tend to have large values on PCHIGH for the same geometric reason as was the case for the vector $\mathbf{d}$.

Now, I will demonstrate that the sample data back up the contention that the length and orientation of standardized *versicolor* observations will tend to be close as those for the vector $\mathbf{d}$ Refer to figure 5.2. This is a histogram of the lengths of the fifty *versicolor* difference vectors, $\mathbf{x} - \overline{\mathbf{x}}_{\text{versicolor}}$ along with the length of the difference of the means vector $\overline{\mathbf{x}}_{\text{versicolor}} - \overline{\mathbf{x}}_{\text{setosa}}$. The histogram provides evidence that the length of $\overline{\mathbf{x}}_{\text{versicolor}} - \overline{\mathbf{x}}_{\text{setosa}}$ is large relative to the lengths of the deviated *versicolor* observation vectors, $\mathbf{x} - \overline{\mathbf{x}}_{\text{setosa}}$. This implies that the lengths of the standardized *versicolor* vectors, $(\mathbf{x} - \overline{\mathbf{x}}_{\text{setosa}})'\mathbf{D}^{-1/2}$ are close to that for $\mathbf{d}$. Refer to Figure 5.3. This is a histogram of the angles between the standardized *versicolor* observations and the plane spanned by the high components from the *setosa* correlation matrix. We see that most of the angles are not far away from the value of 11.61°, the angle between $\mathbf{d}$ and the plane spanned by the high components from $\mathbf{R}_{\text{setosa}}$.

In summary, we have established that the *versicolor* contaminants were detected by PCHIGH primarily because the large mean shift from the *setosa* distri-

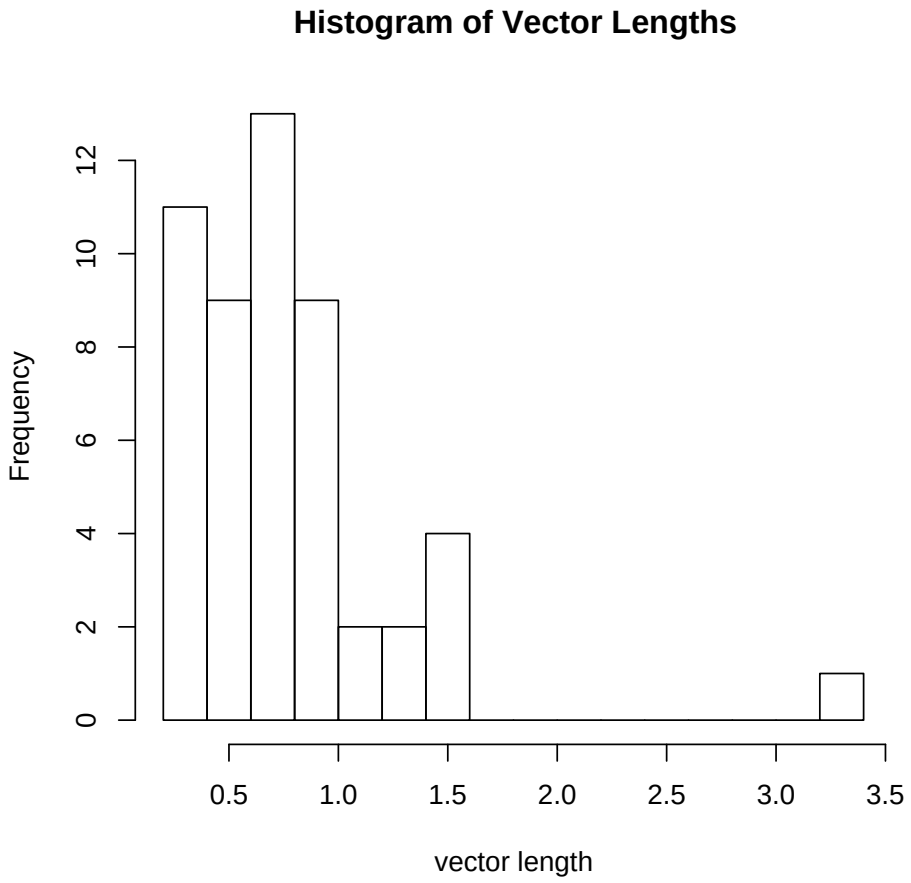

Figure 5.2: Histogram of the lengths of the fifty versicolor difference vectors, $\mathbf{x} - \overline{\mathbf{x}}_{\text{versicolor}}$, along with the length of the vector $\overline{\mathbf{x}}_{\text{versicolor}} - \overline{\mathbf{x}}_{\text{setosa}}$ vector

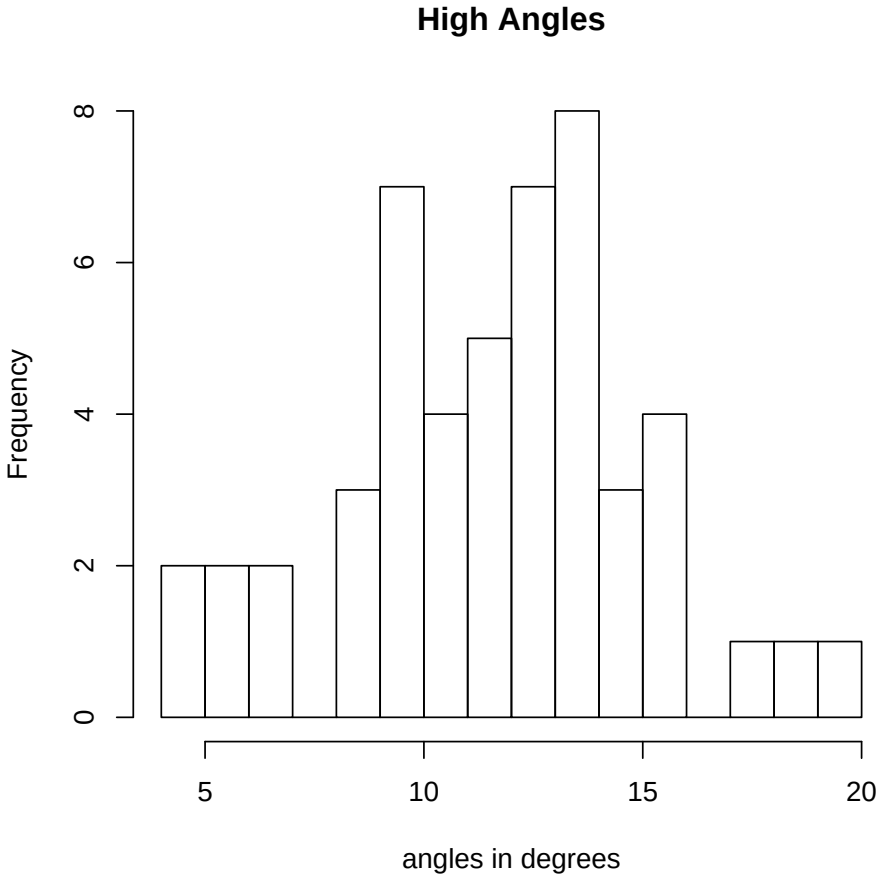

Figure 5.3: Histogram of the angles between the standardized *versicolor* observations, $(\mathbf{x} - \overline{\mathbf{x}}_{\text{setosa}})'\mathbf{D}^{-1/2}$ and the plane spanned by the high components of $\mathbf{R}_{\text{setosa}}$

bution to the *versicolor* distribution was in a direction that was nearly coincident with the plane spanned by the high components. In other words, the mean shift was largely in the direction of the high components. I believe that this finding has broad implications for the general topic of multivariate outlier detection as will be discussed later. Now it is time to consider why the *versicolor* contaminants were detected by PCLOW as well.

The geometric reasons why the *versicolor* contaminants were detected by PCLOW can be discerned from the properties of the standardized difference of the means vector $\mathbf{d}' = (\overline{\mathbf{x}}_{\text{versicolor}} - \overline{\mathbf{x}}_{\text{setosa}})'\mathbf{D}^{-1/2}$ and its relation to the space spanned by the low components just as in the case with PCHIGH. Again this is due to the fact that the mean-shift is so large that it renders the fact that the correlation matrices, $\mathbf{R}_{\text{setosa}}$ and $\mathbf{R}_{\text{versicolor}}$, are different as insignificant. Recall that the extraction of principal components from the correlation matrix, $\mathbf{R}_{\text{setosa}}$ yielded two variances greater than one and two less than one, so the square of the length of the projection of the standardized *versicolor* observations into the space spanned by the low components, Rao's Metric (Equation 5.1), depends on the two eigenvectors (principal components) associated with the eigenvalues (variances) that are less than one. Since there are two low components, the space spanned by them is a plane just as the space spanned by the high components. Because of the orthogonality between the plane spanned by the high components and the one spanned by the low components, the angle between the vector $\mathbf{d}$ and the plane spanned by the low components is ninety degrees minus the angle between $\mathbf{d}$ and the plane spanned by the high components. The value of that angle is $90° - 11.61° = 78.39°$. I verified this relationship by calculating the angle through the inverse cosine of the inner product between the standardized vector of $\mathbf{d}$ and the standardized vector of its projection into the plane spanned by the low components. This computation yielded the same angle. I previously demonstrated that the angles between the

standardized *verisicolor* vectors and the plane spanned by the high components will be in general close to the value for **d**, 11.61°. An implication of this is the fact that the angle between standardized *versicolor* vectors and the the plane spanned by the low components will be close to the value of 78.39°. The foregoing demonstrates that the behavior of the *versicolor* observations with respect to PCLOW can be discerned by the length of **d** and its angle with the plane spanned by the low compnents. Therefore, the length of the projection of the standardized *versicolor* obervations will approximately be the cos(78.39°) times the length of the standardized observation. Even though the value of the cosine just approximately 0.201 the sheer magnitude of the standardized *versicolor* observations ensures that the length of their projections into the plane spanned by the low components will be large. Add to this the fact that the squared projections onto each of the two eigenvectors, $\mathbf{e}_3$ and $\mathbf{e}_4$ are divided by eigenvalues (variances) λ_3 and λ_4 (refer to Equation 5.2) and we have an accounting of why *versicolor* observations have large PCLOW values. I agree with Jolliffe that the fact that an observation has a large PCLOW value indicates that it is outlying with respect to the correlational structure of the majority of the data (Jolliffe, 1986). Thus, the *versicolor* observations were outlying with respect to the correlation structure of the *setosa* observations, and this I believe is the reason why the Carrig D detected the *versicolor* observations as well. Like the geometric explanation for large PCHIGH values, the explanation for large PCLOW values led me to some heretofore unstated observations about the nature of multivariate outliers. A big picture conclusion here is that a mean-shift that moves a large distance in the space spanned by the high principal components observations from the outlying distribution will tend to have large PCHIGH values. Thus, PCHIGH will be effective in identifying these outliers.

Table 5.5: Metric Values for Fisher Data

D^2	MCD	Carrig D	PCHIGH	PCLOW	MLD	Species
0.43	0.40	6.47	0.00	0.39	0.64	*setosa*
2.30	1.99	5.80	0.68	2.00	0.85	*setosa*
0.77	1.18	6.90	0.91	0.42	0.79	*setosa*
1.67	1.68	3.64	0.87	0.83	0.81	*setosa*
0.65	0.69	6.07	0.00	0.75	0.72	*setosa*
2.23	6.23	5.98	3.44	1.44	0.62	*setosa*
2.49	4.17	6.08	0.12	3.73	0.61	*setosa*
0.53	0.30	6.86	0.02	0.26	0.63	*setosa*
2.66	2.84	3.33	2.27	1.03	0.84	*setosa*
3.27	3.54	4.23	0.91	2.61	0.82	*setosa*
1.60	1.82	4.65	0.42	1.45	0.93	*setosa*
2.06	2.17	3.99	0.06	1.81	0.64	*setosa*
3.01	3.39	3.85	1.71	1.66	0.91	*setosa*
2.69	7.39	5.96	5.06	1.99	0.92	*setosa*
6.79	10.75	27.80	1.14	8.86	1.04	*setosa*
8.25	8.51	9.05	6.29	2.10	0.70	*setosa*
6.06	6.57	14.01	1.84	4.76	0.73	*setosa*
0.87	0.91	5.10	0.11	0.92	0.64	*setosa*
2.81	5.28	5.41	2.95	2.06	0.89	*setosa*
1.81	2.13	6.00	0.62	1.21	0.60	*setosa*
3.01	5.16	6.86	0.32	4.82	0.84	*setosa*
2.51	4.93	6.76	1.08	3.20	0.59	*setosa*
3.35	10.76	15.34	0.99	9.56	0.80	*setosa*
2.47	13.51	21.65	1.46	11.82	1.00	*setosa*

Continued on next page

D^2	MCD	Carrig D	PCHIGH	PCLOW	MLD	Species
6.99	10.80	24.00	0.02	9.36	0.60	*setosa*
2.75	3.59	6.09	0.22	3.94	0.79	*setosa*
0.98	5.39	8.06	0.51	4.41	0.61	*setosa*
0.81	0.76	5.98	0.04	0.72	0.82	*setosa*
1.22	1.30	5.60	0.00	1.32	0.78	*setosa*
1.96	2.30	4.69	0.34	1.69	0.75	*setosa*
1.59	2.00	4.57	0.34	1.69	0.82	*setosa*
4.30	6.32	8.52	1.09	6.30	0.82	*setosa*
7.45	8.41	18.21	0.26	10.53	0.67	*setosa*
4.39	5.15	13.87	1.44	4.83	0.74	*setosa*
1.30	1.16	4.52	0.35	1.15	0.83	*setosa*
2.48	3.52	4.20	0.55	3.05	0.76	*setosa*
4.58	5.97	9.14	0.08	5.54	0.99	*setosa*
2.51	3.63	5.43	0.27	3.81	0.76	*setosa*
2.29	3.06	4.83	2.31	1.10	0.83	*setosa*
0.65	0.52	0.54	0.00	0.54	0.61	*setosa*
1.51	1.88	4.67	0.01	2.04	0.68	*setosa*
12.70	12.87	28.35	3.34	15.32	0.79	*setosa*
2.58	3.95	5.84	1.68	2.30	0.68	*setosa*
8.22	22.73	46.07	2.43	19.17	1.18	*setosa*
3.81	13.42	7.72	2.73	7.35	0.39	*setosa*
2.27	2.71	6.20	0.33	3.72	0.82	*setosa*
3.14	2.89	4.39	0.30	2.84	0.62	*setosa*
1.26	1.38	5.85	0.89	0.52	0.76	*setosa*

Continued on next page

D^2	MCD	Carrig D	PCHIGH	PCLOW	MLD	Species
1.22	1.20	4.44	0.29	1.00	0.86	*setosa*
0.61	0.44	7.29	0.14	0.40	0.70	*setosa*
10.61	568.12	878.82	105.81	390.88	0.73	*versicolor*
6.34	551.99	900.03	94.50	379.34	0.74	*versicolor*
9.38	650.07	1045.23	113.97	451.85	0.93	*versicolor*
7.36	423.58	887.03	40.12	338.20	0.88	*versicolor*
7.39	589.92	1022.99	90.45	429.81	0.88	*versicolor*
9.37	508.91	969.91	60.01	371.29	0.80	*versicolor*
9.42	631.35	1030.36	106.81	427.66	0.73	*versicolor*
6.70	232.52	535.93	14.40	191.04	1.05	*versicolor*
9.23	524.98	897.96	81.09	378.81	1.19	*versicolor*
12.32	431.37	867.17	44.67	326.68	0.94	*versicolor*

5.2 Observations from the Iris Example

My geometric interpretation of the results from this example led me to a conclusion that is not in the literature. I think that a distinction should be made between three kinds of purely mean-shift outliers (by purely mean-shift I mean outliers that arise from a distribution with a different mean vector from the majority population but having the same correlation matrix as the majority population). There are mean shift outliers in which the mean-shift moves a large distance in the space spanned by the high principal components but moves a small distance in the space spanned by the low principal components. Outliers of this nature will be flagged by PCHIGH but not by PCLOW. It is possible that MCD may not spot this kind of outlier because the PCLOW part of the sum will be noise. I call this type of outlier "high-slide-only–mean-shift outliers. "Another kind of mean shift occurs when the mean moves a substantial distance in both the space spanned by the high components and

the space spanned by the low components. I call these "high-and-low-slide–mean-shift outliers. "The contaminants in the Iris example were essentially of this variety. This type of outlier will be detected by both PCHIGH and PCLOW. Since it will be identified by both there will be no noise in the MCD. Thus, MCD will spot this kind of outlier as well. The key thing to take away from this is that the "high-and-low-slide–mean-shift outliers" while coming from a distribution with the same correlation matrix as the majority will yet produce observations that are outlying with respect to the correlational structure of the majority distribution. The fact that some purely–mean-shift outliers can be outlying with respect to the correlation structure of the majority population has not been explicitly stated in the literature. Finally there is the "low-slide-only–mean-shift outlier." In this case the mean shift is substantial in the space spanned by the low components but insubstantial in the space spanned by the high components. This type will be spotted by PCLOW only and again a mean-shift produces observations that are outlying correlationally with respect to the majority population. Thus, I contend that the distinction between purely–mean-shift outliers and purely-correlational outliers does not provide insight into the problem of multivariate outliers since purely–mean-shift outliers can manifest themselves in a way that has been previously attributed to purely-correlational outliers only.

I should mention that I arrived at these conclusions by thinking about a sample cloud of data for each population both plotted together in the same coordinate system or equivalently constant-density ellipsoids. Using this kind of visualization, one can imagine scenarios for purely correlational outliers in which PCHIGH only will spot the contaminants, others in which PCLOW only will flag the contaminants, and also others in which both PCLOW and PCHIGH will both detect the contaminants (in this case MCD would detect the contaminants also). The significance of these conjectures is that the forgotten part of the literature on multivariate outliers,

$$
\begin{aligned}
y_1 &= \text{Reading Speed} \\
y_2 &= \text{Reading Power} \\
y_3 &= \text{Arithmetic Speed} \\
y_4 &= \text{Arithmetic Power}
\end{aligned}
$$

$$
\mathbf{R} = \begin{bmatrix}
1 & 0.701 & 0.266 & 0.084 \\
0.701 & 1 & -0.059 & 0.092 \\
0.266 & -0.059 & 1 & 0.596 \\
0.084 & 0.092 & 0.596 & 1
\end{bmatrix} \tag{5.8}
$$

Table 5.6: Correlation Matrix from Hotelling 1933.

the relevancy of principal compoents analysis in multivariate outlier detection, is important but not in the ways previously thought. They cause us to rethink the notions that the Mahalanobis D^2 or its robust version the MCD are only useful in spotting mean-shift outliers (Bacon, 1995) or that PCHIGH is of no utility, or noise in the detection of correlational outliers (Jolliffe, 1986) or the importance of dividing outliers into the categories of mean-shift and correlational (Gnanadesikan & Kettenring, 1972; Comrey, 1985; Bacon, 1995).

5.3 Hotelling Correlation Matrix

It is conceivable that we could run into a high-slide-only–mean-shift outlier in the field of psychometrics. Consider the famous correlation matrix from Hotelling in Table 5.3 (McDonald, 1985):

I performed a principal components analysis of this correlation matrix and obtained $\lambda_1 = 1.85$ and $\lambda_2 = 1.46$ as the variances of the first two principal compoonents. The coefficients of the corresponding principal components are contained

in the two eigenvectors:

$$\mathbf{e}_1' = [0.598\ 0.506\ 0.450\ 0.428] \tag{5.9}$$

$$\mathbf{e}_2' = [0.366\ 0.517\ -0.553\ -0.542] \tag{5.10}$$

It appears that the first principal component represents general intelligence and that the second is a reading versus math variable

These were the only principal components that had variances greater than one. Summing the two variances and dividing by four, we see that the two components account for approximately 82% of the variance. Suppose the correlation matrix came from a large block of post-elementary students who had scored in the average range on a IQ test administered upon entry into the first grade. Now suppose that a certain percentage of these students, say five to ten percent, became regular users of inhalant drugs after the first grade. It is well documented that inhalants such as airplane glue and gasoline cause irreversible neurological damage. It is conceivable that the multivariate distribution for the inhalant abusers has a different mean vector but the same correlation matrix. It could be that the shift is substantial on the first principal component only; in other words, affecting just general intelligence only. Geometrically, the mean shifted a large distance in the space spanned by the first two principal components. PCHIGH would flag the contaminants, but PCLOW would not. The point is that PCHIGH may offer something that PCLOW does not in possible real-world scenarios.

Chapter 6

Results

6.1 Criterion for Significance

Because of the tremendous power in my study's design (14256 degrees of freedom in the between-samples error term and 71280 degrees of freedom in the within-samples error term), I decided to employ a measure of practical significance to determine which effects would make a difference in the real-world application of the metrics. I used the strength of association measure ω^2 introduced by Hays. (Kirk, 1995). The measure indicates the proportion of population variance in the dependent variable that is accounted for by the effect in question. Mathematically,

$$\omega^2 = \frac{\sigma_\alpha^2}{\sigma_\epsilon^2 + \sigma_\alpha^2} \tag{6.1}$$

where σ_α^2 is the variance of the treatment effects and σ_ϵ^2 is the error variance. I used the observed mean squares to get estimates of σ_α^2 and σ_ϵ^2. I used the following guidelines for assessing strength of association published by Cohen (Kirk, 1995):

$$\begin{aligned} \omega^2 &= 0.010 \text{ is a small association} \\ \omega^2 &= 0.059 \text{ is a medium association} \end{aligned}$$

$$\omega^2 \quad = \quad 0.138 \text{ or larger is a large association}$$

I performed two separate analysis of variances, one for the dependent variable hit rate and one for the dependent variable false-alarm rate. The ANOVA for hit rate is discussed first

6.2 Hit Rate ANOVA

The source table for the hit rate is in Table 6.2. Since the experiment had a split-plot design the table is divided into the two sections: between-samples variance and within-samples variance. Recall that samples is the only random factor and it is nested within the between-samples cells. Metric, factor H, along with all interactions involving H are the within-samples effects.

6.2.1 Between-Samples Effects on Hit Rate

Looking at Table 6.2, we see that there are main effects for communality, scenario, number of variables and sample size. However, each of these is a part of a practically significant interaction effect involving metric. Thus, they will not be discussed because their effects are qualified by other factors. Their qualified effects will be discussed in the within-samples effects section (Section 6.2.2) The communality-scenario and communality—number-of-variables interaction effects are practically significant; however, each of these effects is qualified by metric. Their qualified effects will be discussed in the within-samples effects section (Section 6.2.2).

Table 6.2: Source Table for the Hit Rate

A Communality

B Scenario

C Number of Variables

D Sample Size

F Fraction of Outliers

G Sample or Replicate

H Metric

Table 6.1: Factors

Source	df	SS	MS	ω^2
Between Samples	(14,399)	(1187.976)		
A	3	476.220	158.740	0.066
B	1	150.710	150.710	0.042
C	1	73.507	73.507	0.021
D	2	48.722	24.361	0.007
F	2	2.217	14.608	0.004
AB	3	116.541	38.847	0.016
AC	3	55.393	18.464	0.008
AD	6	7.156	1.193	0.001
AF	6	15.060	2.510	0.002
BC	1	1.459	1.459	< 0.0005
BD	2	0.064	0.032	< 0.0005
BF	2	8.703	4.352	0.001
CD	2	14.419	7.209	0.002

Continued on next page

Source	df	SS	MS	ω^2
CF	2	0.720	0.360	< 0.0005
DF	4	0.038	0.009	< 0.0005
ABC	3	0.843	0.281	< 0.0005
ABD	6	0.712	0.119	< 0.0005
ABF	6	6.990	1.165	0.001
ACD	6	1.832	0.305	< 0.0005
ACF	6	0.277	0.046	< 0.0005
ADF	12	0.100	0.008	< 0.0005
BCD	2	0.218	0.109	< 0.0005
BCF	2	0.005	0.002	< 0.0005
BDF	4	0.082	0.020	< 0.0005
CDF	4	0.027	0.007	< 0.0005
ABCD	6	0.636	0.106	< 0.0005
ABCF	6	0.413	0.069	< 0.0005
ABDF	12	0.332	0.028	< 0.0005
ACDF	12	0.286	0.024	< 0.0005
BCDF	4	0.070	0.017	< 0.0005
ABCDF	12	0.104	0.009	< 0.0005
G(A B C D F)	14256	177.120	0.012	0.021
Within Samples	(72000)	(2980.444)		
H	5	1770.087	354.017	0.495
AH	15	247.140	16.476	0.069
BH	5	78.605	15.721	0.022
CH	5	87.609	17.522	0.025

Continued on next page

Source	df	SS	MS	ω^2
DH	10	103.291	10.329	0.029
FH	10	15.226	1.523	0.004
ABH	15	67.514	4.501	0.016
ACH	15	59.181	3.945	0.019
ADH	30	17.255	0.575	0.005
AFH	30	8.499	0.283	0.002
BCH	5	1.083	0.217	< 0.005
BDH	10	0.674	0.067	< 0.0005
BFH	10	6.057	0.606	0.002
CDH	10	30.404	3.040	0.008
CFH	10	0.784	0.078	< 0.0005
DFH	20	0.743	0.037	< 0.0005
ABCH	15	3.109	0.207	0.002
ABDH	30	1.611	0.054	< 0.0005
ABFH	30	8.866	0.296	0.002
ACDH	30	4.540	0.151	0.001
ACFH	30	1.017	0.034	< 0.0005
ADFH	60	0.868	0.014	< 0.0005
BCDH	10	0.239	0.024	< 0.0005
BCFH	10	0.114	0.011	< 0.0005
BDFH	20	0.191	0.010	< 0.0005
CDFH	20	0.160	0.008	< 0.0005
ABCDH	30	1.396	0.047	< 0.0005
ABCFH	30	0.983	0.033	< 0.0005

Continued on next page

Source	df	SS	MS	ω^2
ABDFH	60	0.627	0.010	< 0.00005
ACDFH	60	0.425	0.007	< 0.0005
BCDFH	20	0.136	0.007	< 0.0005
ABCDFH	60	0.314	0.005	< 0.0005
Error	71280	4168.420		
Total	86399	4168.420		

6.2.2 Within-Samples Effects on Hit Rate

Communality-Metric Interaction Effect on Hit Rate

Notice that the communality-metric (AH) interaction qualifies as a medium association under the criteria used with ω^2. The two three way interactions involving AH, communality-scenario-metric (ABH) and communality–number-of-variables–metric interaction (ACH), only qualify as low associations. Therefore, I thought it would be instructive to look at the communality-metric interaction plot even though this effect is somewhat qualified by both scenario and number of variables. Refer to Figure 6.1. We see the same general trend for the four metrics PCLOW, MCD, Carrig D and the MLD. The communality-metric interaction effect seems to be mainly due to the fact that the Mahalanobis D^2 and PCHIGH do not have the same trend lines as the other four. Another indication of an interaction effect is the steeper downward slope in the first part of the curves for PCLOW and MCD and the greater upward slope at the end of the Carrig D curve. PCHIGH and the Mahalanobis D^2 perform poorly in all of the communality conditions. Henceforth, I shall limit the discussion of the interaction effect to the other four metrics.

All four of the metrics perform best in the condition where the communality in the majority population is higher than that for the contaminant population. It is interesting to note that the mean hit rate for PCLOW is equal to the mean hit rate

for the Carrig D in the higher condition. As can be seen from the plot, the MLD has the highest hit rate in all of the communality conditions. All of the metrics perform poorest in the condition where the communality of the majority population is lower than that for the contaminants. It is especially interesting to note that PCLOW and MCD plummet to a mean hit rate of approximately two percent in the lower condition. While the mean hit rate for PCLOW and the Carrig D were equal in the higher condition, the mean hit rate for the Carrig D in the lower condition is twelve percent compared to the two percent for PCLOW. The Carrig D is more resistant to the negative effect of the lower communality condition. While the difference between the higher and lower communality conditions is quite pronounced, the difference between hit rates for the high-same and low-same conditions appears to be not very great. The effect of going from high same to low same is nearly identical for three of the metrics: MLD, MCD and PCLOW. The hit rate drops from somewhere between 0.05 and 0.07 for each of these metrics. Interestingly, the Carrig D's hit rate increased by 0.05 in going from high same to low same, from 0.19 to 0.24. It is also interesting to note that the Mahalanobis D^2 outperforms the PCHIGH in the higher communality condtion. PCHIGH is a robust metric whereas the Mahalanobis D^2 is not. However this difference is not practically significant. They both perform very poorly in this condition. Finally, in regard to the communality-metric interaction there is the interesting behavior of PCHIGH. The hit rate for PCHIGH actually increases in going from the higher communality condition to the lower communality condition. Its hit rate in the former condition is 0.0005 and its hit rate in the latter is 0.02. This effect is in the reverse direction of the effect for MCD, PCLOW, Carrig D and MLD. Interesting, but it is of no practical consequence; a hit rate of 0.02 in the lower communality condition is still miserable.

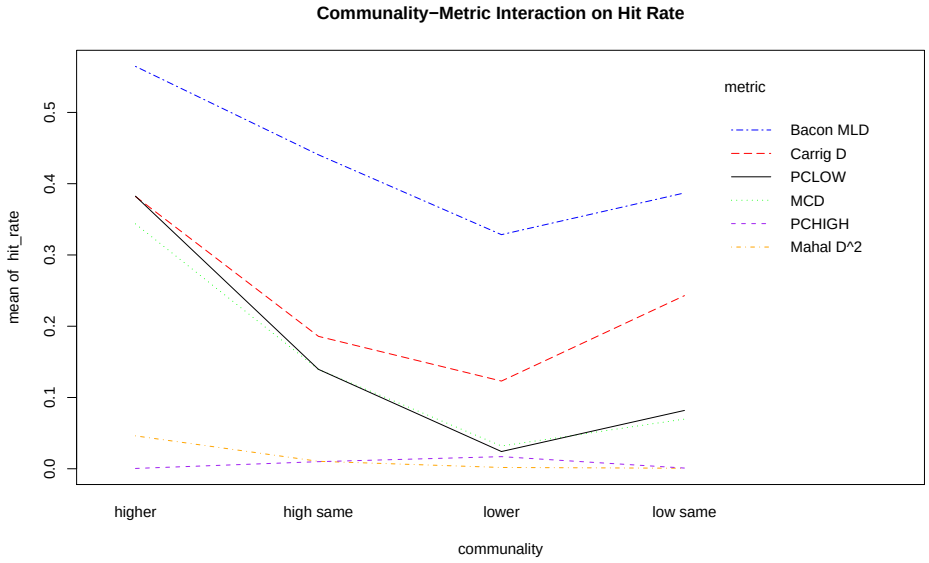

Figure 6.1: Communality-Metric Interaction on Hit Rate

Communality-Scenario-Metric Interaction Effect on Hit Rate

The communality-scenario-metric interaction effect is depicted in Figures 6.2 and 6.3. The performance of PCHIGH and the Mahalanobis D^2 is consistently abysmal across all of the communality-scenario combinations. Therefore it will be excluded from the discussion of the comunality-scenario-metric interaction effect.

Fundamentally, the observed communality-scenario-metric interaction can be characterized by two statements: there is no observed communality-metric interaction under scenario I, there is an observed communality-metric interaction under scenario II. The plot of mean hit rate versus communality level for each metric is depicted in Figure 6.2.

The effect of communality is essentially constant across all levels of metric. The effects for the different levels of communality averaged across the four metrics are high same, 0.02; higher, 0.25; low same, -0.09; lower, -0.17.

Plot 6.3 depicts an observed communality-metric interaction under scenairo II. The mean hit rate versus communality curves for the Carrig D and the MLD changed shape in going from scenario I (one-factor model for legitimate population and two-factor model for contaminant) to scenario II (four-factor model for legitimate population and two-factor model for contaminant) while the shape of the mean hit rate versus commuanality for the MCD and PCLOW remained essentially the same.

The curves of mean hit rate versus communality are essentially parallel are the same shape for the Carrig D and the MLD. Therefore, the effects of communality on the mean rates of these metrics is essentially the same. The effects averaged over both the MLD and Carrig D are high same, -0.05; higher, 0.05; low same, 0.04; lower, -0.03. The effect for high same is in the reverse direction for scenario II compared to scenario I; the same can be said of low same. The higher effect, while in the same direction, is not nearly as great; the same can be said for the lower effect. The magnitude or absolute value of the communality effects is smaller under scenario II than under scenario I.

The Communality–Number-of-Variables–Metric Interaction

Figures 6.4 and 6.5 depict the Communality–Number-of-Variables–Metric Interaction. Specifically Figure 6.4 depicts the communality-metric interaction under the condition that the number of variables is eight. We see that there would not really be an interaction between communality and metric under number of variables equal to eight if it were not for PCHIGH and the Mahalanobis D^2. As in the three-way interaction discussed in the preceding section, the performance of PCHIGH and the Mahalanobis D^2 is so consistently poor (their trend lines are flattened out) among all of the three-way combinations that it will not enter into the discussion that follows. The curves for the four metrics, MLD, Carrig D, PCLOWand MCD are nearly

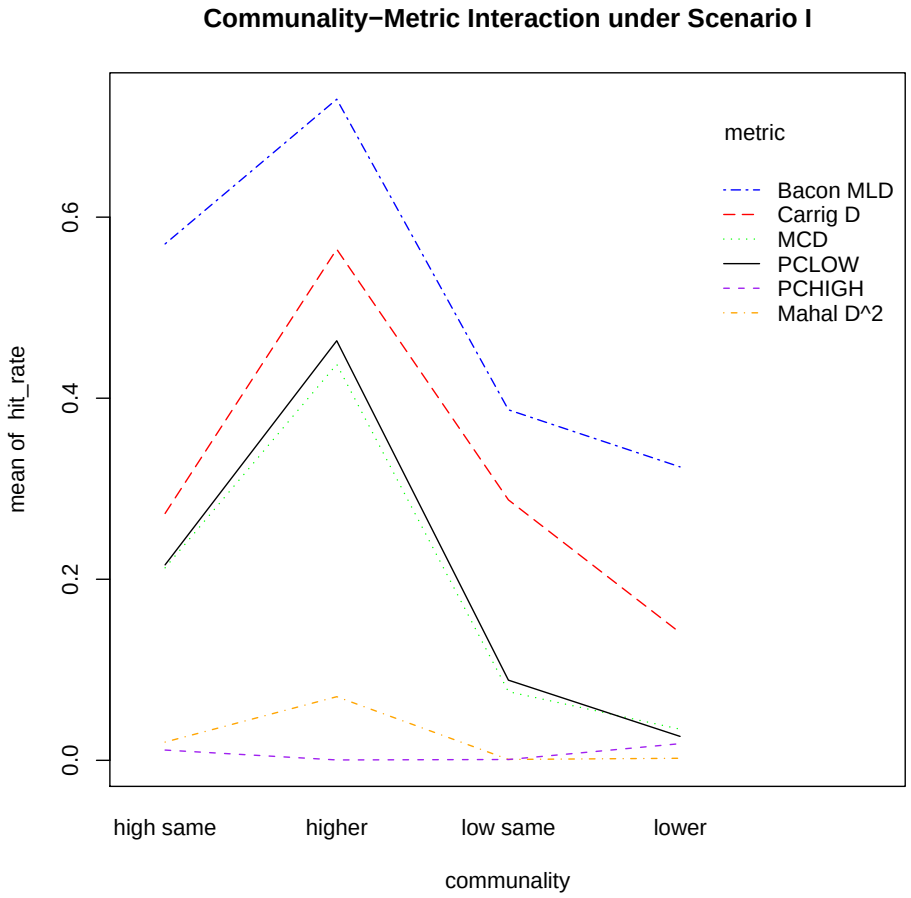

Figure 6.2: Communality-Metric Interaction on Hit Rate under Scenario I, n=1800

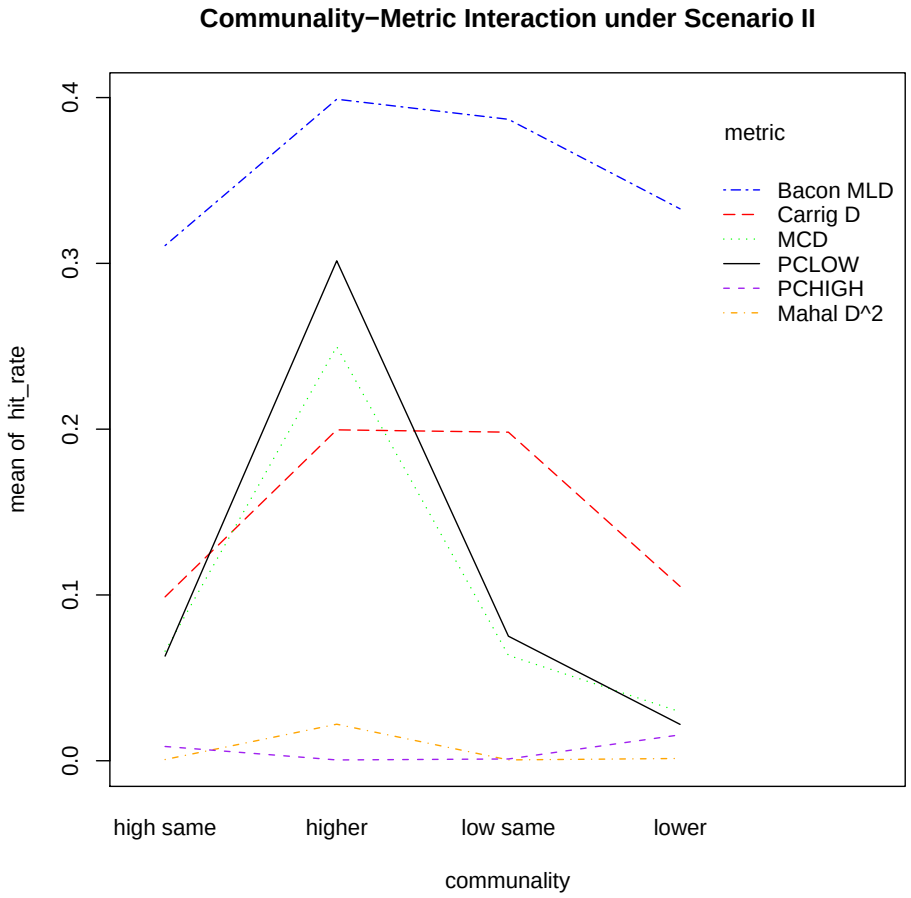

Figure 6.3: Communality-Metric Interaction on Hit Rate under Scenario II, n=1800

parallel. This indicates that there is a pretty constant metric effect across the levels of communality when the number of variables equals eight. The curves for PCLOW and MCD are not perfectly coincident, however. This is due to the slightly higher mean hit rate for PCLOW under the higher communality condition. This mean is approximately seventeen percent for MCD and is approximately twenty-one percent for PCLOW. In contrast to the number of variables equals eight condition their does seem to be a communality-metric interaction when the number of variables equals sixteen.

Under $p = 16$, the metric effects are not constant across the levels of communality. Compare, for example, the difference in mean hit rates for MLD and PCLOW when the communality is higher to the difference when the communality is lower. When the communality is higher, the mean hit rate for PCLOW is approximately fifty-six percent. The corresponding mean hit rate for the MLD is fifty-eight percent. In this instance, the effect size for each of the metrics is essentially the same. Now, when the communality is lower, the mean hit rate for PCLOW is roughly four percent. The mean hit rate for the MLD is thirty-three percent. The effects for MLD and PCLOW are quite different in this case. Therefore, the effect for metric is qualified by the communality level when $p = 16$ unlike the case when $p = 8$. However, the difference in effect sizes for the Carrig D and the MLD are pretty consistent across levels of communality when $p = 16$. Their curves in the plot are nearly parallel. It is the shapes of the PCLOW and MCD curves that change radically from the $p = 8$ to the $p = 16$ plot. This is the main reason why there is a three-way interaction for communality, sample size and metric. PCLOW and MCD clearly perform at their best when the communality is higher and $p = 16$. The higher communality effect on PCLOW and MCD is much more pronounced when $p = 16$. Increasing the number of variables to sixteen improves the performance of PCLOW and MCD at all levels of communality with the exception of the lower

condition. Even when $p = 16$ the mean hit rate is just three percent.

Number-of-Variables–Sample-Size–Metric Interaction

I chose to look at this interaction in terms of number of variables qualifying the sample-size–metric interaction. The sample-size–metric interaction when the number of variables equals eight is discussed first. Refer to Figure 6.6. The figure shows that there is a two-way sample-size–metric interaction present when the number of variables is equal to eight. The trend lines for PCLOW and MCD are different from the ones for the other four metrics. The lines for the other four metrics are essentially horizontal indicating that sample size has no effect on hit rate for each of these four metrics. Sample size does affect the hit rate for for the MCD and PCLOW, however. Surprisingly, the hit rate is highest for the $n = 80$ sample size. If there is an effect for sample size we would expect that higher hit rates would be associated with larger samples sizes, so it is interesting that the highest hit rates for MCD and PCLOW occur at $n = 80$. Since there is an observed three-way interaction we expect to see evidence that the two-way sample-size–metric interaction for number of variables equal to sixteen is different from the interaction for number of variables equal to eight. In 6.7, the plot does provide evidence that the two-way interaction is indeed different at this level for the number of variables. The decline in hit rates for PCLOW and MCD from $n = 80$ to $n = 160$ is much sharper in this plot. It is interesting to note that hit rates at $n = 80$ is nearly equal to that for the MLD, the overall best performing metric with respect to hit rate. They also exceed the hit rate for the Carrig D in the $n = 80$ condition. The mean hit rates for MCD and PCLOW did not exceed the mean hit rate for the Carrig D in the previous plot. Another interesting result is the fact that the hit rates for MCD and PCLOW drop from $n = 160$ to $n = 320$. Again this effect is the reverse of what we would expect.

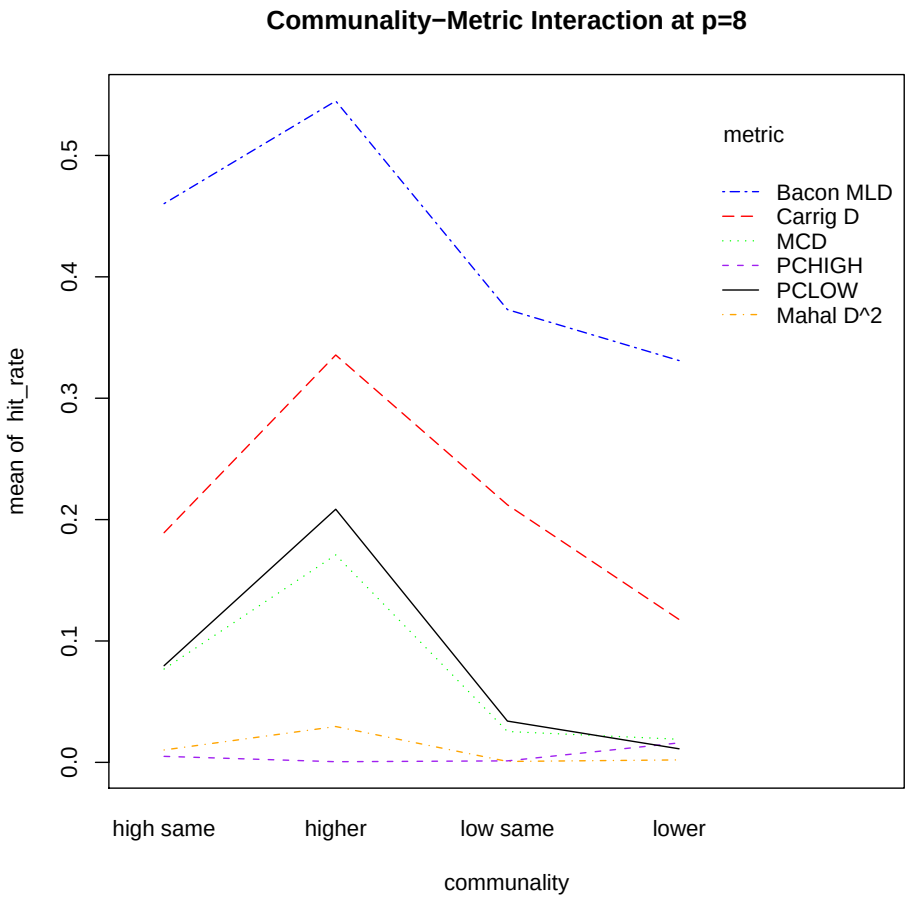

Figure 6.4: Communality-Metric Interaction on Hit Rate When Number of Variables=8

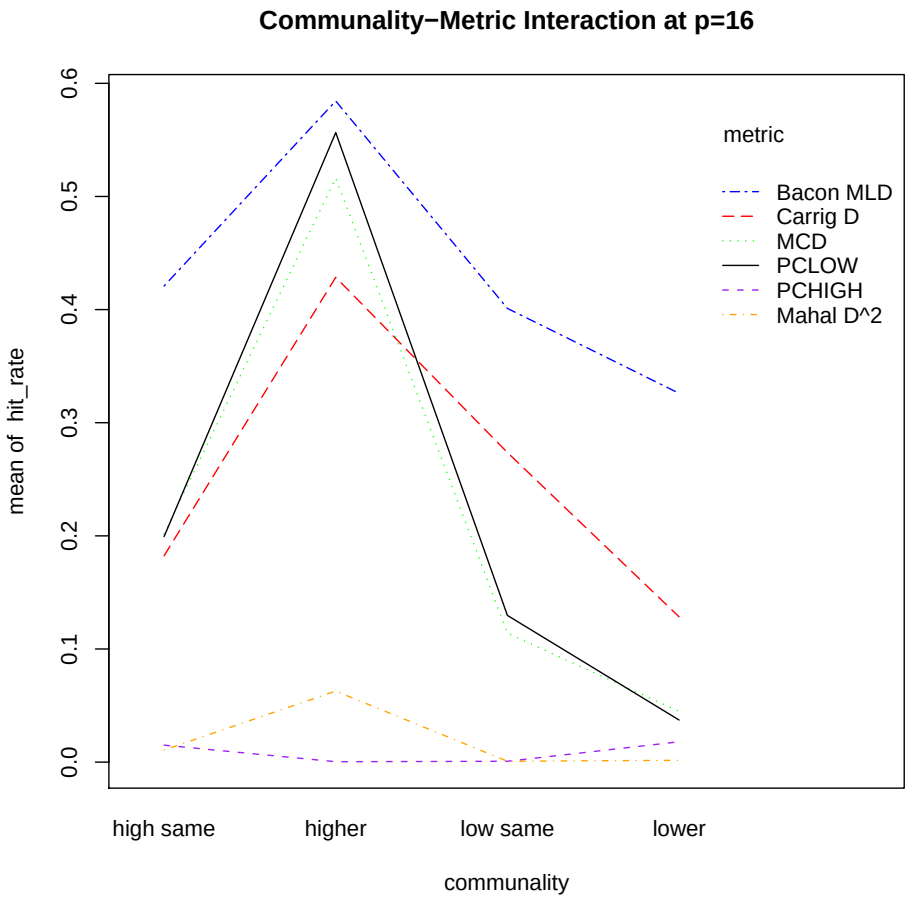

Figure 6.5: Communality-Metric Interaction on Hit Rate When Number of Variables=16

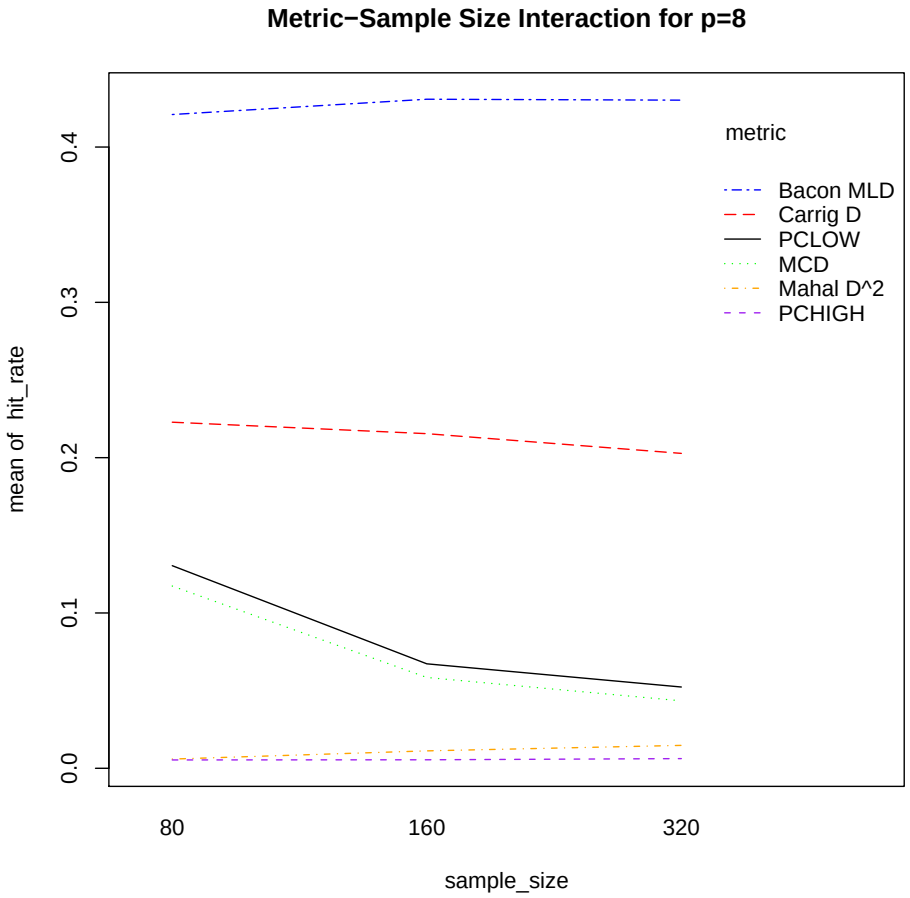

Figure 6.6: Metric–Sample-Size Interaction for Number of Variables Equal to Eight

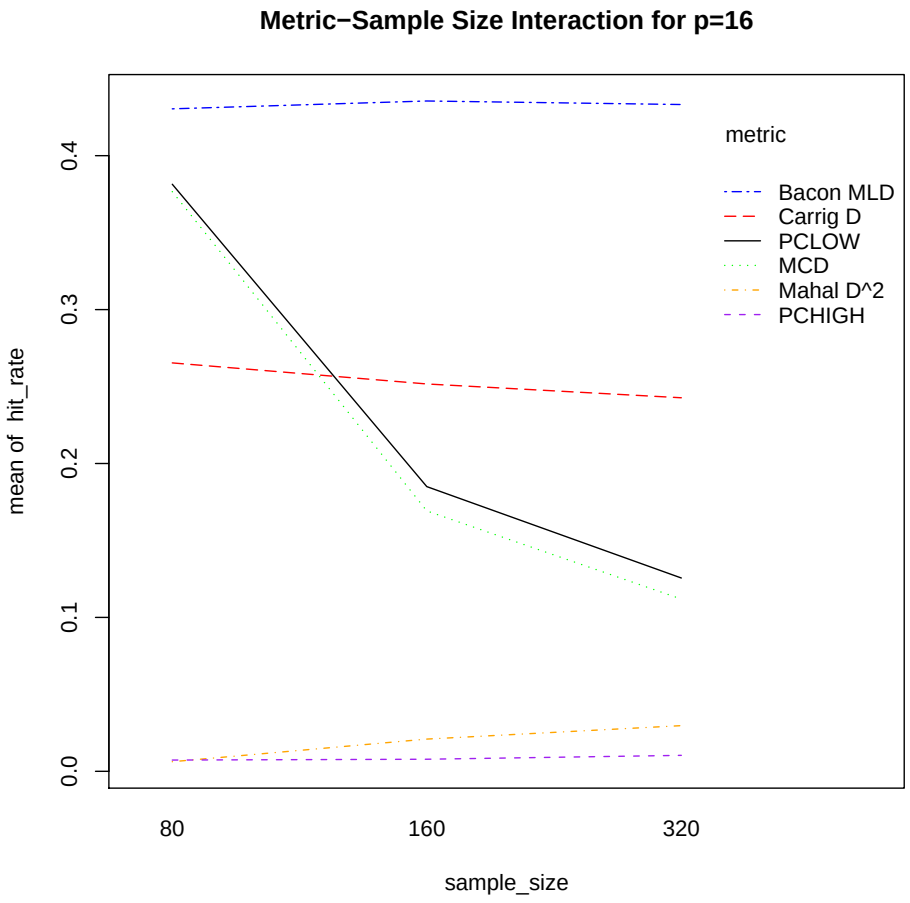

Figure 6.7: Metric–Sample-Size Interaction for Number of Variables Equal to Sixteen

PCLOW versus PCHIGH

It was hoped that the study would shed some light on the question of whether PCHIGH or PCLOW is more effective in spotting correlational outliers.

An answer in regard to mean hit rate can be inferred by looking at the plotsfor all of the practically significant effects: PCLOW outperforms PCHIGH in terms of mean hit rate except when the communality is lower. When communality is lower both PCHIGH and PCLOW had the same miserable mean hit rate, 0.02.

6.3 False-Alarm Rate ANOVA

Refer to Table 6.3. As was the case for the hit-rate ANOVA, the source table is divided into twoparts: between-samples effects and within-samples effects. Samples is the only random factor. The others are fixed factors.

6.3.1 Between-Samples Effects on False-Alarm Rate

Even though their effects were deemed practically significant, there will be no exploration of the main effects for communality and number of variables.This is due to the fact that the strength of association between these factors and the false-alarm rate is small combined with the fact that each of these effects is involved in interactions that have just as large strengths of association. The other between-samples main effect that was deemed practically significant was sample size. Additionally, its strength of association with false-alarm rate was large. However, this main effect will not be explored either as it is a part of interaction effects with medium strengths of association with false-alarm rate. While deemed practically significant, the number-of-variables–sample-size interaction had a low strength of association and is involved in a three-way interaction with metric whose strength of association is just as large. Therefore, this interaction will not be explored.

6.3.2 Within-Samples Effects on False-Alarm Rate

Unlike the large strength of association between the metric factor and hit rate, metric has only medium strength of association with false-alarm rate. Because of this and the fact that the metric effect is qualified by an interaction effect with sample size that also has a medium strength of association, the metric main effect will not be explored.

The Communality-Metric Interaction Effect

Refer to Figure 6.8. The significance of this effect is probably due in large part to the fact that the trend lines for PCHIGH and the Mahalanobis D^2 are horizontal and the lines for the other four metrics are clearly not. The trend line for the Carrig D is interesting. The false alarm rates for both the high same and higher communality conditions are both roughly equal to 0.1. The false-alarm rates double when we move to the low same and lower communality levels. The false-alarm rate is roughly 0.22 for both the low same and lower communalities. PCLOW and MCD exhibit roughly the same trend, but it is not as dramatic as for the Carrig D. However, the false-alarm rate for the lower communality is almost double the rate for the higher communality, 0.1 to 0.06. The false-alarm rates for the MLD are interesting. The rate is roughly 0.34 for the high same higher and lower. It is slightly higher for the low same condition, about 0.38. That said, the false-alarm rate for the MLD is essentially constant across levels of the communality. It should be noted that there was also a communality-metric interaction effect on hit rate, so the interaction of communality and metric has an effect on both hit rate and false-alarm rate. However, unlike the analysis of variance results for hit rate, the communality-metric interaction effect on false-alarm rate is not qualified by any of the other factors. One important thing to note is that despite the fact that the metric effect is qualified by communality the direction of the effect for metric is

consistent across all levels of communality. The false-alarm rate for the Carrig D is always higher than that for the PCLOW and MCD. The MLD has the highest false-alarm rate of all for all of the communality conditions.

Scenario-Metric Interaction

Like the analysis of variance for hit rate, there is a practically significant scenario-metric interaction effect. Refer to Figure 6.9. It appears that the scenario-metric interaction is due mainly to the trend line for the Carrig D. The trend lines for PCHIGH, Mahalanobis D^2, PCLOW and the MCD are almost perfectly horizontal. In other words, scenario has no effect on the false-alarm rates for these metrics. While the line for the MLD is not horizontal, the difference in false-alarm rates between scenario I and scenario II is not practically significant. The difference between the two is roughly 0.02. The Carrig D is the only metric that has a practically significant difference in false-alarm rate across the two levels of scenario; the false-alarm rate is about 0.08 higher for scenario II. Although scenario qualifies the metric effect the direction of the metric effect is consistent across both levels of scenario. MLD has the highest false-alarm rate for both scenarios by a substantial margin and the Carrig D is always larger than PCLOW and MCD although the difference is not that great under scenario I, roughly 0.04.

6.3.3 Number-of-Variables–Sample-Size–Metric Interaction

Both of the two factor interactions, sample-size–metric and number-of-variables–metrics were deemed to be practically significant. However, each of these of interactions is qualified in the form the number-of-variables–sample-size–metric interaction, so this three-way interaction will be discussed exclusively.

Refer to Figures 6.10 and 6.11. From these figures, it can be seen that a sample-size–metric interaction exists for both levels of number of variables, $p = 8$

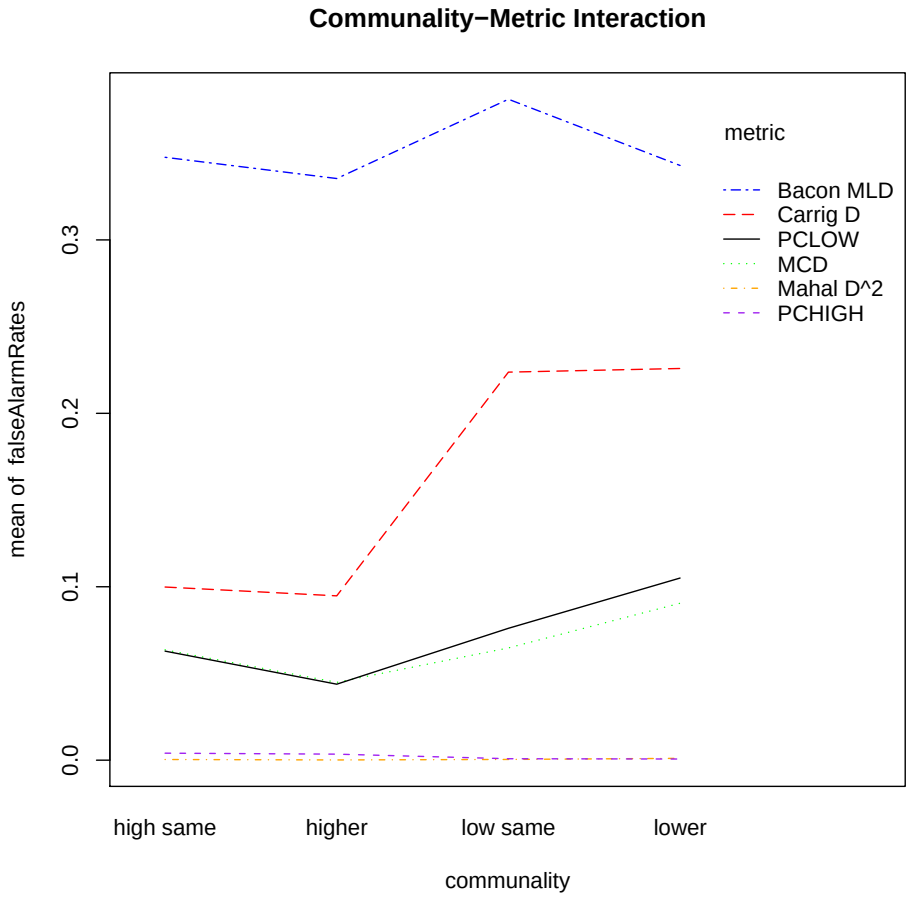

Figure 6.8: Depiction of Communality-Metric Interaction on False-Alarm Rate.

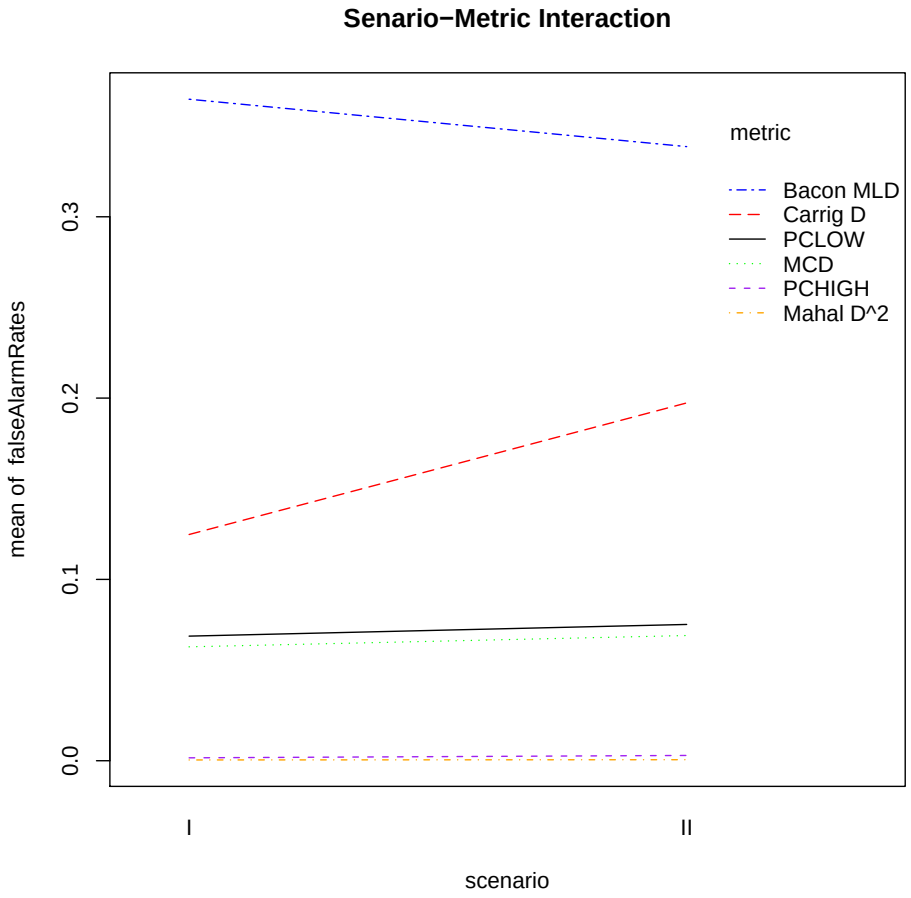

Figure 6.9: Scenario-Metric Interaction Effect on False-Alarm Rate.

and $p = 16$. In both cases, the interaction is due to fact that false-alarm rate varies across sample size for MCD and PCLOW only. The curves for MCD and PCLOW are essentially coincident in both of the plots. The false-alarm rate is essentially constant across sample size for the other four metrics. For $p = 8$, Figure 6.10, the false-alarm rate for MCD and PCLOW does not change very much in going from $n = 160$ to $n = 320$,0.01 to 0.005 for MCD, for example. However, there is a practically significant effect for the $n = 80$ sample size; the false-alarm rate jumps to 0.05 for this condition. Moving to the $p = 16$ plot, we see why there is a three-way interaction: the false-alarm rate skyrockets to 0.27 for MCD under the $n = 80$ condition, and there appears to be a practically significant difference in false-alarm rate between $n = 160$ and $n = 320$, 0.04 and 0.01, respectively for MCD.

6.4 PCLOW versus PCHIGH

The difference between the mean false-alarm rate for PCLOW and the mean false-alarm rate for PCHIGH where each rate is average over all experimental conditions is approximately 0.07. The mean of the false-alarm rate for PCLOW is approximately 0.07 averaged across all 144 experimental conditions. The mean false-alarm rate for PCHIGH averaged over all 144 experimental conditions is approximately 0.002. The mean false-alarm rate for PCLOW is greater in all of the 144 experimental conditions. The ratio of n to p makes the difference greater or less. When $n/p = 80/16 = 5$, the difference in false-alarm rate is approximately 0.27. When $n/p > 10$, the false-alarm rate for both PCHIGH and PCLOW is less than 0.01.

A communality

B scenario

C number of variables

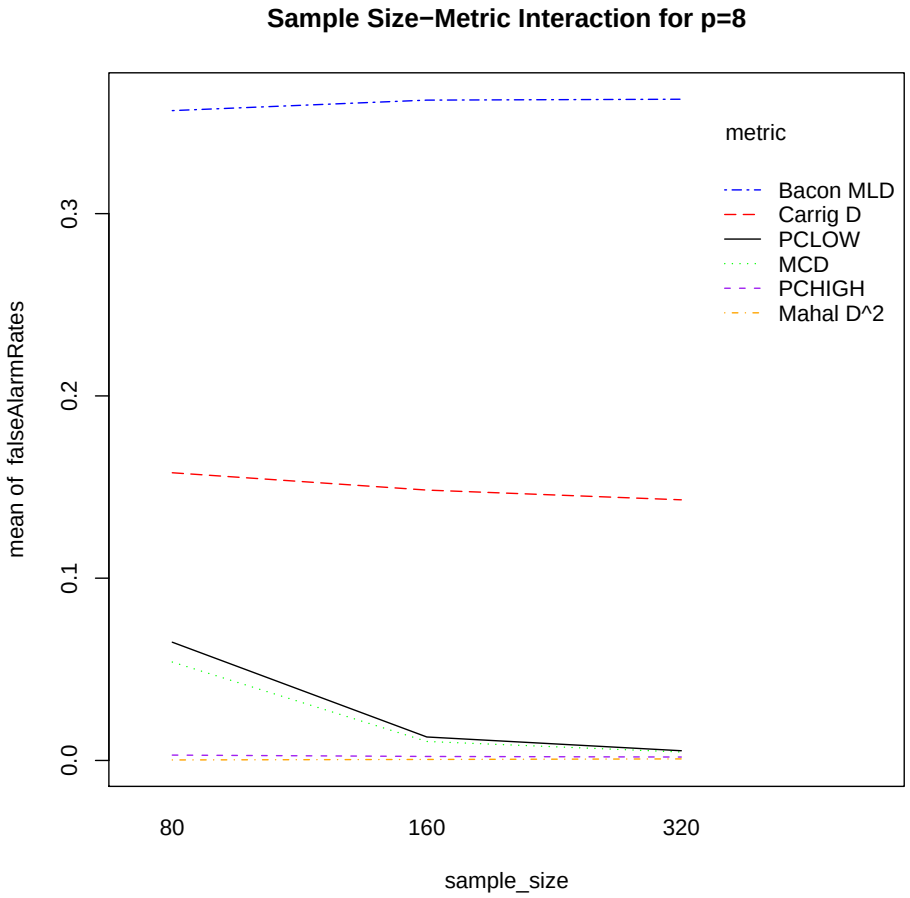

Figure 6.10: Sample-Size—Metric Interaction for $p = 8$

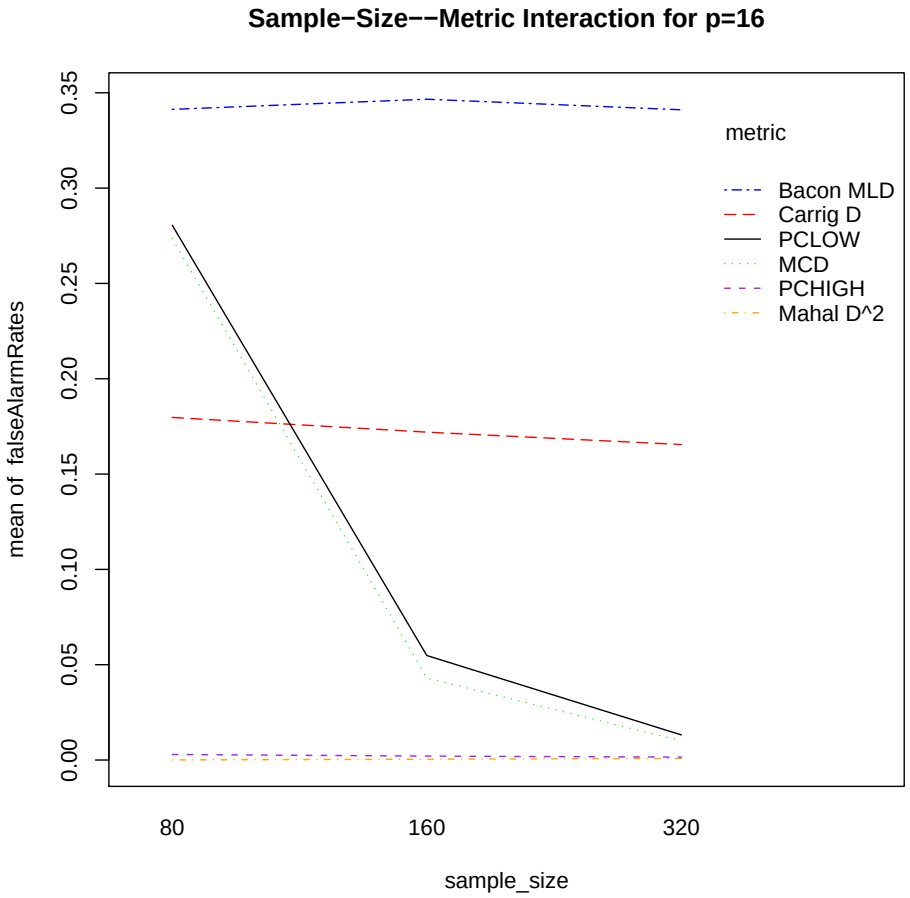

Figure 6.11: Sample-Size—Metric Interaction for $p = 16$

D sample size

F fraction of outliers

G sample or replicate

H metric

Table 6.3: Source Table for False Alarms

Source	df	SS	MS	ω^2
Between Samples	(186.550)			
A	3	26.519	8.840	0.024
B	1	2.201	2.201	0.002
C	1	19.258	19.258	.017
D	2	50.930	25.465	0.045
F	2	1.536	0.768	0.001
AB	3	1.058	0.353	0.001
AC	3	5.086	1.695	0.005
AD	6	1.764	0.294	0.002
AF	6	3.697	0.616	0.002
BC	1	0.193	0.193	< 0.0005
BD	2	0.166	0.083	< 0.0005
BF	2	2.617	1.308	0.002
CD	2	21.116	10.558	0.019
DF	4	0.025	0.006	< 0.0005
ABC	3	0.137	0.046	< 0.0005
ABD	6	0.124	0.021	< 0.0005
ABF	6	1.496	0.249	0.001

Continued on next page

Source	df	SS	MS	ω^2
ACD	6	0.384	0.064	< 0.0005
ACF	6	0.396	0.066	< 0.0005
ADF	12	0.329	0.027	< 0.0005
BCD	2	0.019	0.010	< 0.0005
BCF	2	0.022	0.011	< 0.0005
BDF	4	0.016	0.004	< 0.0005
CDF	4	0.010	0.002	< 0.0005
ABCD	6	0.010	0.002	< 0.0005
ABDF	12	0.057	0.005	< 0.0005
ACDF	12	0.196	0.016	< 0.0005
BCDF	4	0.008	0.002	< 0.0005
ABCDF	12	0.047	0.004	< 0.0005
G(A B C D F)	14256	47.707	0.003	
Within Samples	(72,000)	(1561.168)		
H	5	1267.739	253.548	0.061
AH	15	47.443	3.163	0.023
BH	5	19.513	3.903	0.009
CH	5	38.405	7.681	0.018
DH	10	93.990	9.399	0.045
ABH	15	16.061	1.071	0.008
ACH	15	4.645	0.310	0.002
ADH	30	3.305	0.110	0.002
AFH	30	5.337	0.178	0.003

Continued on next page

Source	df	SS	MS	ω^2
BCH	5	0.249	0.050	< 0.0005
BDH	10	0.232	0.023	< 0.0005
BFH	10	4.381	0.438	0.002
CDH	10	41.316	4.132	0.020
CFH	10	0.242	0.024	< 0.0005
DFH	20	0.301	0.015	< 0.0005
ABCH	15	0.206	0.014	< 0.0005
ABDH	30	0.429	0.014	< 0.0005
ABFH	30	4.877 0.163	0.002	
ACDH	30	1.162	0.039	0.001
ACFH	30	1.051	0.035	< 0.0005
ADFH	60	0.995	0.017	< 0.0005
BCDH	10	0.036	0.004	< 0.0005
BCFH	10	0.076	0.008	0.0005
BDFH	20	0.222	0.011	< 0.0005
CDFH	20	0.110	0.006	< 0.0005
ABCDH	30	0.032	0.001	< 0.0005
ABCFH	30	0.133	0.004	< 0.0005
ABDFH	60	0.455	0.008	< 0.0005
ACDFH	60	0.356	0.006	< 0.0005
BCDFH	20	0.030	0.001	< 0.0005
ABCDFH	60	0.138	0.002	< 0.0005
Error	71280	175.901	0.002	
Total	86399	1747.718		

Continued on next page

Source	df	SS	MS	ω^2

Chapter 7

Discussion

7.1 Summary of Results

7.1.1 Effects on Hit Rate

Most of the conclusions that follow were arrived at by looking at the plots for all of the practically significant effects on mean hit rate simultaneously. Refer to Figures 6.1, 6.2, 6.3, 6.4, 6.5, 6.6 and 6.7. Once again, because of the consistently miserable performance of the Mahalanobis D^2 and PCHIGH, the discussion that follows only deals with the MCD, PCLOW, Carrig D and Bacon MLD unless other wise stated.

All of the metrics performed best under the higher communality condition and worst under the lower communality condition. This was consistently true across all of the levels of other factors. This was consistent with Bacon's findings and was thus not a surprise (Bacon, 1995).

The difference in mean hit rates between the higher communality condition and the high same communality condition was larger than I expected. This effect was not really qualified by metric or any of the other factors. The mean hit rate under the communality higher condition was approximately 0.29. For the high same communality condition it was almost fifty percent less at approximately 0.15. I will

elaborate later on why this result surprised me and just why it did occur despite my expectations.

Compared to other metrics, MCD and PCLOW had the most dramatic difference in mean hit rates between the higher communality and lower communality conditions. This was not a surprise given Bacon's findings (Bacon, 1995).

Although PCLOW outperformed MCD under the factor combination of $p = 8$, scenario II and communality higher the difference was not as great as I expected. Under this combination PCLOW had a mean hit rate of 0.30 while MCD had a hit rate of 0.25.

The effects of communality on hit rate were essentially constant across the four metrics under scenario II. This was surprising given that one of the main conclusions of Bacon's study was that the change in communality did not have as great an effect on the MLD as the Mahalanobis D^2 (Bacon, 1995). The communality effects on mean hit rate were also essentially constant across metric when $p = 8$. Again, surprising given Bacon's finding just mentioned.

The effect of p, the number of variables, on hit rate was only substantial for MCD and PCLOW. The mean hit rates were higher under $p = 16$ compared to $p = 8$. This is consistent with statements made by Carrig. In regard to MCD, Carrig states that an increase in p results in an increase in information available to the metric leading to better estimates. In regard to the lack of a p effect on the Carrig D and MLD, Carrig contends that this is not surprising given that the MLD and Carrig D are sums of bivariate statistics (Carrig, 2005).

The result that the MCD and PCLOW had higher mean hit rates as the n/p ratio decreased (for MCD, mean hit rate = 0.37 for $n/p = 80/16 = 5$ and 0.04 for $n/p = 320/8 = 40$) suprised me at first. My reasoning was that higher n/p ratios lead to better parameter estimates and thus should yield better mean hit rates. When I took into account that the mean false-alarm rates also skyrocketed when

$n/p < 10$, I found an explanation for this curious effect. The explanation is detailed below.

The result that surprised me the most initially was that the MLD had the highest mean hit rates in all of the 144 experimental conditions. Could it be that the MLD is resistant to the masking effect? This would be surprisiing indeed given that the MLD uses a term, $\overline{\text{MLCE}}_{.jk}$ (see Equation 2.29), in its computation that is a equally weighted mean across all observations. This matter will be explored below.

PCLOW clearly outperformed PCHIGH in nearly all of the experimental conditions with the exception of the lower communality condition where their mean hit rates are similar. This result is consistent with the assertions of Jolliffe and Hawkins that PCLOW will tend to spot observations that are outlying with respect to correlational structure (Jolliffe, 1986; Hawkins, 1974). Also note that this result is not consistent with the assertion of Gnanadesikan and Kettenring that observations that are outlying with respect to correlational structure will be outlying with respect to the high principal components (Gnanadesikan & Kettenring, 1972).

7.1.2 False-Alarm Rates

As was the case in the discussion of effects on mean hit rate, PCHIGH and the Mahalanobis D^2 are excluded from the discussion unless otherwise noted. Again, their consistently abyssmal performance in terms of mean hit rate makes their inclusion in the discussion moot.

The effect of metric on the mean false-alarm rates is, in general, consistent across all of the experimental conditions. The overall mean false-alarm rate for the four metrics: MCD, PCLOW, Carrig D and Bacon MLD are 0.06, 0.07, 0.16 and 0.35, respectively. Clearly, the mean false-alarm rate for the MLD is unacceptably large. Using the MLD in the way that it was used in this study obviously cannot

be recommended. The fact that the MLD had the largest mean false-alarm rate by far is surprising given a result from Carrig's study. She reported that the MLD had the smallest mean false-alarm rate under her "Different Shape/No Mean Slippage ", "Natural drop-off "condition. The outliers in this condition came from a distribution with the same mean vector as the majority population but having a covariance matrix with a shape different from the one for the majority distribution. So, the outliers in this condition were purely correlational in nature just like all of the conditions in my study. "Natural drop-off "refers to the fact that she used the k-means algorithm to separate the observed MLDs into two clusters. Observations in the cluster with the higher mean were flagged as potential outliers. This is exactly the same way in which I used the MLDs to spot potential outliers in my study. Thus my surprise that the MLD had the smallest false-alarm rate under these conditions in her study. However, the mean false-alarm rate she reported for her Different Shape/No Mean Slippage condition was 0.22, not as high as the overall mean in my study but still unacceptably large.

Refer to Figure 6.9. Why is it that scenario only affects the mean false-alarm rates for the Carrig D? The mean false-alarm rate under scenario I is approximately 0.12. Under scenario II it is approximately 0.20.

The intolerably high false-alarm rate for the MLD is fairly constant across all of the experimental conditions. Refer to Figures 6.8, 6.9, 6.10 and 6.11. Ironically, the mean false-alarm rates are high for the same reason that the mean hit rates for the MLD are high. The reason is discussed below.

The effect of scenario on the false-alarm rate for the Carrig D was interesting. Refer to Figure 6.9. The Carrig D was the only metric that was substantively affected by scenario in regard to false-alarm rate. The false-alarm rate under scenario II is almost double that for Scenario I.

7.2 n/p Effect

One of the results that seemed surprising was the fact that MCD and PCLOW had their highest hit rates by far when the sample size was 80. I expected the reverse; highest mean hit rates at $n = 320$. The effect was most pronounced when $p = 16$. However, the false-alarm rates were also at their highest levels when $n = 80$ and $p = 16$, (See Figures 6.6 and 6.7). The mean false-alarm rates are so high in fact that I would not recommend using MCD or PCLOW when the n/p ratio is five. These results for hit rate and false-alarm rate suggest that the mechanism at work is the n/p ratio. I found it useful to think in terms of MCD and PCLOW having more power when $n/p = 80/16 = 5$ than when $n/p = 320/16 = 20$. MCD and PCLOW for the majority observations are assumed to be distributed as chi-square random variables, and the degrees of freedom for each are the same across the two aforementioned combinations of n and p. This assumption along with the increased power at $n/p = 80/16 = 5$ led me to believe that the distributions MCD of PCLOW at $n/p = 80/16 = 5$ are not the same as their counerparts when $n/p = 320/16 = 20$.

I investigated this conjecture by generating random samples from a $n/p = 80/16 = 5$ experimental condition and a $n/p = 320/16 = 20$ experimental condition. For each sample I computed the MCDs for the majority observations. I then constructed two Q-Q plots of the MCDs of the majority observations versus the expected values under the assumption that the MCDs came from a chi-square distribution with sixteen degrees of freedom. Recall this assumption is based on the idea that the MCDs for the majority observations will be approximately equal to

$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2_p \tag{7.1}$$

Specifically, the assumption is that for a "large "sample the estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ obtained from the MCD robust estimation algorithm will be close to the true

population values, and, thus, the MCD distances for the majority observations will be approximate chi-square random variables with p degrees of freedom. The MCDs for the contaminants will not be realizations of the chi-square distribution for the majority MCDs because the population correlation matrix for the contaminants is different from the one the one for the majority population. So, the Q-Q plots are for the MCDs of the majority observations only. This does not mean that I did not bother to generate the contaminants. Recall that the robust estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are estimated from the entire data set of both majority and contaminant observations. I will next describe the specifics of how I generated random samples for the $n/p = 80/16 = 5$ and $n/p = 320/16 = 20$ conditions.

I selected two experimental conditions in which the number of variables equaled sixteen and differed only with respect to sample size. Specifically I chose the condition of higher communality, scenario I, $p = 16$, $n = 80$, and fraction of outliers equal to 0.16 and the condition with the same values on all of the factors except that $n = 320$. The first condition is representative of $n/p = 80/16 = 5$ and the second is reperesentative of the $n/p = 320/16 = 20$. The correlation matrices for the majority distribution and the contaminant distribution for this combination of communality, scenario and number of variables are contained in Equations 4.7 and 4.9. Since all of the factors except sample size are the same for both conditions comparison of a random sample from one condition to a random sample from the other will give us an idea of how the distributions of MCDs compare for the two n/p values.

Since the fraction of outliers is 0.16 the random sample for $n/p = 80/16 = 5$ consisted of 67 majority observations and 13 contaminants. I computed the MCDs for the majority observations in the same manner that I did in the main experiment. The Q-Q plot for a random sample from this condition is contained in Figure 7.1

The Q-Q plot tells us that the 67 majority MCDs most certainly did not

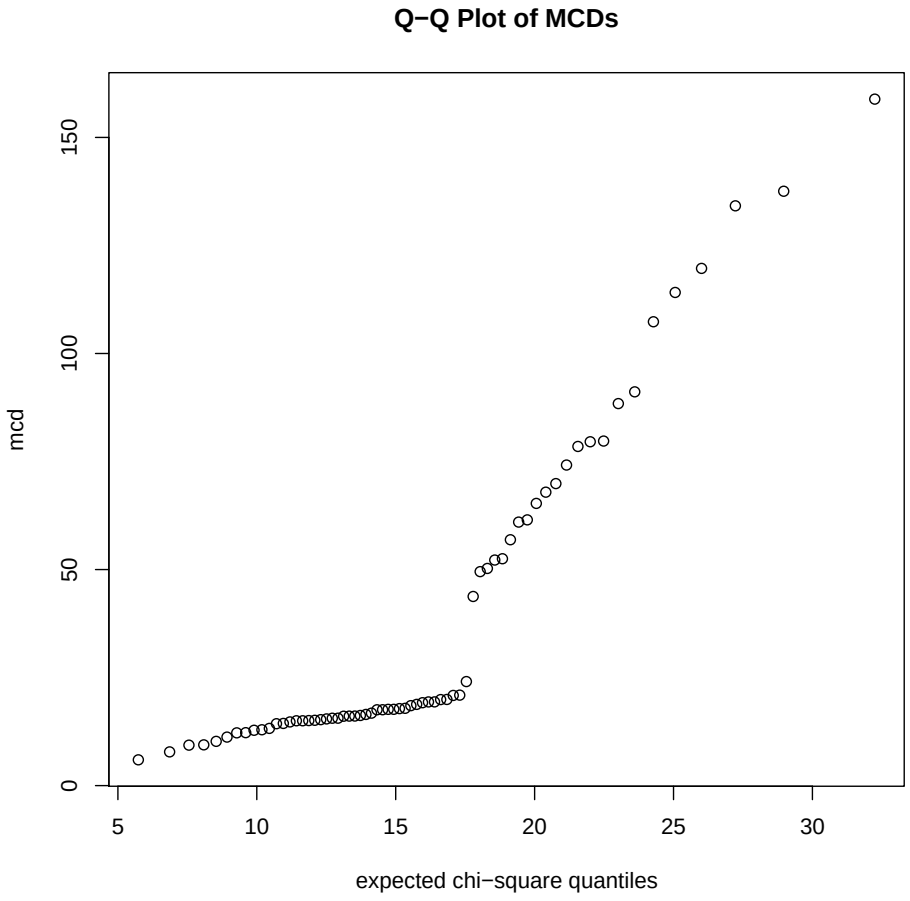

Figure 7.1: Q-Q plot of MCDs for the 67 majority observations taken from a sample of 80 observations under communality higher, scenario I, $p = 16$ and fraction of outliers equal to 0.16

come from a chi-square distribution with sixteen degrees of freedom. We see that the largest fifteen MCDs are considerably larger than the expected values of their corresponding chi-square order statistics. Recall that the cut-off value for the MCD in the main experiment was $\chi^2_p(0.999)$ which in this case is approximately 39. We see from the plot that roughly 13 or 14 of the majority MCDs exceed this value. How could this happen? I conjectured that the correlation matrix output by the MCD algorithm and subsequently used in the computation of MCD distances underestimates the generalized variance of the true multivariate normal population. Before investigating this let us look at the Q-Q plot for $n/p = 320/16 = 20$. The Q-Q for a random sample from this condition is contained in Figure 7.2.

The Q-Q plot supports the assumption that the 268 majority MCDs are realizations of a chi-square random variable with 16 degrees of freedom. Next, I tested my conjecture that the correlation matrix output by the MCD algorithm will, in general, underestimate the generalized variance of the majority population.

Johnson and Wichern define generalized variance to be the determninant of a covariance matrix (1992). The larger the determinant of a sample correlation matrix the more "variable "it is in p-space where p is the dimensionality or the number of variables. Consider Figure 7.3. The blue points came from a sample in which x_1 and x_2 were uncorrelated, $\rho = 0$, and the red points came from a bivariate normal distribution with $\rho = 0.9$. For each sample rotate the figure until you have an elliptical pattern of uncorrelated variables. Draw the axes of this ellipse through the centroid which we take to be $(0, 0)$ for both samples. The assumption is being made that the centroids for both data sets are the same. This is not unreasonable with this data. This is precisely the same exercise in which we defined new coordinate axes for correlated data in the discussion of statistical distances (Section 2.3.2). Now, for one of the samples, imagine an ellipse defined by the axes constructed that tightly contains all of the sample points. Roughly speaking, the determinant of the

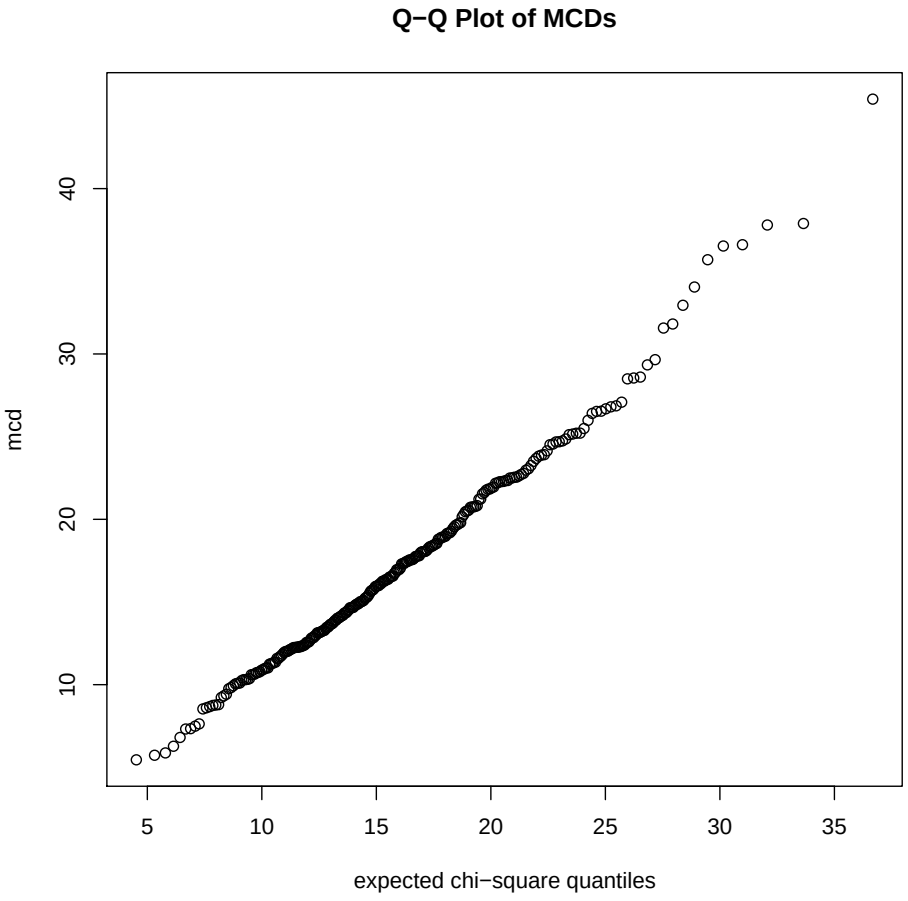

Figure 7.2: Q-Q plot of MCDs for the 268 majority observations taken from a total sample of size 320 under communality higher, scenario I, $p = 16$ and fraction of outliers equal to 0.16

correlation matrix is proportional to the area of that ellipse. The area of the ellipse for the blue sample is clearly larger than the ellipse for the red sample. Thus the determinant, or generalized variance, for the blue sample is larger. This jibes with a first impression of the scatterplot namely that the blue sample is spread out over a larger area than the red sample.

It is obvious in the univariate case that if our estimate of population of the population standard deviation, $\hat{\sigma}$, is a substantial underestimate, then the statistic, $\frac{(x-\overline{x})}{\hat{\sigma}}$ will flag some legitimate or majority observations as outliers, but it will also have increased power in the detection of real outliers. The same is true in the multivariate case when the estimate of the population covariance or correlation matrix, $\mathbf{S}$ or $\mathbf{R}$, has a generalized variance that is substantially less than the generalized variance corresponding to the population covariance or correlation matrix. Now, back to the task of establishing that under the $n/p = 80/16 = 5$ condition the MCD algorithm outputs estimated correlation matrices that underestimate the multivariate variance of the majority population.

I conducted a small-scale—Monte Carlo experiment to investigate whether the MCD algorithm under the $n/p = 80/16 = 5$ will generally produce correlation matrices that underestimate the generalized variance of the majority population. I generated 500 samples from each of the two experimental conditions used in the generation of the Q-Q plots. For each sample, the determinant of the correlation matrix output by the MCD algorithm was calculated. Figure 7.4 contains boxplots which compare the distributions of the 500 determinants for the $n/p = 80/16 = 5$ condition and the 500 determinants for the $n/p = 320/16 = 20$ condition.

We see that the results of the simulation suggest that the MCD algorithm consistently underestimates the determinant, and, thus, the generalized variance, of the population correlation matrix for the majority distribution (note that the red line in the plot corresponds to the determinant of the population correlation matrix).

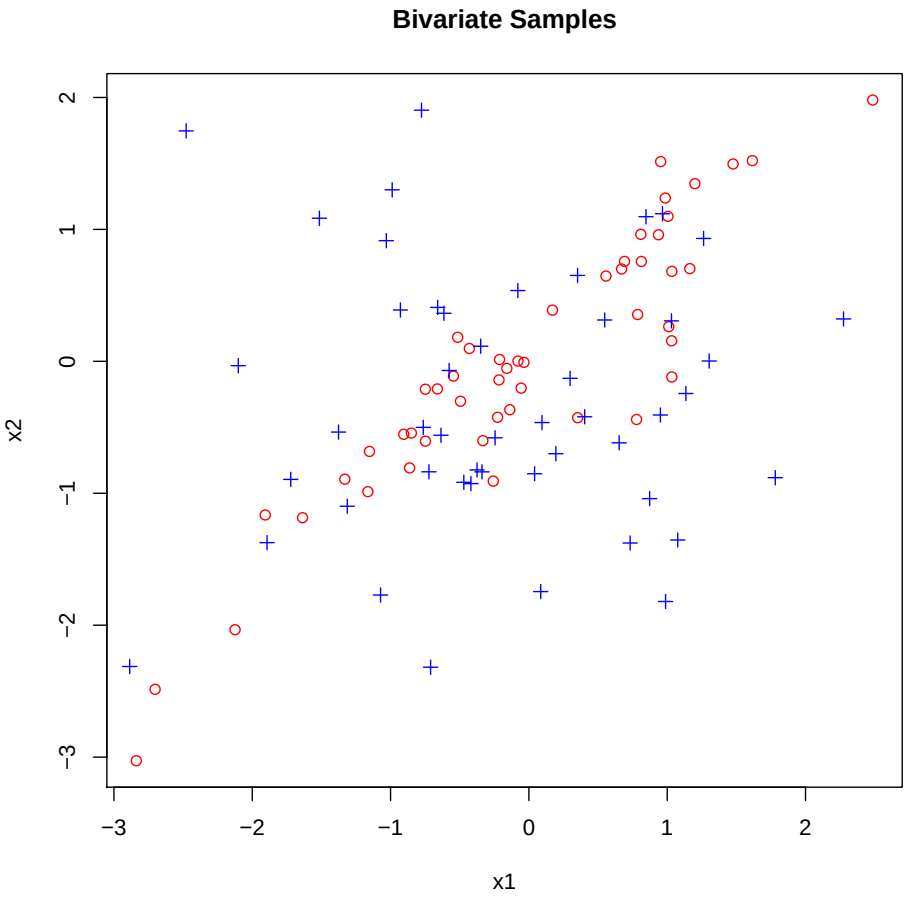

Figure 7.3: Scatterplot of two bivariate samples. The blue one was generated from bivariate normal distribution with $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\rho} = \mathbf{I}$. The distribution for the red sample had the same mean vector but its correlation matrix had $\rho = 0.9$ as the off-diagonal element.

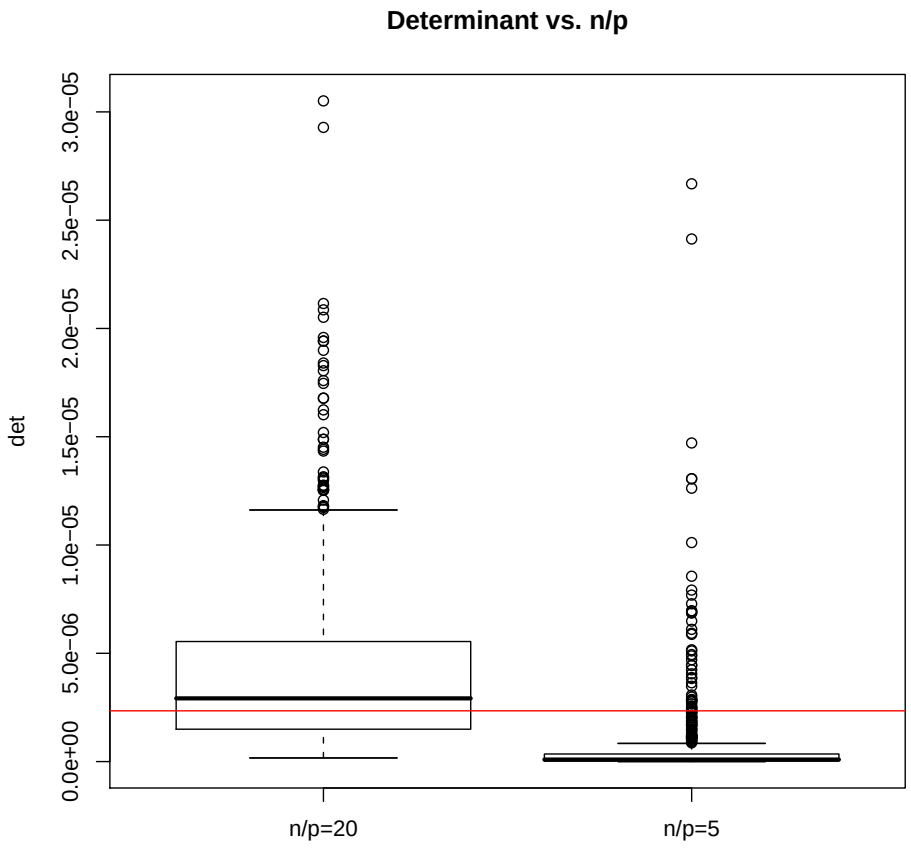

Figure 7.4: Comparative boxplots of the distributions of the 500 determinants for the $n/p = 80/16 = 5$ condition and the 500 determinants for the $n/p = 320/16 = 20$ condition. Red line corresponds to the determinant of the population correlation matrix.

For the $n/p = 80/16 = 5$ condition, 90.6% of the determinants are less than the determinant of the population correlation matrix. Thus, the MCD and PCLOW have higher mean hit rates and higher false-alarm rates under the $n/p = 80/16 = 5$ condition due to the underestimation of the generalized variance of the majority population. It is rather interesting to note that the IQR for $n/p = 80/16 = 5$ is many times smaller than the IQR for $n/p = 320/16 = 20$.

The skew to the right of the $n/p = 320/16 = 20$ is interesting. The third quartile, Q_3, of the distribution is approximately 5.54×10^{-6} more than double the value of the true population generalized variance. This indicates that in a substantial number of cases there is large overestimation of the generalized variance. It is reasonable to conclude that in these cases the power of the MCD distance to spot the outliers will be greatly diminished. The MCD will perform poorly in these samples. I added PCLOW and the Carrig D to my conjecture because, like the MCD, they depend on the correlation matrix output by the MCD algorithm. I conjectured that this large variability in the estimate of the generalized variance would lead to a larger amount of variability in the performance of the MCD distance, PCLOW and Carrig D in terms of both hit rate and false-alarm rate in the $n/p = 320/16 = 20$ condition in comparison to the $n/p = 80/16 = 5$ condition. Note that my conjecture implies that the variability in the performance of the metric is due in part to the application of the metric and not sample to sample variability within the experimental cell solely. In fact, I conjecture that there will be more variability in hit rate under the $n/p = 80/16 = 5$ condition for MCD, PCLOW and the Carrig D because all of these metrics depend on the robust correlation matrix that is output by the MCD algorithm. I decided to look at the distributions of the 100 hit rates and false-alarm rates for the three metrics within the experimental cells corresponding to the levels of the factors used in the generation of the 500 determinants for both the $n/p = 320/16 = 20$ and $n/p = 80/16 = 5$ example. Specifically, to test my

conjecture I needed to compare the variability in hit rates and false-alarm rates across the three n/p conditions, $n/p = 5$; $n/p = 10$; $n/p = 20$, for the MCD, PCLOW and the Carrig D. I made these comparisons via a comparative boxplot for each of the three metrics. Refer to Figures 7.5, 7.6, 7.7, 7.8, 7.9 and 7.10. Within each of these figures a single boxplot is of the 100 hit rates or false-alarm rates for the experimental cell: communality higher, scenario I, $p = 16$, fraction of outliers equal to 0.16 and the sample size on the abscissa.

The comparative boxplot for MCD hit rates refuted my conjecture (see Figure 7.5). We see that the distribution of hit rates is less variable under the $n/p = 320/16 = 20$ condition than the $n/p = 80/16 = 5$ in terms of the interquartile range. The interquartile range for the former is approximately 0.101 and 0.231 for the latter. It is interesting to note that the distributions of the hit rates under the $n/p = 80/16 = 5$ and $n/p = 160/16 = 10$ are skewed to the left while the distribution under $n/p = 320/16 = 20$ is roughly symmetrical. Perhaps an explanation for these results is that while there is little variability in the estimate of generalized variance under the $n/p = 80/16 = 5$ condition there is variability in the estimation of the directions of the principal components. And, under the $n/p = 320/16 = 20$ condition while we have relatively large variability in the estimate of generalized variance there is little variability in the estimates of the directions of the principal components. We would expect to see the same relationships for false alarm rates. The relationships are of the same nature but are more dramatic. The drop in median false-alarm rate is percipitous in going from the $n/p = 80/16 = 5$ condition to the $n/p = 160/16 = 10$ condition. Moreover, the variability of false-alarm rate within the $n/p = 160/16 = 10$ and $n/p = 320/16 = 20$ conditions is drastically less than the variability within the $n/p = 80/16 = 5$. It is not surprising that the comparative boxplots for PCLOW are almost identical to those for the MCD given their almost identical performance across most of the experimental conditions.

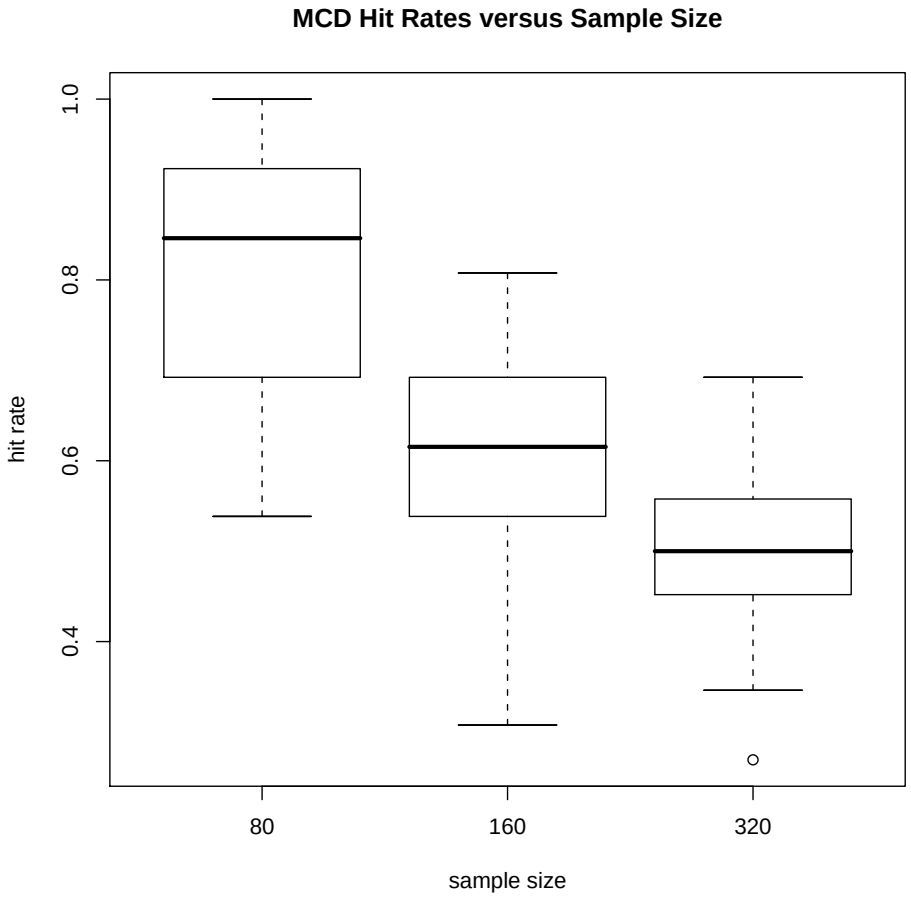

Figure 7.5: Distributions of MCD hit rates under three different experimental conditions that differ only with respect to sample size. The levels of the other factors are communality higher, scenario I, $p = 16$, fraction of outliers equal 0.16.

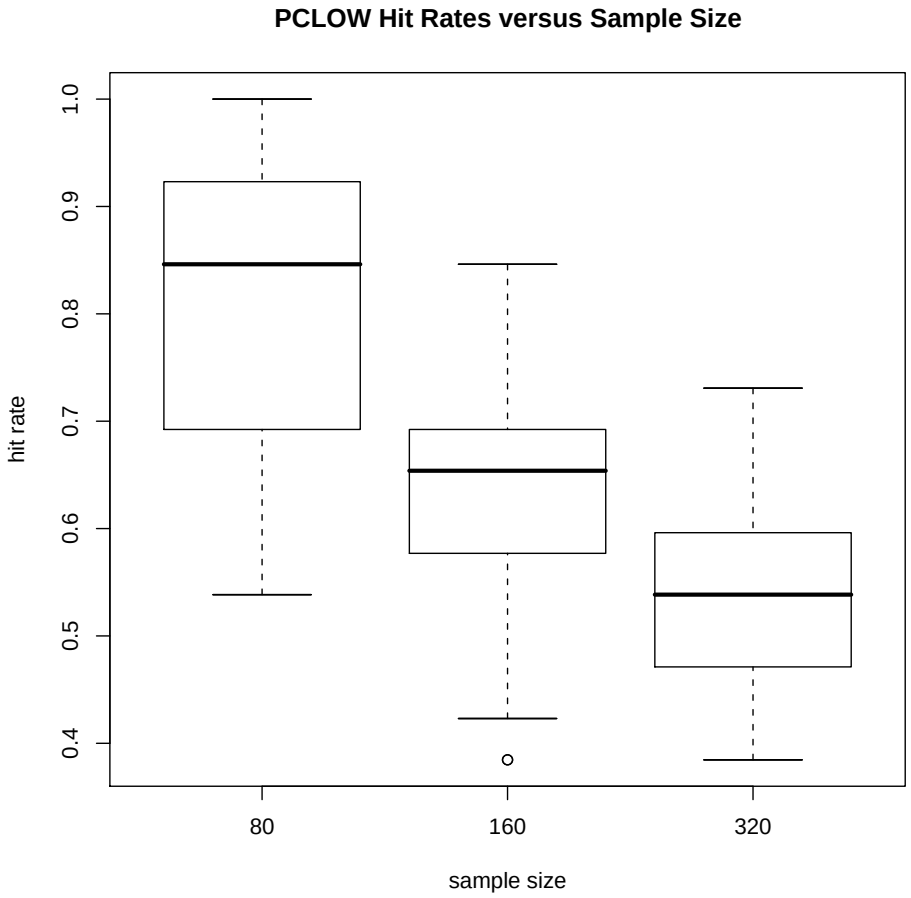

Figure 7.6: Distributions of PCLOW hit rates under three different experimental conditions that differ only with respect to sample size. The levels of the other factors are communality higher, scenario I, $p = 16$, fraction of outliers equal to 0.16

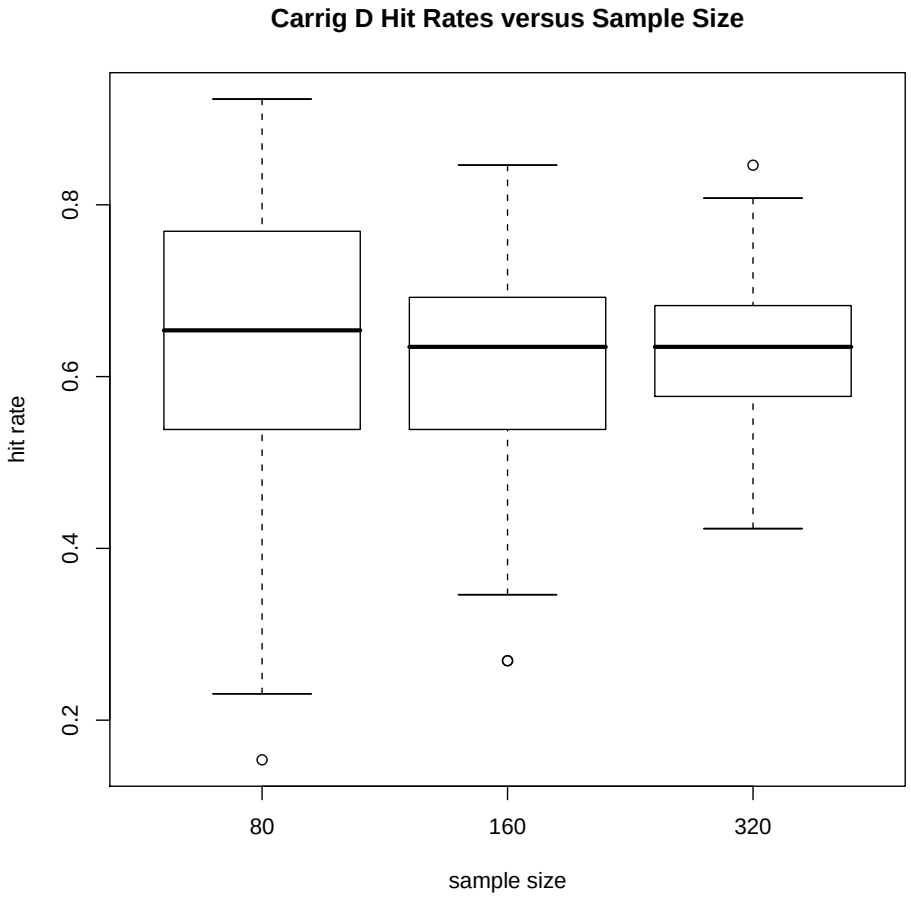

Figure 7.7: Distributions of Carrig D hit rates under three different experimental condition that differ only with respect to sample size. The levels of the other factors are communality higher, scenario I, $p = 16$, fraction of outliers equal to 0.16

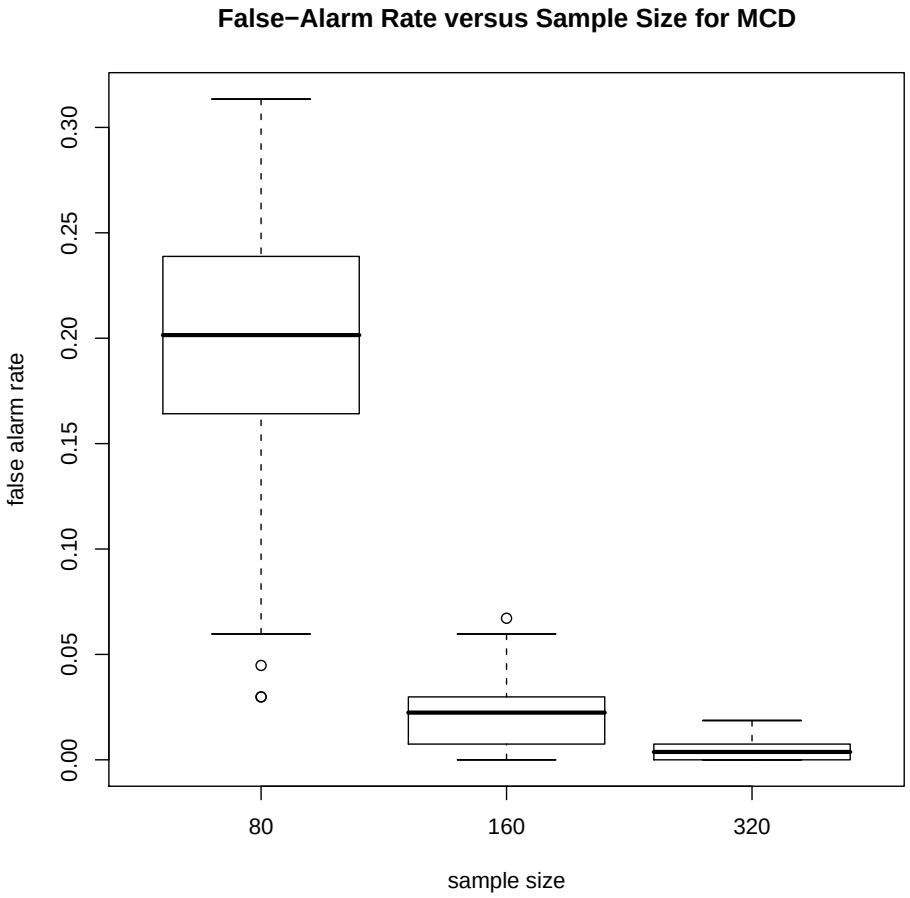

Figure 7.8: Distributions of MCD false-alarm rates under three different experimental conditions that differ only with respect to sample size. The levels of the other factors are communality higher, scenario I, $p = 16$, fraction of outliers equal to 0.16.

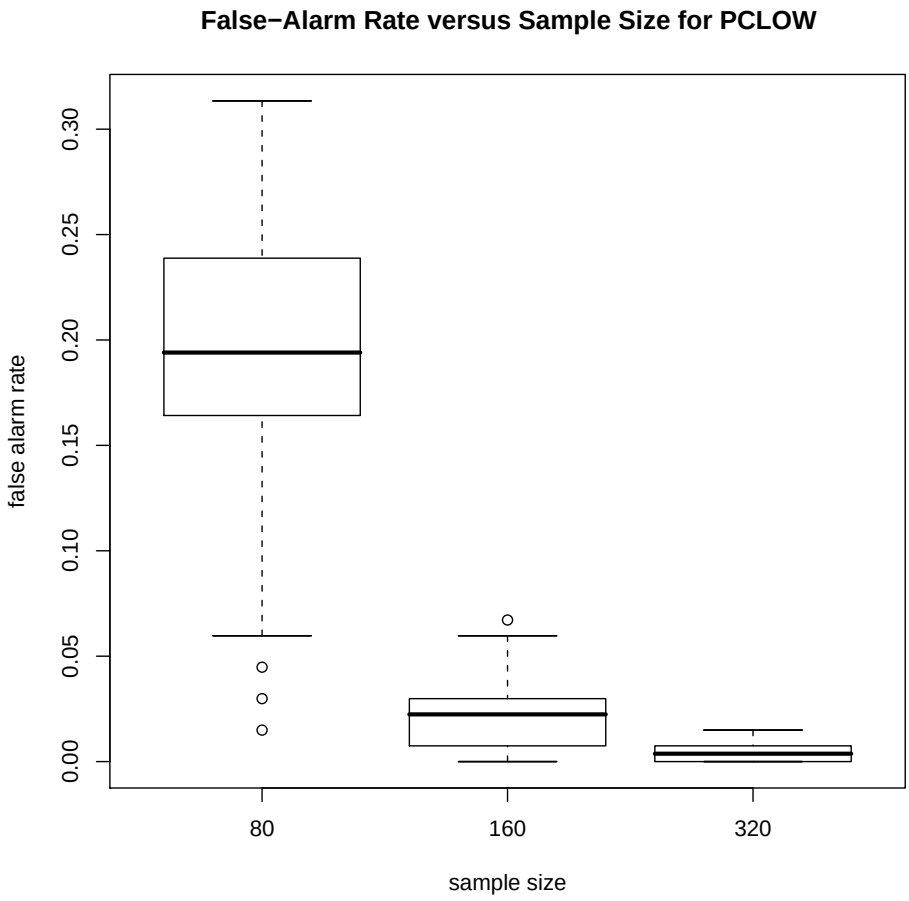

Figure 7.9: Distributions of PCLOW false-alarm rates under three different experimental conditions that differ only with respect to sample size. The levels of the other factors are communality higher, scenario I, $p = 16$, fraction of outliers equal to 0.16.

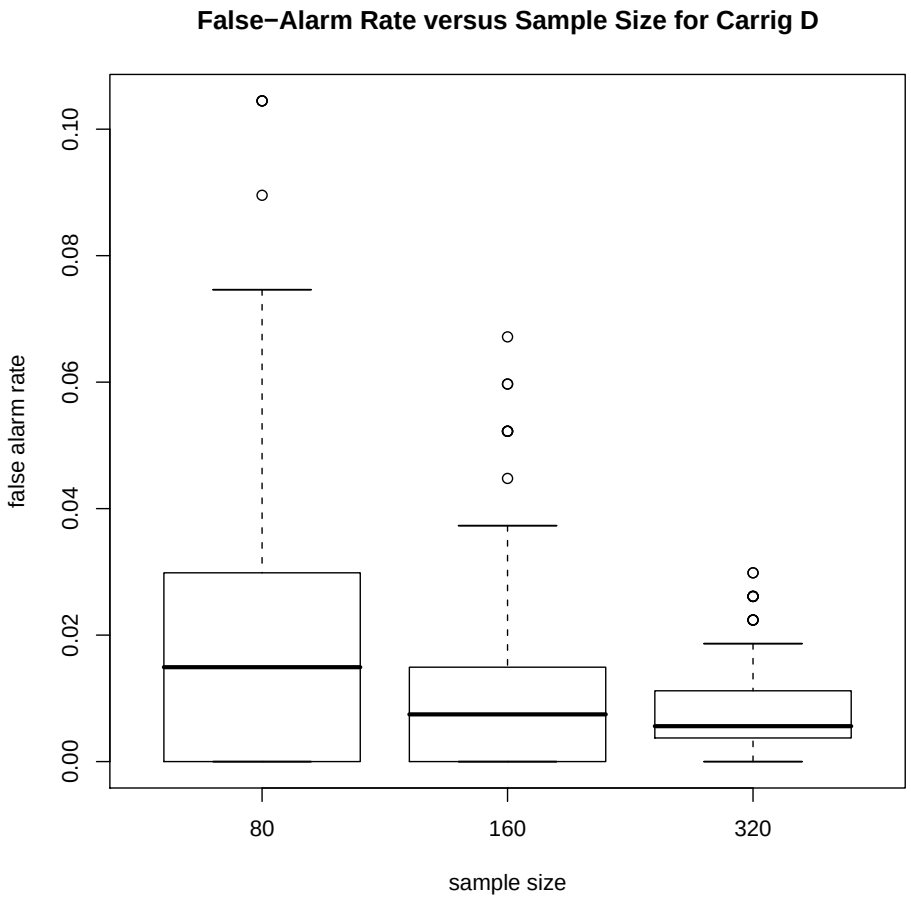

Figure 7.10: Distributions of Carrig D false-alarm rates under three different experimental conditions that differ only with respect to sample size. The levels of the other factors are communality higher, scenario I, $p = 16$, fraction of outliers equal to 0.16.

Regardless of the reasons why the foregoing relationships between performance of the MCD and the n/p ratio exist, one thing is clear, the MCD and PCLOW should probably not be used when the n/p ratio is around five. This is consistent with n/p guidelines for various kinds of multivariate analyses. In his widely used book, *Applied Multivariate Statistics for the Social Sciences*, Stevens suggests a minimum of five observations per variable when conducting a factor analysis. However, he also adds that others have recommended a minimum of twenty observations per variable (n.d.).

Figures 6.6, 6.7, 6.10 and 6.11 demonstrated that there is not a n/p effect on Carrig D hit rates nor false-alarm rates. However, the variability in Carrig D hit rates and false-alarm rates decreased with increasing sample size as was the case for MCD and PCLOW Also, notice that the false-alarm rates for the Carrig D are relatvely low in the three experimental conditions compared to its mean across all experimental conditions of 15.5%. At least under this very restricted set of conditions, communality higher; scenario I; $p = 16$; fraction of outliers equal to 0.16, we would recommend using the Carrig D as a detection metric over MCD and PCLOW due to the considerably lower false-alarm rate.

7.3 Performance of the Bacon MLD

It was readily apparent in Chapter 6 that the MLD had the highest hit rate across all of the experimental conditions. The high hit rates for the MLD across all levels of fraction of outliers at first led me to believe that the MLD was resistant to the masking effect. This was surprising given the fact that the MLD has the term $\overline{\text{MLCE}}_{.jk}$ in its formula (see Equation 2.29). This term is the mean of all the single observation maximum likelihood correlation estimates. Like the sample mean of a set of observations, it should be highly susceptible to a masking effect exerted by a relatively high proportion of outliers. A closer examination of the MLDs within

the 144 experimental conditions led me to the conclusion that the MLD was not resistant to a masking effect. Its high hit rates were due to the fact that in general the k-means algorithm divided the MLDs such that there were a large number of observations in the group with the higher mean. Recall that the group with the higher mean MLD constitutes the observations that are flagged as being potential outliers. Say the group with the higher mean MLD contains 40% of the observations, then there would be a relatively high probability that a true outlier would fall into this group even if the MLD did not measure degree of discordancy. Now for the analysis that led me to this conclusion.

From previous exploratory analyses, I noticed that the MLDs for a sample of observations tended to be symmetrically distributed. It makes sense that the k-means algorithm with the number of clusters set at two will divide a symmetric distribution of values into groups of equal size.

The Bacon MLD had the largest mean false-alarm rate in every experimental cell. Looking at the plots of the practically significant interaction effects on false-alarm rates involving the metric factor (Figures 6.8, 6.9, 6.10 and 6.11), we see that the mean false-alarm rate for the Bacon MLD is consistent across the experimental cells. It does vary much from the overall mean false-alarm rate of 0.35. I contend that the hypothesized mechanism responsible for the high mean hit rates is also responsible for the high false-alarm rates. I investigated whether there was evidence of the mechanism I proposed as being responsible for both the high hit rates and high false-alarm rates. Specifically, for each case study, a sample was generated according to the corresponding levels of the experimental condition. Then the distribution of the MLDs for each observation in the sample was examined. It was hoped that the distributions would shed light on whether the scores are somewhat symmetrically distributed (thus, the k-means algorithm divides the sample into two equally sized groups) and to what extent the MLD scores are measuring degree of discordancy.

I will now go over the reasoning used to select certain experimental conditions as case studies.

The range in MLD mean hit rates across the experimental conditions was larger than the range for the mean false alarm rates. The largest mean hit rate was 0.9, and it occurred in the higher communality, scenario I, $p = 16$, $n = 160$, fraction of outliers equal to 0.08 experimental cell. The false-alarm rate for this experimental condition was approximately 0.45 which is fairly close to the maximum false alarm rate of 0.52 The smallest hit rate was approximately 0.24, and it occurred in the communality high-same, scenario II, $p = 16$, $n = 320$, fraction of outliers equal to 0.32 experimental cell. The false-alarm rate for this condition was approximately 0.26 which is not that far away from the minimum false-alarm rate of 0.18. I chose the foregoing two conditions for study because they yielded extremes in performance in terms of hit rate and false-alarm rate. I also wanted to have as a case study an experimental cell that was "typical "in terms of both hit rate and false-alarm rate. I used a rather crude criterion for choosing a "typical "experimental cell. From a vector of the 144 MLD hit rate cell means, I formed a vector of the absolute values of the deviations of cell means from the grand mean. I did the same thing for the MLD false-alarm means. I then summed these two absolute-value-of-the-deviations vectors and found the element with the minimum value in this vector. The experimental condition that corresponded to this minimum value was the communality higher, scenario II, $p = 8$, $n = 160$, fraction of outliers equal to 0.16 condition. The hit rate for this condition was approximately 0.16. The false-alarm rate was 0.35.

I generated a random sample corresponding to each one of these three experimental conditions and did some exploratory data analysis of the resulting MLD-values. Recall that one of the reasons for using the k-means algorithm to select "outlying "values on the Bacon MLD and the Carrig D is that it presumably mimics the action of spotting the outliers from a plot, so it will be instructive to look at

a plot of the MLD scores for each of the three case study samples. The distributions for each of the three case studies will be displayed in the form of a stem-and-leaf plot. The leaves corresponding to contaminant observations will be distinguished by the color red.

I generated a random sample for the experimental condition corresponding to the case study one. Using the k-means algorithm in the application of the Bacon MLD, 68 of the 160 observations were flagged as possible outliers. Thus, 42.5% of the observations were flagged. Recall that in this experimental condition only 13 of the observations were contaminants. With the k-means method, the MLD flagged 10 of the 13 as possible outliers. However, the method also identified 58 of the 147 legitimate observations as possible outliers as well. The false-alarm rate is clearly unacceptable. Since 42.5% of the observations were identified as possible outliers it could be that the MLD metric does not provide much information about the discordancy of an observation: if the metric provides no information about discordancy, but yet you set aside almost fifty percent of the data, you are bound to catch some of the true outliers. A stem-and-leaf plot of the MLDs will help us infer from the sample if the MLD does provide some information about an observation's discordancy. Table 7.1 contains a stem-and-leaf plot of the MLDS. The leaves corresponding to the true outliers are colored red.

It does appear that observations from the contaminant distribution tend to have relatively high values on the MLD. Unfortunately, the high values do not distinguish themselves in the plot. If one were using the plot to spot outliers, they would probably conclude that there are no outliers and that the MLDs are approximately normally distributed. If one's motivation for spotting outliers is the influence they would exert on parameter estimation, a trimming approach would remove some of the outliers. However, as discussed previously, a trimming approach is not an acceptable practice. Note that the distibution of the MLD values in this

6	234
6	55778999
7	0122334
7	5566677778899
8	0000122222223333444444
8	55555666677778888888889999
9	000001122222222333444
9	555677777888888888999
10	0001111112333344
10	567777789999
11	022244
11	56
12	24
12	
13	
13	9
14	2

Table 7.1: MLDs for Case Study One. True Outliers Indicated by Yellow Leaves.

sample are symmetric. Thus, the k-means algorithm will tend to break up the values into two groups of roughly equal size, so this particular example lends credence to my hypothesized mechanism.

Again, as was the case in the Fisher Iris Data example, the reason why the MLD procedure yielded such a high false-alarm rate is that the k-means algorithm divides the data into essentially two equal halves. The k-means algorithm tends to do this when there are not two distinctive groups and the distibution is symmetric.

Let us take a look at the stem-and-leaf plots for the other two case studies. They are in Tables 7.2 and 7.3.

Looking at the stem-and-leaf plot for the second case study (Table 7.2), we see evidence of masking effect. There are a lot of the true outliers on the low end of the plot. The percentage of true outliers whose MLD is less than the median value of the entire sample in case 2 is approximately 44.7%. This is a strong indication of

7	
7	5
8	001234
8	5555667788899999999999
9	000001111111111112222222222222222233333333333333344444444444444
9	555555555555555555566666666666667777777777777778888888888999999999
10	000000000011111111111111122222222222223333333333344444444444444
10	555555555555566666677777888888888888999999
11	000000000001111122223344
11	557777999
12	0013344
12	
13	112
13	9
14	13
14	9
15	1

Table 7.2: MLDs for the Second Case Study. True Outliers indicated by Yellow Leaves.

7	4
7	5677888899
8	00123333334
8	57788889
9	00111111222222233344
9	55555556666667777788888899999
10	000000000001112222333333444
10	55556777778888888999999
11	011223334
11	677779
12	0344444
12	5578
13	03
13	9
14	2
14	6
15	
15	
16	0

Table 7.3: MLDs for Case Study 3. True Outliers indicated by Yellow Leaves.

a masking effect. The true number of outliers whose MLDs fell above the median is close to what would be expected if only chance were operative in assigning values to the MLDs. In other words, in this example, the MLD does not seem to measure an observation's degree of discordancy at all. It is not surprising that we saw a making effect in this case given that the fraction of outliers was at its highest level 0.32. Once again, none of the observations really distinguish themselves in the plot. Under this experimental condition, the graphical technique using MLDs utterly fails as a method for spotting outliers.

The level for fraction of outliers in case 3 was 0.16. Looking at the stem-and-leaf plot corresponding to this case, we see that there is not as strong of a masking effect at work. There is mild evidence that the MLD does somewhat measure the degree of discordancy. Within the top 10% of the values on the MLD, 43.75% correspond to true outliers. However, 38.5% of outliers fell below the median, so there is some degree of a masking effect at work.

In summary, the case studies suggest the presence of a masking effect increases with the fraction of outliers present. This is as one would expect given the computation of the MLD for each observation uses the sample mean of the single-observation maximum likelihood correlation estimates. Based on the exploratory analysis of the MLD scores for the three case studies, I hypothesize that the MLD had the highest hit rates and false-alarm rates because the k-means algorithm assigns a large proportion of the total observations to the potential outlier group. Furthermore, the MLD is not especially robust to the masking effect as I first suspected. I computed the mean proportion of observations assigned to group representing potential outliers for each experimental cell. A stem-and-leaf plot of these mean proportions is presented in Table 7.4 Note that the first quartile of the distribution is approximately 0.32. Thus, the k-means algorithm does assign a lot of the observations to the potential outlier group.

18	50
20	
22	3
24	3039
26	03801
28	679017
30	1280223333455889
32	00123346678990001244455566789
34	011233334401112377999
36	1234566777899011224566
38	012346791233346679
40	01517
42	68
44	45934
46	45
48	37
50	37
52	1

Table 7.4: Mean proportion of observations flagged as potential outliers by the MLD. One mean for each of the 144 experimental cells.

7.4 The Performance of the Carrig D

Like all metrics, with the exception of the MLD whose performance was consistent across all experimental conditions, the Carrig D performed best in terms of hit rate under the higher commuality condition. However, unlike the Bacon MLD, Carrig D had a significantly (practically significant) lower false-alarm rate under the higher and high-same conditions compared with the lower and low-same conditions (about 0.1 to 0.2 refer to Figure 6.8). This made me wonder if some of the true outliers distinguished themselves in plots of the Carrig D under the higher and high-same conditions. Two of the experimental cells I decided to generate random samples from were communality higher, scenario I, p=16, n=160, fraction of outliers = 0.08 and the experimental cell differing only from the first in the fraction of outliers: 0.32. I also wanted to investigate the robustness to the masking effect as well. I chose the communality higher and scenario I conditions because of the low mean false-alarm rate under that combination, approximately 0.03. The stem-and-leaf plot of Carrig Ds for a randomly generated sample coming from majority and contaminant populations specified by the factor levels in case study one is contained in Table 7.5 The stem-and-leaf plot of Carrig Ds for a randomly generated sample coming from majority and contaminant populations specified by the factor levels in case study two is contained in Table 7.6.

The stem-and-leaf plots imply that the Carrig D is measuring degree of discordancy as the contaminants tend to have higher scores on the metric. This is true even for the sample corresponding to the case 2 in which 32% of the observations were contaminants. Thus, we infer that the Carrig D is robust to the masking effect under the conditions of communality higher, scenario I, $p = 16$ and $n = 160$ at the very least. However, the decision of where to split the plots into groups of legitimate observations and potential outliers would largely be subjective. The contaminants do not clearly unequivically distinguish themselves in the plots. This is partly due

5	466788888999999
6	000111122222333333344444556677888899999999999
7	00111222333344445566666778889999
8	11112233333455566678888899
9	0017788
10	00001122334567
11	01122
12	03
13	1459
14	677
15	9
16	68
17	37
18	7
19	
20	
21	
22	
23	1
24	
25	4

Table 7.5: Carrig Ds for a random sample generated from the higher communality, scenario I, $p = 16$, $n = 160$, fraction of outliers = 0.08 condition. True outliers have red leaves.

6	57888889900011122222233333444445555666667777778899
8	000112222344445566788889999911124455689
10	011333577805588
12	1122447881113589999
14	29113789
16	
18	68
20	3646
22	0357775
24	02237799
26	3
28	78
30	3
32	0
34	51
36	1
38	
40	
42	
44	9

Table 7.6: Carrig Ds for communality higher, scenario I, $p = 16$, $n = 160$, fraction of outliers equal to 0.32. Contaminants have red leaves.

to the fact that we do not know anything about the distribution of the Carrig D for the legitimate observations. It is unknown whether it has a long, flat tail or not. Thus, the decision of where to divide the plots would largely be subjective.

I decided to generate a sample for a third case, one in which the Carrig D has a high false-alarm rate. The combination of communality lower and scenario II has a mean false-alarm rate of approximately 0.24. Specifically, I generated a sample from the communality lower, scenario II, $p = 16$, $n = 160$ and fraction of outliers equal to 0.16 experimental condition. Table 7.7 contains the stem-and-leaf plot of the Carrig Ds from just such a random sample.

The stem-and-leaf plot is very interesting. The contaminants do not tend to have high values on the Carrig D. The Carrig D utterly fails as a measure of discordancy under this experimental condition. Is this due to a masking effect or something else? I decided to investigate whether a random sample of MCDs from the same experimental condition would have a similar plot.

The stem-and-leaf of the MCDs for a sample drawn from the experimental condition (communality lower, scenario II, $p = 16$, $n = 160$, fraction of outliers $= 0.16$) in which the Carrig D performed so poorly is displayed in Table 7.8. Like the Carrig D, the MCD fails utterly when it comes to spotting outliers. Does it fail in the sense that it does not measure discordancy, or are the contaminants not discordant? The answer to this question will come when the principal component structureof the majority population correlation matrix is compared to the principal component structure of the contaminant population correlation matrix under the lower communality condition. The comparison of majority and contaminant principal component structures occurs in the next section, the discussion of the communality-metric interaction.

12	7
14	0001112255566789
16	3591267789
18	2345557899001123 5668899
20	12335900234455699
22	223589900149
24	00333556669990036899
26	344445134
28	9166
30	11811
32	037
34	50
36	75
38	
40	4804
42	145
44	9
46	
48	
50	
52	
54	2
56	
58	9
60	3

Table 7.7: Carrig Ds for a sample generated from the communality lower, scenario II, $p = 16$, $n = 160$, fraction of outliers equal to 0.16 experimental cell. Contaminants have red leaves.

0	
0	56677888889999
1	00001111111223333334444444444
1	55555555556666666677777777788888888888999999
2	00000011111111222222233333444
2	5567788
3	00111334
3	55555667889
4	1344
4	79
5	13
5	
6	
6	5

Table 7.8: Distribution of MCDs for a random sample drawn from the communality lower, scenario II, $p = 16$, $n = 160$ and fraction of outliers equal to 0.16 experimental condition. Contaminants are indicated by red leaves.

7.5 Communality-Metric Interaction

I explored the communality-metric interaction on hit rate for MCD and PCLOW first. My higher communality (majority population has a higher communality than the contaminant population) and lower commuality (majority population has a lower communality than the contaminant population) are similar to the two levels of a factor in Bacon's study. Recall that Bacon did not generate his data with a factor-analytic model as the basis. All of his data, both majority and contaminants, came from populations whose correlation matrices had all off-diagonal elements equal to one another. He had two higher (majority having higher correlations than contaminants) conditions: one in which all of the majority bivariate correlations were 0.9 and all of the bivariate contaminant correlations were 0.2. He had another higher condition where the all of the bivariate majority correlations were 0.7 and all of the bivariate contaminant correlations were zero. The two lower conditions can be obtained by switching the majority and contaminant correlations in the higher con-

ditions. The higher and lower levels of my communality factor are similar because, of course, the higher the communalities the higher the bivariate correlations.

Since the MCD distance can be thought of as a robust Mahalanobis D^2 we expect that the performance of the MCD under the higher and lower communality conditions will mirror the performance of the Mahalanobis D^2 under the higher and lower bivariate correlations conditions in Bacon's study. Recall that the fraction of outliers was a constant 0.10 across all of the experimental conditions in Bacon's study. Thus, the masking effect was not strong enough to make the Mahalanobis D^2's performance the same across the levels of the higher-lower factor.

Recall that in Bacon's study the Mahalanobis D^2 performed much better in the higher condition than in the lower condition. In my study the MCD and PCLOW performed markedly better in the higher communality condition than in the lower communality condition. Bacon used bivariate plots to illustrate why the Mahalanobis D^2 performed better in the higher condition. I thought it would be instructive to recreate these plots.

First let us consider the higher condition in which the correlation between the two variables is 0.9 in the majority population and the correlation between the two variables in the outlying population is 0.2. I generated samples of seventy-five observations from each of the populations. Observations from the majority population are indicated by blue characters, and observations from the outlying population are indicated by red characters. Now, imagine we draw a sample that has both majority and contaminant observations. Assume that the number of observations from the outlying population is less than 50%. The MCD is computed for each of the observations and the resulting values are compared to the 99.9th percentile of a χ^2_2 random variable. The observation is flagged as a potential outlier if its MCD exceeds the foregoing cut-off value. Assume that the robustly computed correlation matrix (i.e. the one outputted by the application of the MCD algorithm) used in

the computation of the MCDs is very close to the true correlation matrix for the majority population. If an observation exceeds the cut-off value, then the point corresponding to it will fall outside of the 99.9% constant density ellipse for the majority distribution. This constant density ellipse will roughly be an ellipse that encompasses all or very nearly all of the blue points in Figure 7.11, and it will do so fairly tightly. Notice that a good deal of the outlying observations indicated by the red characters will fall outside of this ellipse. Thus, realizations from the outlying population will tend to be spotted by the MCD.

Now consider the lower communality condition. We will do so by switching the correlations for the majority and contaminant populations (0.2 for the majority and 0.9 for the contaminants). Refer to Figure 7.12. The same line of reasoning as was used in the higher communality example is applied to this lower communality example. In order for the MCD to detect an outlier, it must fall outside the 99.9% constant density ellipse. However, notice that the constant density ellipse in this case will be larger in terms of area than the one from the previous example. Also, note that it appears that very few of the contaminant observations will fall outside of this constant density ellipse. Thus, in this case, the MCD will not tend to flag observations that are realizations from the outlying population.

The same line of reasoning can be applied to distributions with a higher number of variables. If the number of variables is greater than three, then instead of constant density ellipses, as in our example, we have constant density hyperellipsoids.

Refer to Figures 6.4, 6.5, 6.2 and 6.3. These are the plots corresponding to all of the practically significant effects involving communality. We do indeed see that both MCD and PCLOW consistently performed better in the higher communality condition as opposed to the lower communality condition. Now let us consider the two conditions, high-same communality and low-same communality, that did not

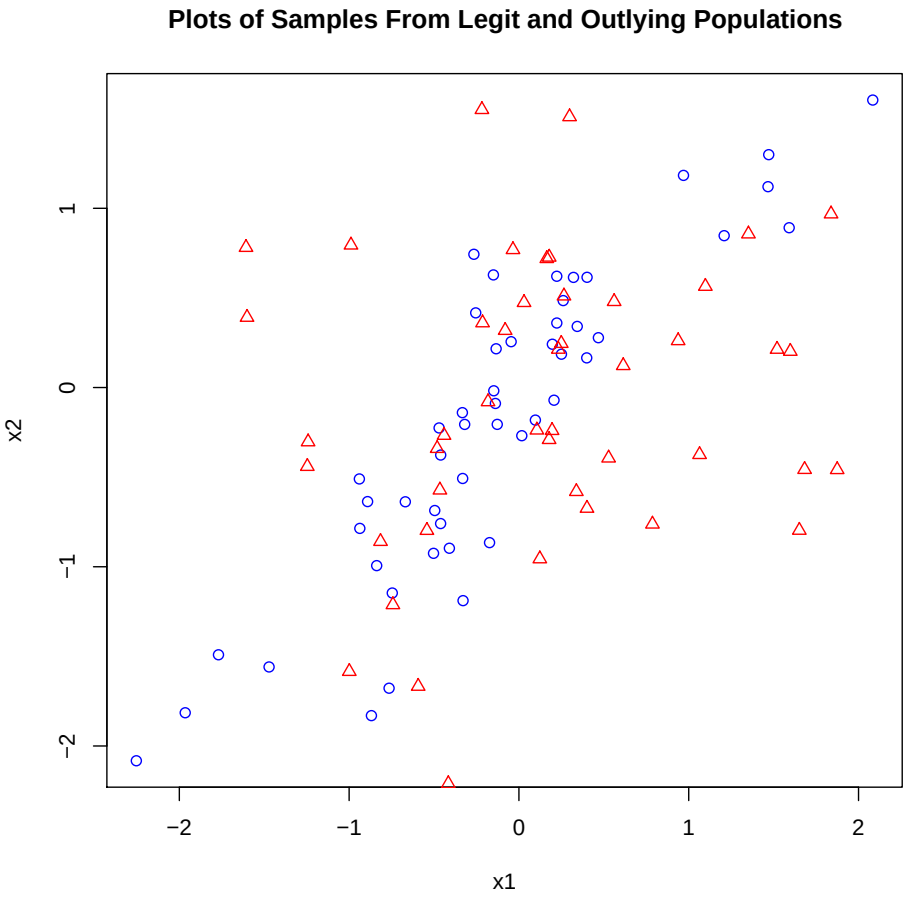

Figure 7.11: Samples of seventy-five observations for both the majority population and the contaminant population. Observations from the majority population are indicated by the blue characters those from the contaminant by red characters. Correlation for the majority is 0.9 and is 0.2 for the contaminants.

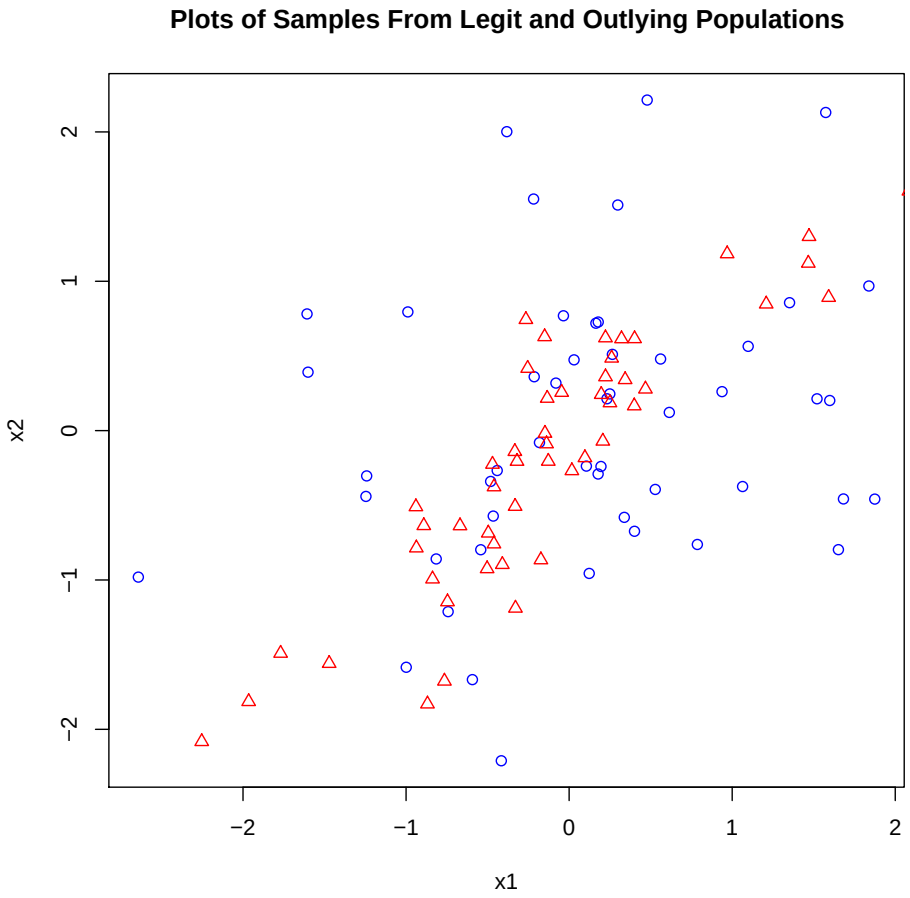

Figure 7.12: Samples of seventy-five observations for both the majority population and the contaminant population. Observations from the majority population are indicated by the blue characters those from the contaminant by the red characters. Correlation for the majority is 0.2 and is 0.9 for the contaminants.

have counterparts in Bacon's study. From the interaction plots of which communality was a part we see that high-same and low-same consistently had mean hit rates less than those for the higher condition but larger mean hit rates than those for lower. We also see that consistently there is not a substantial difference in mean hit rates between high-same and low-same. This last point puzzled me because if I apply the same reasoning employed to explicate the difference in hit rates between the lower and higher communality conditions I come to the conclusion that MCD should have substantially higher hit rates under the high-same condition as opposed to the low-same condition. Let me illustrate with a plot. Consider the bivariate case where the majority population has a correlation of 0.8 and the contaminant population has a correlation coefficient of -0.8. I generated two samples of seventy-five observations, one from the majority population and one from the contaminant population. The plot of the samples is contained in Figure 7.13 Remember that under the high-same condition in my main experiment the majority observations and the contaminants had two different underlying factor models. Therefore their principal component structure was different. This is why I made the majority and contaminant distributions that generated the data in Figure 7.13 have different principal component structures so has to mirror the conditions in the main experiment.

It looks like the contaminant observations will tend to fall outside of the critical constant density ellipse. Next, I generated a plot to reflect the low-same commuality condition. This plot is in Figure 7.14. The correlation for the legitimate population is 0.2 and -0.2 for the contaminant. Thus, the principal component structures are different but both reflect a situation in which both of the communlities are low. Based on this plot, it looks as though the contaminants will not tend to fall outside the critical constant density ellipse.

It should be mentioned that in the bivariate plot exemplifying the high-same condition (Figure 7.13 the high-variance principal components, one from the ma-

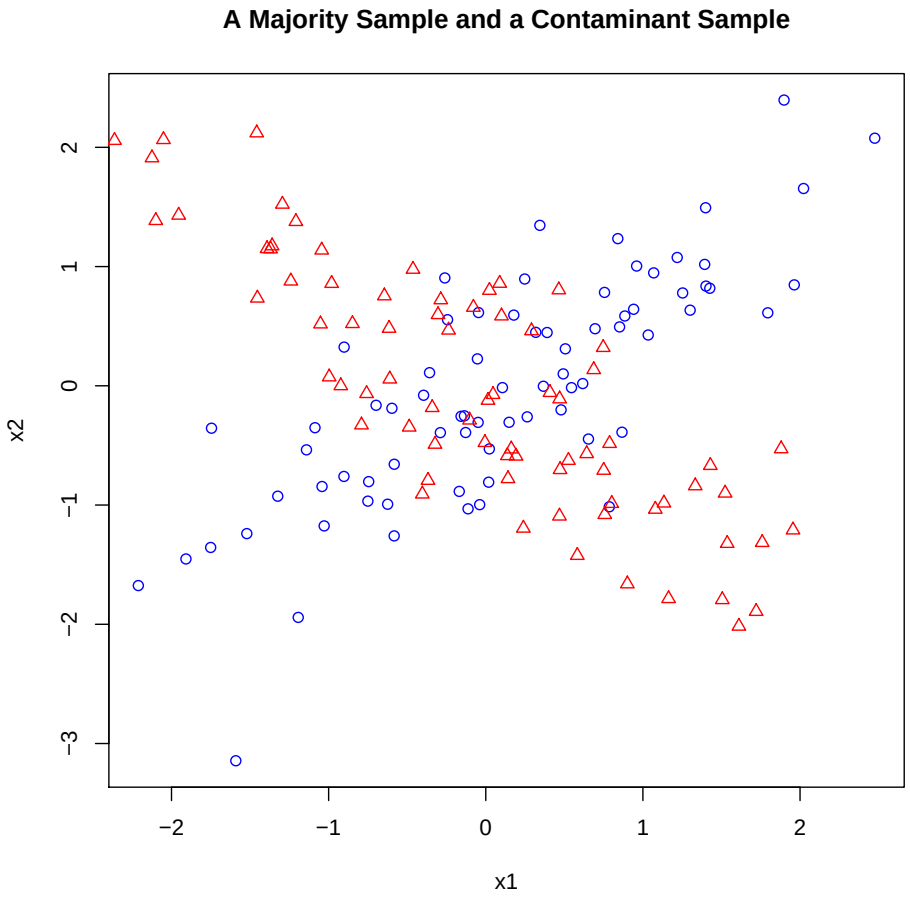

Figure 7.13: Samples of sevent-five observations for both the majority population and the contaminant population. Observations from the majority population are indicated by blue characters, contaminants by red characters. Correlation for the majority is 0.8. Correlation for the contaminants is -0.8

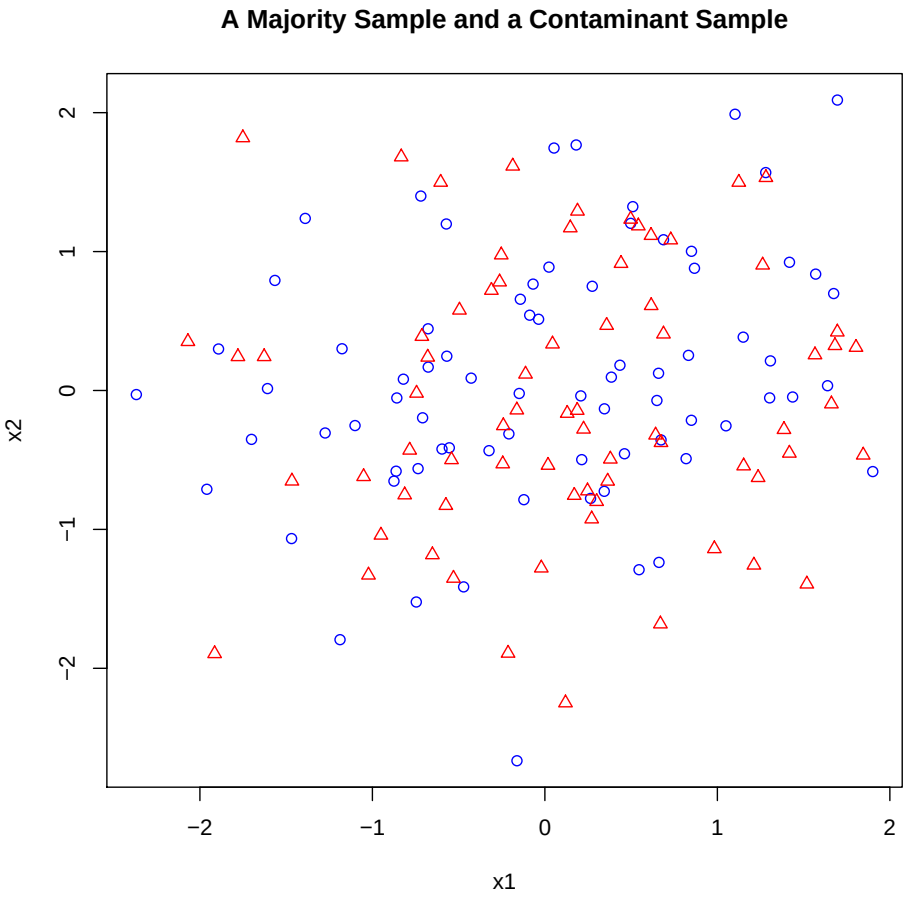

Figure 7.14: Samples of seventy-five observations for both the majority population and the contaminant population. Observations from the majority population are indicated by blue characters, contaminants by red characters. Correlation for the majority is 0.2 and -0.2 for the contaminants.

jority population and one from the contaminant population, are orthogonal. If the principal components were coincident, the majority and contaminant observations would come from the same population. As we move from orthogonality to coincidence the performance of the MCD will decline [1]. Could it be that in my high-same conditions the space spanned by the the high principal components for the majority populations are far from being orthogonal to the spaces spanned by the high principal components for the contaminant populations?

Recall that the factor-analytic models underlying the majority and contaminant populations differ across the two levels of the sceanrio factor, so the principal component structures for the majority and contaminant populations will differ across the two levels of this factor. Therefore, I looked at the principal component structures under the high-same communality for both levels of scenario in order to investigate the surprising under performance of the MCD under the high-same communality level. Specifically, I looked at the at the principal component structures with $p = 16$ for communality high same under both scenario I and scenario II. Initially, I decided to compare principal component structures by looking at the angles between the high principal components from the majority population and those from the contaminant population. I also looked at the variances for the high principal components.

For the the high-same communality, $p = 16$, scenario I conditions, I extracted principal components from the majority and contaminant correlation matrices implied by their respective factor-analytic models. Refer to Equations 4.7 and 4.28 for the majority and contaminant correlation matrices respectively. The majority population has one principal component with a variance of 10.60. The fifteen remaining

[1] In the bivariate case, the eigenvectors for a correlation matrix will always be the same $\mathbf{e}_1 = (1/\sqrt{2}, 1/\sqrt{2})$ and $\mathbf{e}_2 = (1/\sqrt{2}, -1/\sqrt{2})$, so in the two-dimensional case we technically cannot move through degrees of non-orthogonality to coincidence. However, it is possible for the constant density ellipsoids in three-dimensional space to assume intermediate states between orthogonality and coincidence.

principal components each had variances of 0.36. The contaminant population had two principal components each with variances equal to 5.48. The fourteen remaining principal components all had variances equal to 0.36. Clearly, the number of high principal components for the majority population is one, and the number of high principal components for the contaminant population is two. The angles between the lone high principal component for the majority population and the two high components for the contaminant population are each 45° Because the angles are 45° one may be led to believe that the orientation of the two spaces spanned by high principal components is intermediate between orthogonality and coincidence. However, under this combination of communality, scenario and p, we can actually compute the angle between the two spaces. The space spanned by the majority high principal component is a line, and the space spanned by the contaminant high principal components is a plane.

The angle between the two spaces is the angle between the eigenvector corresponding to the high majority principal component and its orthogonal projection into the plane spanned by the two eigenvectors corresponding to the two high contaminant principal components. I followed this procedure and found that the eigenvector corresponding to the high majority component lies in the plane spanned by the eigenvectors corresponding to the high contaminant principal components. The angle between the two spaces is thus zero. Remember that aforementioned angles between singular components were 45°. This example shows that when one is dealing with more than one dimension the whole is not the sum of the parts.

Even though the space spanned by the high majority component lies within the space spanned by the high contaminant components MCD will catch a fair amount of contaminant because the dimensionality of the high contaminant space is two and the dimensionality of the high majority space is one. Realizations of the majority populations will cluster very closely to the line corresponding to the high

component. Observations from the contaminant distribution have a high variance in two orthogonal directions. Thus some realizations from the contaminant distribution will fall outside the constant density hyperellipsoid that tightly encloses the lone principal component for the majority distribution. But, why did the MCD perform better in the higher communality condition under scenario I compared to the high same condition? The reason why the MCD performed better in the higher communality condition as opposed to the high-same condition is that while the constant density hyperellipsoids for the contaminant populations in both conditions have the same orientation they will have larger volumes in the higher condition. The fact that the constant density hyperellipsoids have the same orientation can be shown by looking at the eigenstructure of the respective correlation matrices in terms of the parameters for the factor-analytic model. In other words $\boldsymbol{\rho}$ in its equivalent form $\mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$ where $\mathbf{L}$ is the loading matrix and $\boldsymbol{\Psi}$ is the diagonal matrix of specific variances. Recall from Section 7.2 that the volume of a constant density hyperellipsoid is proportional to the determinant of the correlation matrix. Now, it can be shown that the determinant of a correlation matrix is the product of its eigenvalues by applying properties of determinants to the spectral decomposition of the correlation matrix.

Constant density hyperellipsoids for contaminant populations from the higher communality condition will cover a greater volume than the corresponding constant density hyperellipsoids for the majority populations. Say we set the probability at 0.999. Consider a point within the larger hyperellipsoid but outside of the smaller hyperellipsoid. The probability that a point with a greater statistical distance than the point specified in the preceding sentence will occur is less than 0.001 for the population with the smaller volume hyperellipsoid but it is as least as large as 0.001 for the population with the higher volume hyperellipsoid. Thus, the probability of points being a greater statistical distance away from the centroid common to all of

the populations is greater for the population with the constant density hyperellipsoid with the larger volume. This would explain why the MCD and PCLOW do not perform as well under the high-same conditions in comparison to the higher condition with scenario equal to I. Under the high same condtion, the constant density hyperellipsoid for the contaminant population will have a smaller volume than it would be under the higher condition because the product of the eigenvalues will be smaller. In our $p = 16$ example the eigenvalues for the contaminant correlation matrix under the high same condition are 5.48 with multiplicity two and 0.36 with multiplicity fourteen. Thus, the value of the determinant is $(5.48)^2(0.36)^{14} = 1.84 \times 10^{-5}$. Now for the contaminant correlation matrix in the higher condition. The eigenvalues for this correlation matrix are 2.12 with multiplicity two and 0.84 with multiplicity fourteen. The determinant of this correlation matrix is $(2.12)^2(0.84)^{14} = 0.391$. Now let us consider high-same communality under scenario II.

As in the case study under scenario I, I set p at sixteen. Refer to Equations 4.16 and 4.28 to see the majority and contaminant correlation matrices implied by their respective factor-analytic models under the factor combination being explored here. The majority population has four principal components each having a variance of 2.92. The remaining twelve principal components have variances of 0.36. The contaminant population has two principal components each with variances of 5.48. The fourteen remaining components each have variances of 0.36. Thus, we have four high components for the majority population and two for the contaminant population. The space spanned by the majority high components is a four-dimensional hyperplane in sixteen-dimensional space, and the space spanned by the high contaminant components is a two-dimensional plane in sixteen-dimensional space. I do not know how to quantify the difference in orientations of these two spaces since one is four dimensional. Let us look at the hit rates under the factor combination in question.

The mean hit rate for communality high same, scenario I, $p = 16$ is approximately 0.29. In contrast, the mean hit rate for communality high same, scenario II, $p = 16$ is approximately 0.11. This leads me to believe that the space spanned by high contaminant components lies within the space spanned by the majority high components. We can test this by seeing if the eigenvectors corresponding to the two high components for the contaminants lie within the hyperplane spanned by the four eigenvectors corresponding to the four high majority components. I can answer this question by finding the orthogonal projection of two contaminant eigenvectors into the space spanned by the four majority eigenvectors. If the projection is identical to the original vector, then, the eigenvector lies in the space. Recall the discussion of orthogonal projections is Section 2.4.1. The equation for the orthogonal projection is

$$\mathbf{x}_{\text{proj}} = \mathbf{x}'\mathbf{e}_1\mathbf{e}_1 + \mathbf{x}'\mathbf{e}_2\mathbf{e}_2 + \mathbf{x}'\mathbf{e}_3\mathbf{e}_3 + \mathbf{x}'\mathbf{e}_4\mathbf{e}_4 \tag{7.2}$$

where $\mathbf{x}$ is one the high contaminant eigenvectors. I found the orthogonal projections of the two high contaminant eigenvectors using Equation 7.2. As I suspected each of the projections was identical to the original vectors, so the space spanned by the high contaminant components lies within the space spanned by the high majority components. I computed the determinants for the majority correlation matrix and the contaminant correlation matrix. They were 3.44×10^{-4} and 1.84×10^{-5} respectively. The constant density hyperellipsoids for the contamiant population are smaller. Combining this with the fact the space spanned by the high components for the contaminant population is within the space spanned by the high components for the majority population, I was led to conclude that realizations from the contaminant population will tend to lie within the 99.9% constant density hyperellipsoid for the majority population and thus will not be detected by the MCD. It can also be said that since realizations from the contaminant populations in these cases do not tend to lie outside of 99.9% constant density hyperllipsoids they are not outlying

with respect to the majority population.

This fact will help us reinterpret a earlier finding with regard to the Carrig D (see Section 7.4). The stem-and-leaf plot in Table 7.7 at first glance might seem to indicate that the Carrig D fails as a measure of multivariate discordancy under the communality lower and Scenario II factor combination. I do not think the Carrig D failed so much as a measure of multivariate discordancy as the observations from the contaminant distribution were not outlying with respect to the majority distribution. The same goes for the poor results exhibited by the MCD under the communality lower, scenario II factor combination. (see Table 7.8).

7.6 The Poor Performance of PCHIGH

As mentioned before, it can be readily discerned from the plots of all of the pratically significant effects on hit rate (Figures 6.2, 6.3, 6.4, 6.5, 6.6 and 6.7)that PCHIGH performed miserably across all experimental conditions. Also, there was almost no variability in its performance across the experimental conditions. The only way PCHIGH would catch these purely correlational outliers was if observations from the contaminant distributions were highly variable in the directions defined by the high variance principal components from the majority population. Now if the contaminant observations were highly variable in these directions, then the covariance matrices would be "very similar. "Since the covariance matrices for the majority and the contaminant populations were not "very similar "under any of the experimental conditions, PCHIGH did not identify the contaminants very well. This is especially true given one of the restrictions in my study. In all conditions the covariance matrices for the majority and contaminant distributions were correlation matrices. Because of this the total variance, defined as the sum of the diagonal elements of a covariance matrix, was fixed at p. I used the special cases of correlation matrices in order to avoid a confound with the effects of variance inflation. So, I contend that

under the restrictions of this study it is a logical deduction that PCHIGH will not spot correlational outliers.

7.7 PCLOW and MCD

PCLOW and MCD clearly outperformed PCHIGH in the context of the entire experiment. PCLOW and MCD did perform almost as poorly as PCHIGH in the lower communality condition. However, we saw in the preceding section that under the lower communality, scenario II combination realizations from the contaminant distribution are not often outlying with respect to the majority population. I do not think that there will ever be a metric that will detect contaminants such as these with great regularity.

The more interesting result to me is the fact that PCLOW did not strongly distinguish itself in terms of higher mean hit rates from the MCD under certain experimental conditions. Consider the higher communality, scenario II, $p = 8$ combination. The correlation matrices for the majority and contaminant populations are Equations, 4.12 and 4.5. The majority correlation matrix has four principal components each with a variance of 1.64 and four components each with a variances of 0.36. The contaminant correlation matrix has two principal components each with variances of 1.48 and six components each with variances of 0.84. It can be shown that the two components with variances of 1.48 (the high contaminant components) along with two of the components with variances of 0.84 from the contaminant population span the same four-dimensional hyperplane as the four components from the majority population having variances of 1.64 (the high majority components). The four components from the majority population mentioned in the preceding sentence will be designated as the high majority components and the four contaminant components as the first four contaminant components. However, none of the four aforementioned contaminant components (eigenvectors) are equal to any of the four

aforementioned majority components (eigenvectors); they just span the same space. Now the variances of the four components from the contaminant population that span said space are smaller than the variances for the four high components from the majority population that span the same space. [2]. This means that the contaminant observations will tend not to have large statistical distances in this space, and the statistical distance of the projection into this space is PCHIGH. PCHIGH is going to pull the total sum that is MCD down so to speak.

Now consider the space spanned by the four low principal components (the ones that are left after the high components are taken away from the set of eight) from the majority population. This space, a four-dimensional hyperplane, has to be the same as the one spanned by the last four components from the contaminant population. The variance of the contaminants in this space is greater than the variance of the majority observations in this space. Therefore some of the projections of the contaminants into the space will have large statistical distances that are based on the variability of the majority observations in this space. The statistical distance is PCLOW. Therefore some of the contaminants should have unusually large PCLOW values and will, thus, be flagged as potential outliers.

Because the PCHIGH part of the sum will in general be small, it seems that PCLOW by itself would have more power than MCD. Therefore, I am surprised by the fact that the difference in mean hit rates for PCLOW and MCD under this condition was not greater. Recall that under this factor combination (communality higher, scenario II, $p = 8$), the mean hit rate for PCLOW was approximately 0.30 and was approximately 0.25 for MCD.

[2]Variability in this space can be quantified in terms of a single number for each of the populations by summing the appropriate eigenvalues or taking the product of the eigenvalues. Either way the variance of the contaminants is smaller in the space.

7.8 Unanswered Questions

7.9 Limitations, Future Research and Recommendations

An obvious limitation of this study is that it dealt with correlational outliers only. Therefore, no recommendations can be made about multivariate outlier detection in general.

One of the main questions that I had, whether PCLOW offers much over and above the MCD, is still largely unanswered due to the fact that only two differences in the orientations of majority and contaminant correlation matrices were investigated, scenario I and scenario II. Further research in which the differences in orientations are more varied are in order before even tentative conclusions can be reached in regard to MCD versus PCLOW.

Also, the question as to whether the Carrig D can offer anything over and above the distance metrics is hardly settled either. My study did provide some indication that using plots of sample Carrig Ds may not be the best implementation of this metric. However, research into the distributional properties of this metric could prove fruitful. Perhaps a statistical criterion based on distributional or approximate distributional properties would lead to lower false-alarm rates than were observed in this study along with a hit rate superior to the distance metrics.

There is a parallel to my question of whether PCLOW offers anything over and above MCD in spotting correlational outliers and that is whether PCHIGH offers anything over and above MCD in spotting multivariate outliers be they mean-shift or combinations of mean-shift and correlational.

A major area for future research would be an investigation into the effects of non-normality on the various metrics. The statistical criterions used by MCD, PCHIGH and PCLOW rely on the assumption of multivariate normality. How robust would these three procedures be to violation of this assumption especially in

comparison to the Carrig D?

Based on the results of my study, the single metric that I would recommend for detecting outliers (or contaminants) is MCD. I do so for a number of reasons. It had a low overall false-alarm rate, approximately 7%, compared to the Carrig D and the Bacon MLD, approximately 16% and 35%, respectively. Also, in the case of the MCD, a cut-off value informed by an approximate distribution of the individual MCDs can be used in contrast to the Carrig D and the Bacon MLD which have unknown distributional properties. The MCD is recommended over the Mahalanobis D^2 because of the Mahalanobis D^2's susceptibility to the masking effect at even low levels of contamination. The Mahalanobis D^2 had a hit rate of 2% under 8% contamination compared to 17% for MCD. MCD is recommended over PCLOW for the reason that the MCD is a robust version of the Mahalanobis D^2. Presumably, users of multivariate methods are already familiar with the Mahalanobis D^2 and the use of a χ^2_p cut-off value with it. Thus, there would be a level of understanding and comfort in using the MCD over PCLOW which demands the approximating hyperplane interpretation of principal components for its comprehension. These recommendations are limited to data that are approximately multivariate normal. Subsequent research into the distributional properties of the Bacon MLD and Carrig D could lead to advocation of their use. Also, further research must be done before PCLOW and PCHIGH can be completely excluded from a researcher's arsenal for spotting multivariate outliers. This is especially true for PCHIGH given my hypothesis of its utility in spotting certain kinds of mean-shift outliers.

Appendix A

Bacon MLD Example

I have reprinted the equations for the Bacon MLD and its uncertainty term as Equations A.1 and A.2, respectively.

$$MLD_i = \frac{2}{p(p-1)} \sum_{j=1}^{p-1} \sum_{k=j+1}^{p} \frac{\left| \overline{MLCE}_{.jk} - MLCE_{ijk} \right|}{U_{ijk}}. \tag{A.1}$$

$$U_{ijk} = \sqrt{\frac{\sum_{q=1}^{m} L(p_{qjk})(p_q - MLCE_{ijk})^2}{\sum_{q=1}^{m} L(p_{qjk})}} \tag{A.2}$$

A data set of IQ scores on twins separated at birth is used to illustrate the computational aspects of the Bacon MLD. Each row in Table A.1 contains an IQ score for the child that was placed in a foster family and the score of her twin who resided with her biological family. I have added a spurious observation, number twenty-eight. It is outlying with respect to the correlational structure of the rest of the data. Refer to Figure **??**. The spurious observation is indicated by the red triangle. Now, let us delve into the details of computing the MLCE_{ijk} term.

At first glance, it may seem that the estimate of a population correlation from one observation may be impossible. In reality the computation of MLCE_{ijk} uses some information from all of the observations because an observation's z-scores

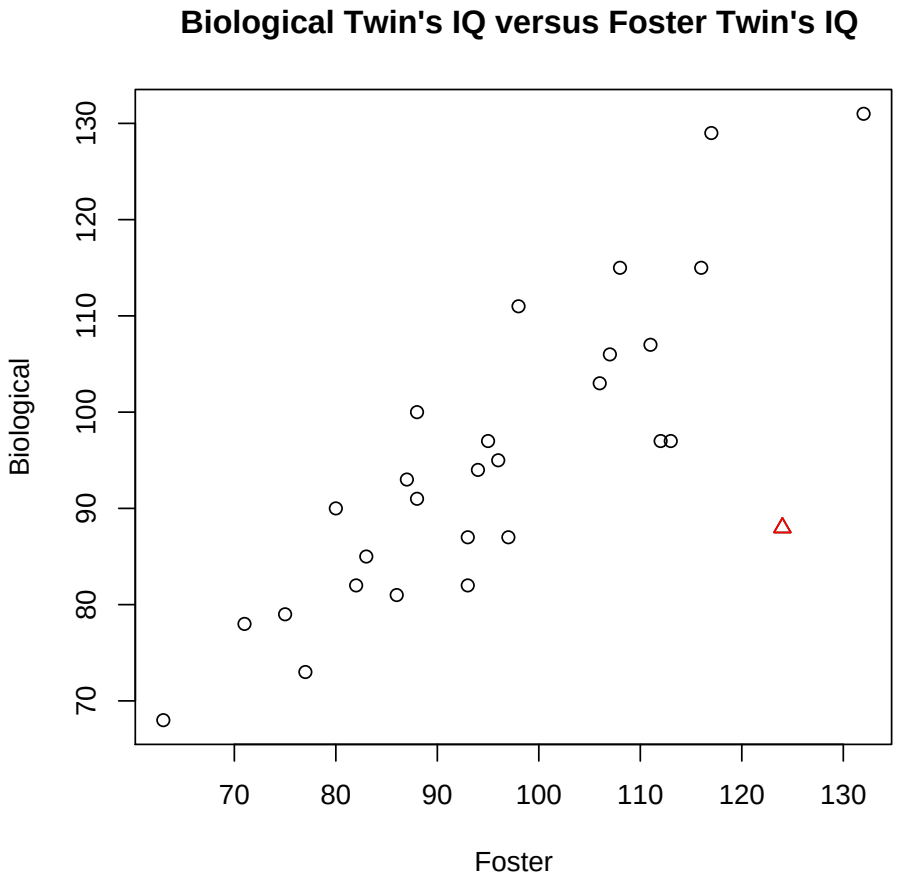

Figure A.1: Scatterplot of the twin data with the spurious observation, number twenty-eight indicated by the red triangle.

are used in the bivariate density. Specifically, the bivariate density is

$$f(z_1, z_2) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} e^{\frac{-(z_1^2 - 2\rho z_1 z_2 + z_2^2)}{2(1-\rho^2)}} \tag{A.3}$$

The variables z_1 and z_2 represent the z-scores for the foster and biological twins, respectively. Note that both σ_1 and σ_2 are assumed to equal one. Consider observation twenty-eight, the spurious data point. Its z-scores are 1.67 and -0.43 for foster and biological IQs, respectively. The $\text{MLCE}_{(28,\ 1,\ 2)}$ is the maximum of the function A.3 with the values 1.67 and -0.43 plugged into z_1 and z_2, respectively. Note that since the two $z - scores$ have been plugged into the expression the result is a function of one variable, ρ. Bacon considers this function to be a likelihood:

$$\text{L}(\rho) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} e^{\frac{-\left(1.67)^2 - 2(1.67)(-0.43)\rho + (-0.43)^2\right)}{2(1-\rho^2)}} \tag{A.4}$$

The graph of this function is depicted in Figure **??**. We see that the likelihood function has a maximum. I used R's `optimise` function to find the maximum for the likelihood function. The maximum likelihood resulted from a correlation value of -0.406 which is thus the result for the MLCE. This value is $\text{MLCE}_{(28,\ ,1,\ 2)}$. Now, let us compute the uncertainty for observation twenty-eight.

I have written some pseudo-code for the computation of the uncertainty, U_{ijk}. Refer to equation A.2 The only values that need to be provided to the algorithm are the z-scores, z_1 and z_2, and the MLCE. I approximated Bacon's use of 250 equally spaced points between -1 and 1 bystarting at -0.999 moving by increments of 0.008 to approximately 0.999, the 250th point. The index `p` in the code cycles through these 250 values. The variables `numerator` and `denominator` refer to the numerator and denominator of the expression under the square root sign in Equation 2.30 Since the terms that are in the numerator and denominator are sums each of them is computed within a loop. The values of `numerator` and `denominator` will contain

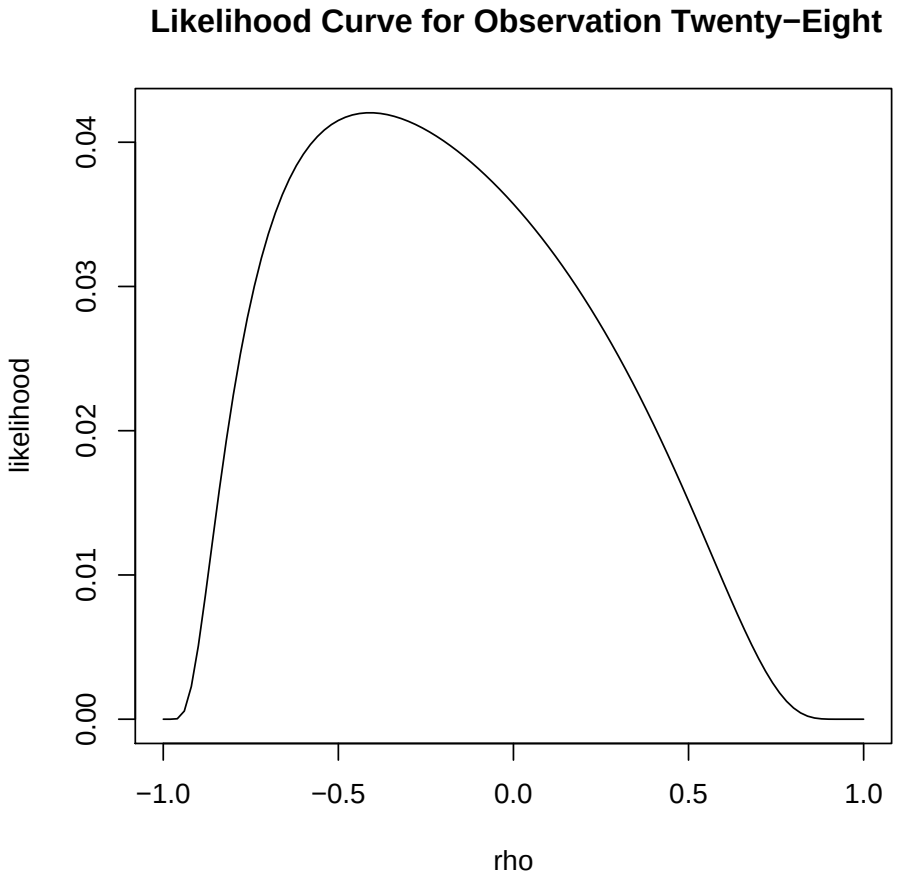

Figure A.2: The "likelihood "function for observation twenty-eight.

their corresponding sums after the exit is made from the loop. The pseudo-code is provided below in typewriter type.

```
%%%%%%%%%%%%%%%COMPUTATION OF THE UNCERTAINTY%%%%%%%%%%%%%%%

numerator=0

denominator=0

FOR i = 0 to 249 {

   p = -0.999 + i*(0.08)

   likelihood_of_p = likelihood function evaluated at z1, z2 and p

   squared_deviation = (p - mlce)^2

   product = likelihood_of_p * squared_deviation

   numerator = numerator + product

   denominator = denominator + likelihood_of_p

} %%%%%END OF LOOP%%%%%%%

quotient = numerator / denominator
```

```
uncertainty = sqrt(quotient)

%%%%%%%%%%%%%%%%%%%%%%END OF ALGORITHM%%%%%%%%%%%%%%%%%%%%%%%%%%
```

I wrote a program in R corresponding to this algorithm and computed the uncertainty for each of the observations. The uncertainty values are contained in the last column of Table A.1. We see that the uncertainty for the twenty-eighth observation is 0.446.

The term, $\overline{\text{MLCE}}_{\cdot jk}$, is just the mean of the twenty-eight individual MLCEs. This mean is approximately 0.570.

Because we have only two variables in this example the Bacon MLD is simply:

$$\text{MLD}_i = \frac{\left|\text{MLCE}_{i12} - \overline{\text{MLCE}}_{\cdot 12}\right|}{\text{U}_{i12}}. \tag{A.5}$$

. Thus, the Bacon MLD for the twenty-eighth observation is simply $|-0.406 - 0.570|/0.446 = 2.19$. If we were computing the MLD for an observation from a sample on four variables for instance, there would be six terms in the sum that is the MLD (one for each possible pairing among the four variables).

Let us compare two observations that had similar MLCEs but different uncertainties. Two such observations are observation seven and twenty-five with MLCEs of 0.986 and 0.997, respectively. The z-scores for observation seven are 2.15 and 2.32 and are 0.59 and 0.51 for observation twenty-five. Observation seven provides more compelling evidence that the population correlation coefficient is close to one because a realization of two z-scores of these magnitudes would be unlikely if the population correlation coefficient were really small or negative. Contrastingly, we cannot say the same thing about observation twenty-five. While the z-scores are close in both magnitude and sign, which leads to an MLCE of close to one, they could have conceivably come from a population with a low positive correlation. Look

at the likelihood curves for each of the observations (Figures A.3 and A.4).

If we loosely think of these curves as densities we see that the one corresponding to observation twenty-five has a larger variance because of the fatter tail. This is because smaller correlation coefficients are more "likely "for observation twenty-five than for observation seven. Thus, the second moment, or its square root, is a measure of uncertainty.

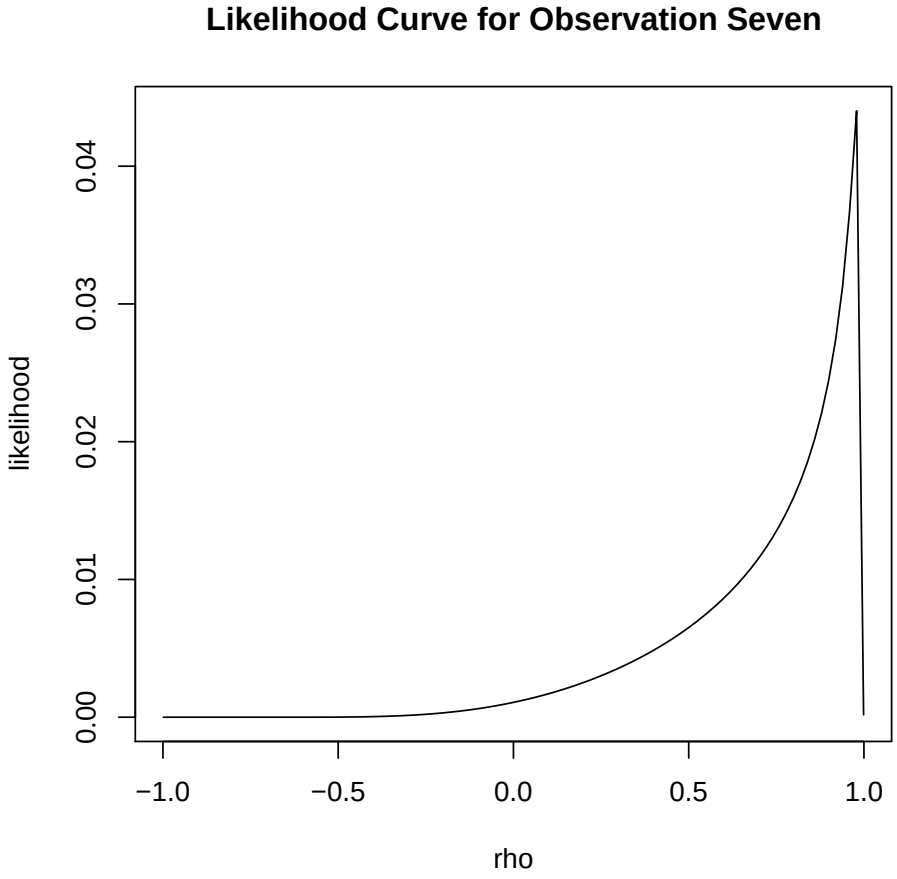

Figure A.3: Likelihood Curve for Observation Seven

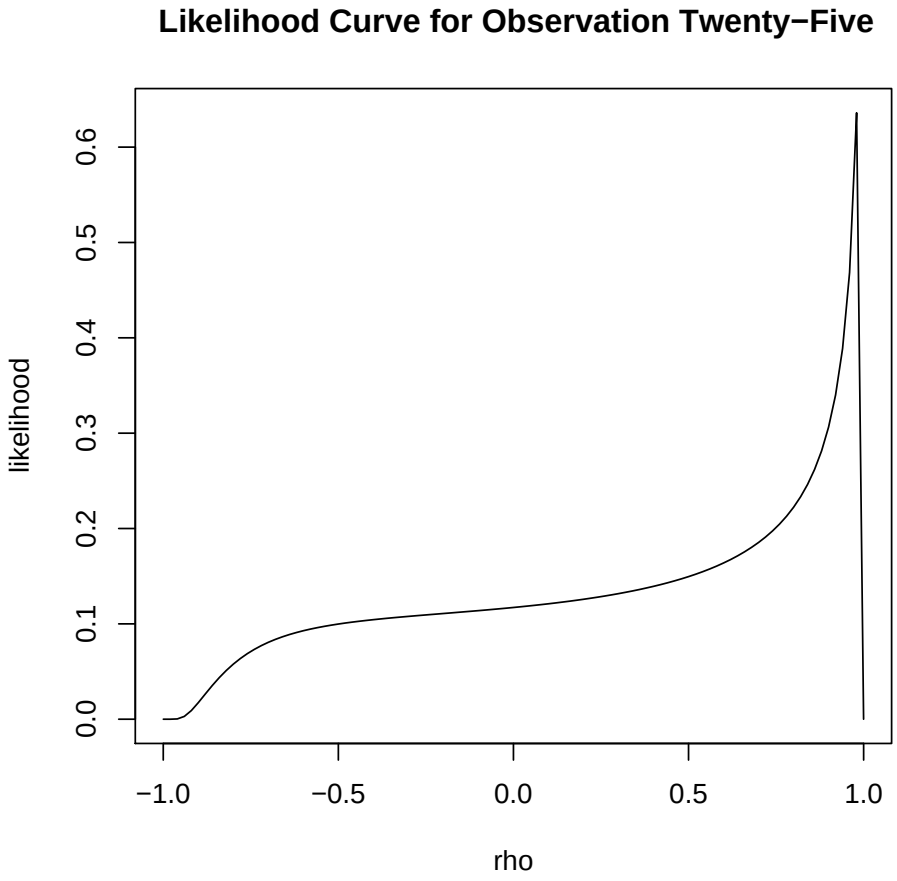

Figure A.4: Likelihood Curve for Observation Twenty-Five

Observation	Foster IQ	Biological IQ	z_1	z_2	MLCE	U
1	82	82	-0.85	-0.85	0.999	0.731
2	80	90	-0.97	-0.32	0.773	0.776
3	88	91	-0.49	-0.26	0.973	0.993
4	108	115	0.71	1.29	0.833	0.656
5	116	115	1.19	1.29	0.995	0.589
6	117	129	1.25	2.19	0.695	0.403
7	132	131	2.15	2.32	0.986	0.360
8	71	78	-1.51	-1.10	0.919	0.580
9	75	79	-1.27	-1.03	0.971	0.633
10	93	82	-0.19	-0.84	0.767	0.837
11	95	97	-0.07	0.13	-0.998	1.154
12	88	100	-0.49	0.32	-0.982	0.972
13	111	107	0.89	0.77	0.993	0.747
14	63	68	-1.98	-1.78	0.972	0.437
15	77	73	-1.15	-1.42	0.964	0.589
16	86	81	-0.61	-0.91	0.955	0.790
17	83	85	-0.79	-0.65	0.990	0.799
18	93	87	-0.19	-0.52	0.944	1.005
19	97	87	0.05	-0.52	-0.883	1.29
20	87	93	-0.55	-0.13	0.908	1.004
21	94	94	-0.13	-0.13	0.998	1.165
22	96	95	0.00	-0.01	-0.999	1.182
23	112	97	0.95	0.13	0.530	0.669
24	113	97	1.01	0.13	0.530	0.669
25	106	103	0.59	0.51	0.997	0.879
26	107	106	0.65	0.71	0.998	0.810
27	98	111	0.11	1.03	0.470	0.635
28	124	88	1.67	-0.43	-0.406	0.446

Table A.1: Twin IQ Data with Intermediate Values Leading to the MLD.

Appendix B

Comparison of a Parent Distribution with an Order Statistic from it.

Consider the following experiment. We query a random number generator for five random numbers between zero and one. We have five random variables $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4, \mathbf{X}_5$ that are independently and identically distributed. Specifically,

$$\mathbf{X}_i \sim U(0,1)$$

In other words, each are independent uniform zero-one random variables. Now the probability that $\mathbf{X}_i$ exceeds 0.5 is equal to 0.5 for each i where i runs from one to five.

Now, consider the experiment where we query the random number generator for five uniform zero-one random variables and we observe the one with the largest value. Obviously, the largest value will vary randomly from experiment to experiment and it will have a distribution. However, it will not have the same distribution

as the individual $\mathbf{X}_i$s.

Now the largest of the five will exceed 0.5 *if and only if* at least one of the five exceed 0.5. The probability of the latter can be computed using a binomial probability law in which success is defined as an observation exceeding 0.5 with five trials because of the fact that each of the realizations is independent. Therefore the probability in question is

$$\sum_{j=1}^{5} \frac{5!}{j!(5-j)!}(0.5)^j(0.5)^{5-j}$$

This probability is 0.96875. This is considerably larger than the probability of one of the individual random variables exceeding 0.5 (Larson, 1982). Thus, we conclude that the distribution of the largest of the five random variables is distributed differently than any of the individual random variables (Larson, 1982). According to Larson this difference in distributions "proved somewhat confusing to early students of science. "(1982) However, I think most people would object to participating to a gambling game based on our example. You get to draw one number from the random number generator and your opponent gets to draw five. Whoever produces the largest number wins. Intuitively, we know that this an unfair game.

Appendix C

R Programs

```
%This function generates an index that will be used to generate a random
%sample within the experimental conditionns specified by the index.
%``corrPair'' refers to which of the sixteen majority-contaminant
%correlation matrix pairs will be used.  These pairs are numbered
%according to the order of their appearance in the Methods Chapter. ``n''
%specifies the size of the sample to be generated, and ``f'' the fraction
%of of contaminants to be generated.

partial.indicies=function() {

   corPair=rep(1:16, rep(5400, 16))

   n=rep(rep(c(80, 160, 320), rep(1800, 3)), 16)

   f=rep(rep(c(0.08, 0.16, 0.32), rep(600, 3)), 48)

   cbind(corPair, n, f)

}
```

```
%This function produces the desired sample corresponding to the index that
%it is given.  It also has the index vector "inOutIndex" appended to the
%data matrix as the last column.  A zero in this vector indicates the
%corresponding row in the data matrix is a legitimate observation.  A
%one indicates an outlier.

for.each.partial.index=function(x) {

   legitCorMatrix=legitMatrixList[[x[1]]]

   outCorMatrix=outMatrixList[[x[1]]]

   p=dim(legitCorMatrix)[1]

   zLegit=t(chol(legitCorMatrix))

   zOut=t(chol(outCorMatrix))

   n=x[2]

   f=x[3]

   numberOuts=floor(f*n) + 1

   numberLegits=floor((1-f)*n)

   inOutIndex=rep(c(0, 1), c(numberLegits, numberOuts))

   legits=zLegit%*%matrix(rnorm(n=numberLegits*p), \\
      nrow=p, ncol=numberLegits)
```

```
   outs=zOut%*%matrix(rnorm(n=numberOuts*p), nrow=p, ncol=numberOuts)

   legits=t(legits)

   outs=t(outs)

   dataMatrix=rbind(legits, outs)

   cbind(dataMatrix, inOutIndex)

}

%This function computes the n carrig Ds for a n x p data matrix.  The
%to this function is the n x p data matrix.

carrigD=function(x) {

   carrigs=c()

   n=dim(x)[1]

   p=dim(x)[2]

   mcdList=cov.rob(x, cor=TRUE, quantile.used=floor((n+p+1)/2), \\
      method="mcd")

   Z=scale(x, center=mcdList$center, scale=sqrt(diag(mcdList$cov)))

   for(i in 1:n) {
```

```
      sum=0
      for(j in 1:(p-1)) {
         for(k in (j+1):p) {
            term=abs(2*(1-mcdList$cor[j, k])-(Z[i, j]-Z[i, k])^2)
            sum=sum+term
         }
      }
      carrigs=c(carrigs, sum)
   }
   return(carrigs)
}

%This program computes MCD, PCHIGH and PCLOW given the n x p sample data
%matrix as input.  The n MCDs PCHIGHs and PCLOWs are top level components
%in the list that is output.  The fourth top-level component in the list
%is the number of principal components retained from the eigen analysis
%of the correlation matrix output by the mcd algorithm.

mcds.low.and.high=function(x) {
```

```
n=dim(x)[1]

p=dim(x)[2]

mcdList=cov.rob(x, cor=TRUE, quantilie.used=floor((n+p+1)/2),
   method="mcd")

robPCsList=eigen(mcdList$cor)

compsRetained=sum(robPCsList$values>=1)

highVals=robPCsList$values[1:compsRetained]

lowVals=robPCsList$values[(compsRetained+1):p]

highPCs=robPCsList$vectors[ , 1:compsRetained]

lowPCs=robPCsList$vectors[ , (compsRetained+1):p]

Z=scale(x, center=mcdList$center, scale=sqrt(diag(mcdList$cov))

mcds=diag(Z%*%solve(mcdList$cor, t(Z)))

pcHighs=(Z%*%highPCs)^2%*%(1/highVals)

pcHighs=as.vector(pcHighs)

pcLows=(Z%*%lowPCs)^2%*%(1/lowVals)

pcLows=as.vector(pcLows)
```

```
}


%This function computes the n Bacon MLDs for the n x p data matrix
%that is %passed to the function.

mld=function(x) {

   n=dim(x)[1]

   p=dim(x)[2]

   Z=scale(x, center=TRUE, scale=TRUE)

   combnMatrix=combn(p, 2)

   mlceArray=array(rep(1, n*p^2), dim=c(p, p, n))

   uArray=array(rep(1, n*p^2), dim=c(p, p, n))

   meanMlceArray=array(rep(1, n*p^2), dim=c(p, p, n))

   for (k in 1:n) {

      for (h in 1:(p*(p-1)/2)) {

         i=combnMatrix[1, h]

         j=combnMatrix[2, h]

         stndScore1=Z[k, i]
```

```
      stndScore2=Z[k, j]

      mlceElement=optimise(neg.likelihood.rho,  \\
         interval=c(-0.999, 0.999), z1=stndScore1, \\
         z2=stndScore2)

      mlceArray[i, j, k]=mlceElement$minimum

      m=seq(-0.999, 0.999, by=0.008)

      uArray=[i, j, k]= \\
         sqrt(sum(-1*neg.likelihood.rho(m, \\
         z1=rep(stndScore1, 250), \\
         z2=rep(stndScore2, 250))* \\
        (mlceElement$minimum-m)^2)/ \\
        sum(-1*neg.likelihood.rho(m, \\
        z1=rep(stndScore1, 250), rep(stndScore2, 250))))

   }

}

meanMlceMatrix=(1/n)*apply(mlceArray, c(1, 2), sum)

meanMlceArray[ , , 1:n]=meanMlceMatrix

distanceArray=abs(meanMlceArray-mlceArray)/uArray

(2/(p*(p-1)))*apply(distanceArray, 3, sum)
```

```
}


neg.likelihood.rho=function(x, z1, z2) {

   -1*exp(-1*(z1^2-2*x*z1*z2+z2^2)/(2*(1-x^2)))/(2*pi*sqrt(1-x^2))

}


%This function calculates the hit rates and false alarm rates for each of
%the metrics.

metrics=function(x) {

   dataMatrix=x[ , 1:(dim(x)[2]-1)]

   numVar=dim(dataMatrix)[2]

   inOutIndex=x[ , dim(x)[2]]

   outIndicies=which(inOutIndex==1)

   legitIndicies=which(inOutIndex==0)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%% Carrig D Block %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

   carrigs=carrigD(dataMatrix)

   carrigClusters=kmeans(carrigs, centers=2)
```

```
   clargerGroup=which.max(carrigClusters$centers[ , 1])

   carrigPositiveIndicies=which(carrigClusters$cluster==clargerGroup)

   carrigHitRate=sum(is.element(carrigPositiveIndicies, outIndicies))
      /length(outIndicies)

   carrigFARate=sum(is.element(carrigPositiveIndicies, \\
      legitIndicies))/length(legitIndicies)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%% Mahalanobis D^2 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

   mahals=mahalanobis(dataMatrix, center=apply(dataMatrix, 2, mean),
      cov=cov(dataMatrix), inverted=FALSE)

   mahalPositives=which(mahals>qchisq(p=0.999, df=numVar,
      lower.tail=TRUE))

   mahalHitRate=sum(is.element(mahalPositives, outIndicies))/
      length(outIndicies)

   mahalFARate=sum(is.element(mahalPositives,
      legitIndicies))/length(legitIndicies)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%% MCD, PCHIGH, PCLOW Block %%%%%%%%%%%%%%%%%%%%%%%%%%%

   mcdsLowsAndHighs=mcds.low.and.high(dataMatrix)
```

```
compsRetained=mcdsLowsAndHighs[[4]]

mcds=mcdsLowsAndHighs[[1]]

mcdPositives=which(mcds>qchisq(p=0.999, df=numVar, lower.tail=TRUE))

mcdHitRate=sum(is.element(mcdPositives, outIndicies))/
   length(outIndicies)

mcdFARate=sum(is.element(mcdPositives, legitIndicies))/
   length(legitIndicies)

pchighs=mcdsLowsAndHighs[[2]]

highPositives=which(pchighs>qchisq(p=0.999, df=compsRetained,
    lower.tail=TRUE))

highHitRate=sum(is.element(highPositives, outIndicies))/
   length(outIndicies)

highFARate=sum(is.elemnent(highPositives, legitIndicies))
   /length(legitIndicies)

pclows=mcdsLowsAndHighs[[3]]

lowPositives=which(pclows>qchisq(p=0.999, df=(numVar-compsRetained),
   lower.tail=TRUE))

lowHitRate=sum(is.element(lowPositives, outIndicies))/
   length(outIndicies)
```

```
   lowFARate=sum(is.element(lowPositives, legitIndicies))/
      length(legitIndicies)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% MLD Block %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

   mlds=mld(dataMatrix)

   baconClusters=kmeans(bacons, centers=2)

   blargerGroup=which.max(baconClusters$centers[ , 1])

   baconPositives=which(baconClusterscluster==blargerGroup)

   baconHitRate=sum(is.element(baconPositives, outIndicies))/
      length(outIndicies)

   baconFARate=sum(is.element(baconPositives, legitIndicies))/
      length(legitIndicies)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%% Assemblage Block %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

   hitRates=c(carrigHitRate, mahalHitRate, mcdHitRate,
      highHitRate, lowHitRate, baconHitRate)

   fARates=c(carrigFARate, mahalFARate, mcdFARate, highFARate,
      lowFARate, baconFARate)

   c(hitRates, fARates)
```

```
}

%These are the commands that I entered at the prompt in the R environment
%to carry out the experiment.  The two assignment statements for
%"sampleCarrigs" and "sampleMlds" enabled me to save the values of these
%metrics for each observation within each of the 14400 samples.  I did this
%because I wanted to investigate the distribution these metrics within a
%sample.

partialIndicies=t(partial.indicies())

set.seed(1961)

samples=apply(partialIndicies, 2, for.each.partial.index)

sampleCarrigs=sapply(samples, carrig2)

sampleMlds=sapply(samples, mld2)

obs=sapply(samples, metrics)
```

References

Anderson, T. (1958). *An introduction to multivariate statistical analysis.* New York: John Wiley and Sons.

Bacon, D. R. (1995). A maximum likelihood approach to correlational outlier detection. *Multivariate Behavioral Research*, *30*, 125–148.

Barnett, V. (1978). The study of outliers: Purpose and model. *Applied Statistics*, *27*, 242–250.

Barnett, V., & Lewis, T. (1984). *Outliers in statistical data.* New York: John Wiley and Sons.

Beckman, R., & Cook, R. (1983). Outlier.....s. *Technometrics*, *25*, 119–149.

Belsley, D., Kuh, E., & Welsch, R. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity.* New York: Wiley.

Bollen, K. (1987). Outliers and improper solutions: A confirmatory factor analysis example. *Sociological Methods and Research*, *17*, 55–64.

Carrig, M. M. (2005). *Detection of multivariate mean vector and covariance matrix outliers in behavioral sciences data.* Unpublished doctoral dissertation, The University of North Carolina at Chapel Hill.

Chatterjee, S., Jamieson, L., & Wiseman, F. (1991). Identifying most influential observations in factor analysis. *Marketing Science*, *10*(2), 145–160.

Comrey, A. (1985). A method for removing outliers to improve factor analytic results. *Multivariate Behavioral Research*, *20*, 273–281.

Devlin, S., Gnanadesikan, R., & Kettenring, J. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika*, *62*, 531–545.

Devlin, S., Gnanadesikan, R., & Kettenring, J. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, *76*, 354–362.

Dillon, W., & Goldstein, M. (1984). *Multivariate analysis: Methods and applications*. New York: John Wiley and Sons.

Dixon, W. (1950). *Analysis of extreme values* (Vol. 21).

Draper, N., & Smith, H. (1981). *Applied regression analysis*. New York: John Wiley and Sons.

Egan, W., & Morgan, S. (1998). Outlier detection in multivariate analytical chemical data. *Analytical Chemistry*, *70*, 2372–2379.

Glass, G., & Hopkins, K. (1996). *Statistical methods in education and psychology*. Needham Heights, MA: Allyn and Bacon.

Gnanadesikan, R., & Kettenring, J. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, *28*, 81–124.

Hawkins, D. (1974). The detection of errors in multivariate data using principal components. *Journal of the American Statistical Association*, *69*(346), 340–344.

Hawkins, D. (1980). *Identification of outliers*. New York: Chapman and Hall.

Hawkins, D., & Fatti, L. (1984). Exploring multivariate data using the minor principal components. *The Statistician*, *33*, 325–338.

Hoaglin, D., & Welsch, R. (1978). The hat matrix in regression and anova. *The American Statistician*, *32*, 17–22.

Hove, H., Liang, Y., & Kvalheim, O. (1995). Trimmed object projections: A nonparametric robust latent-structure decomposition method. *Chemometrics and Intelligent Laboratory Systems*, *27*, 33–40.

Huber, P. (1983). *Robust statistics*. New York: John Wiley and Sons.

Johnson, R., & Wichern, D. (1992). *Applied multivariate statistical analysis*. Saddle River, New Jersey: Prentice Hall.

Jolliffe, I. (1986). *Principal components analysis*. New York: Springer-Verlag.

Khattree, R., & Naik, D. (2000). *Multivariate data reduction and discrimination with sas software*. New York: John Wiley and Sons.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (Third ed.). Belmont, California: Wadsworth.

Larson, H. J. (1982). *Introduction to probability theory and statistical inference* (Third ed.). New York: Wiley.

Mahalanobis, A. (1983). *Prasanta chandra mahalanobis*. New Delhi: National Book Trust, India.

Mardia, K. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*, 519–530.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Minitab reference manual release 11 [Computer software manual]. (1999).

Nair, K. (1948). *The distribution of the extreme deviate from the sample mean and its studentized form* (Vol. 35).

Nair, K. (1952). *Tables of the percentage points of the 'studentized 'extreme deviate from the sample mean* (Vol. 39).

Novotny, J. A. (2001). *Detection and evaluation of multivariate outliers*. Unpublished doctoral dissertation, University of Denver.

Rao, C. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya: The Indian Journal of Statistics*, *26*, 329–358.

Rassmussen, J. (1988). Evaluating outlier identification tests: Mahalanobis d squared and comrey d. *Multivariate Behavioral Research*, *23*, 189–202.

Rocke, D., & Woodruff, D. (1996). Identification of outliers in multivariate data. *Journal of the American Statistical Association*, *91*, 1047–1061.

Rousseeuw, P. (1985). Mathematical statistics and applications. In W. GROSSMAN, G. PFLUG, I. VINCE, & W. WERTZ (Eds.), (Vol. B, pp. 283–297). Dordrecht: Reidel Publishing.

Rousseeuw, P., & Zomeren, B. van. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, *85*, 633–639.

Rudra, A. (1996). *Prasanta chandra mahalanobis a biography.* Delhi: Oxford University Press.

Schott, J. (1997). *Matrix analysis for statistics.* New York: John Wiley and Sons.

Schwager, S., & Margolin, B. (1982). Detection of multivariate normal outliers. *Annals of Statistics*, *10*, 943–954.

Siotani, M. (1959). *The extreme value of the generalized distance of the individual points in the multivariate normal sample* (Vol. 10).

Stapleton, J. (1998). *Linear statistical models.* New York: John Wiley and Sons.

Stevens, J. (n.d.). *Applied multivariate statistics for the social sciences.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Tabachnick, B., & Fidell, L. (1996). *Using multivariate statistics* (Third ed.). New York: HarperCollins.

Verzani, J. (2005). *Using r for introductory statistics.* Boca Raton: Chapman and Hall/CRC.

Walczak, B., & Massart, D. (1995). Robust principal components regression as a detection tool for outliers. *Chemometrics and Intelligent Laboratory Systems*, *27*, 41–54.

Woodruff, D., & Rocke, D. (1994). *Computable robust estimation of multivariate location and shape in high dimension using compound estimators* (Vol. 89).

Vita

Paul Muse Ritter was born in Polk County, Florida on October 25, 1961, the son of Paul Muse Ritter, Sr. and Mayre Louise McDowell. He earned a Bachelor of Arts in Sociology (1984) and a Bachelor of Science in Mathematics (1988) from the University of Central Florida. He received a Master of Science Statistics from the Mathematics Department at the University of Texas at Austin in 1998. From 1989 to 1994 he worked as a field statistician on the Florida Citrus Crop Forecast Survey. He currently works as a nightwatchman where he has plenty of time to pursue his academic interests.

Permanent Address: 2602 Hidden Oaks Drive Austin, TX 78745

This dissertation was typeset with LaTeX 2_ε[1] by the author.

[1] LaTeX 2_ε is an extension of LaTeX. LaTeX is a collection of macros for TeX. TeX is a trademark of the American Mathematical Society. The macros used in formatting this dissertation were written by Dinesh Das, Department of Computer Sciences, The University of Texas at Austin, and extended by Bert Kay, James A. Bednar, and Ayman El-Khashab.