The Distribution of Robust Distances

Johanna Hardin

Division of Statistics

University of California at Davis

hardin@wald.ucdavis.edu

David M. Rocke

Center for Image Processing and Integrated Computing

University of California at Davis

dmrocke@ucdavis.edu

September 1999

Mahalanobis-type distances in which the shape matrix is derived from a consistent, high-breakdown robust multivariate location and scale estimator have an asymptotic chi-squared distribution as is the case with those derived from the ordinary covariance matrix. However, even in quite large samples, the chi-squared approximation is poor. We provide an improved F approximation that gives accurate outlier rejection points for various sample sizes.

1. Introduction

In one or two dimensions, outlying data that are sufficiently far from the main mass of data are easily identified from simple plots, but detection of outliers is more challenging in higher dimensions. In multivariate applications, with three or more dimensions, outliers can be difficult or impossible to identify from coordinate plots of observed data. Although the outliers may lie at a great distance from the main body of data in a certain projection, identification of this projection can be difficult.

Various methods for detecting outliers have been studied (Atkinson, 1994; Barnett and Lewis, 1994; Gnanadesikan and Kettenring, 1972; Hadi, 1992; Hawkins, 1980; Maronna and Yohai, 1995; Penny, 1995; Rocke and Woodruff, 1996; Rousseeuw and Van Zomeren, 1990). One way to identify possible multivariate outliers is to calculate a distance from each point to a "center" of the data. An outlier would then be a point with a distance larger than some predetermined value. A conventional measurement of quadratic distance from a point X to a location Y given a shape S, in the multivariate setting is:

$$d_S^2(X,Y) = (X-Y)'S^{-1}(X-Y)$$

This quadratic form is often called the Mahalanobis Squared Distance (MSD). If there are only a few outliers, large values of $d_S^2(x_i, \bar{X})$, where $\bar{X}$ and S are the conventional sample mean and covariance matrix, indicate that the point x_i is an outlier (Barnett and Lewis, 1994). The distribution of the MSD with both the true location and shape parameters and the conventional location and shape parameters is well known (Gnanadesikan and Kettenring, 1972). However, the conventional location and shape parameters are not robust to outliers, and the distributional fit to the distance breaks down when robust measures of location and shape are used in the MSD (Rousseeuw and Van Zomeren, 1991).

Determining exact cutoff values for outlying distances continues to be a difficult problem.

In trying to detect single outliers in a multivariate normal sample, $d^2_S(x_i, \bar{X})$ will identify sufficiently outlying points. In data with clusters of outliers, however, the distance measure $d^2_S(x_i, \bar{X})$ breaks down (Rocke and Woodruff, 1996). Data sets with multiple outliers or clusters of outliers are subject to problems of masking and swamping (Pearson and Chandra Sekar, 1936). Masking occurs when a group of outlying points skews the mean and covariance estimates toward it, and the resulting distance of the outlying point from the mean is small. Swamping occurs when a group of outlying points skews the mean and covariance estimates toward it and away from other inlying points, and the resulting distance from the inlying points to the mean is large. As an example, consider a data set due to Hawkins, Bradu, and Kass (Hawkins et al., 1984). These data consist of 75 points in dimension three. We can only see one outlying point, but 14 of the points were constructed to be outliers. By using the mean and variance of all the data, we have masked the remaining 13 outliers. (See Figure 1)

Problems of masking and swamping can be resolved by using robust estimates of shape and location, which by definition are less affected by outliers. Outlying points are less likely to enter into the calculation of the robust statistics, so they will not be able to influence the parameters used in the MSD. The inlying points, which all come from the underlying distribution, will completely determine the estimate of the location and shape of the data. Some robust estimators of location and shape include the minimum covariance determinant (MCD) and the minimum volume ellipsoid (MVE) of Rousseeuw (Hampel et al., 1986; Rousseeuw, 1984; Rousseeuw and Leroy, 1987) and M-estimators and S-estimators. (Campell, 1980; Campell, 1982; Huber, 1981; Kent and Tyler, 1991; Lopuhaä, 1992; Maronna, 1976; Rocke, 1996; Tyler, 1983; Tyler, 1988; Tyler, 1991). By using a robust location and shape estimate in the MSD, outlying points will not skew the estimates and can be identified as outliers by large values of the MSD.

The MSD can take as its arguments any location and shape estimates. In this paper we are interested in robust location and shape estimates, which are better suited for detecting outliers. In particular, we are interested in the MCD location and shape estimates. Given n data points, the MCD of those data is the mean and covariance matrix based on the sample of size h ($h \leq n$) that minimizes the determinant of the covariance matrix (Rocke and Woodruff, 1996; Rousseeuw, 1984).

$$\begin{aligned} MCD &= (\bar{X}^*_J, S^*_J) \\ \text{where} \qquad \text{J} &= \{\text{set of } h \text{ points} : |S^*_J| \leq |S^*_K| \quad \forall \text{ sets K} \;\; \text{s.t. } |K| = h\} \\ \bar{X}^*_J &= \frac{1}{h}\sum_{i \in J} x_i \\ S^*_J &= \frac{1}{h}\sum_{i \in J}(x_i - \bar{X}^*_J)(x_i - \bar{X}^*_J)^t \end{aligned}$$

The value h can be thought of as the minimum number of points which must not be outlying. The MCD has its highest possible break down at $h = \lfloor \frac{(n+p+1)}{2} \rfloor$ where $\lfloor \cdot \rfloor$ is the greatest integer function (Rousseeuw and Leroy, 1987; Lopuhaä and Rousseeuw, 1991). We will use $h = \lfloor \frac{(n+p+1)}{2} \rfloor$ in our calculations and refer to a sample of size h as a half sample. The MCD is computed from the "closest" half sample, and therefore, the outlying points will not skew the MCD location or shape estimate. Large values of MSDs, using the MCD location ($\bar{X}^*$) and shape estimate (S^*), will be robust estimates of distance and will correctly identify points as outlying. Recall the constructed data by Hawkins, Bradu, and Kass. Using the MCD estimates, the distances give a clear identification of the 14 outlying points. (See Figure 2)

Not every data set will give rise to an obvious separation between the outlying and non-outlying points. Consider the data given by Daudin, Dauby and Trecourt and analyzed by Atkinson (Daudin et al., 1988; Atkinson, 1994). The data are eight measurements on 85 bottles of milk. Using the robust MCD estimates, we are not subject to masking or swamping, but we are not sure which group of points should be considered as outlying. (See Figure 3)

In Figure 2, points were identified as obvious outliers, but in many situations (including Figure 3) it will be important to construct a minimum outlying distance. For some known constant c, $c \cdot d^2_{S^*}(x_i, \bar{X}^*)$ are asymptotically distributed as χ^2_p, but the asymptotic convergence is very slow, and the χ^2_p quantiles will be larger than the corresponding MSD quantiles for even quite large samples. Use of χ^2_p quantiles as cutoff points will often lead to identifying too many points as outliers (Rousseeuw and Van Zomeren, 1991).

Finding a good approximation to the distribution of $d^2_{S^*}(x_i, \bar{X}^*)$ will lead to cutoff values that identify minimum outlying values, even for clusters of outliers. If $\{x_i\}$ come from a multivariate normal sample, and mS comes from an independent Wishart, then $d^2_S(x_i, \mu)$ will have a distribution that is a multiple of an F statistic (Mardia et al., 1979). Since the MCD shape and location estimates are calculated using only the inlying points, $\bar{X}^*$ and S^* can be thought of as asymptotically independent from the extreme values in the sample. We can also approximate the distribution of S^* by matching the first two moments of a Wishart. Accordingly, the $d^2_{S^*}(x_i, \bar{X}^*)$ will be approximately distributed as a multiple of an F statistic. This insight allows us to find cutoff values for outlying points is in the estimation of the degrees of freedom associated with the F statistic. We will examine various cutoff values for MSD with MCD shape and location estimates for multivariate normal data given different values of n and p.

2. Robust Estimators for Outlier Detection

The estimation of multivariate location and shape is one of the most difficult problems in robust statistics (Campell, 1980; Campell, 1982; Davies, 1987; Devlin et al., 1981; Donoho, 1982; Hampel et al., 1986; Huber, 1981; Lopuhaä, 1989; Maronna, 1976; Rocke and Woodruff, 1993; Rousseeuw, 1985; Rousseeuw and Leroy, 1987; Stahel, 1981; Tyler, 1983; Tyler, 1991). For some statistical procedures, it is relatively straightforward to obtain estimates that are resistant to a reasonable fraction of outliers– for example, one dimensional location (Andrews et al., 1972) and regression with error-free predictors (Huber, 1981). The multivariate location and shape problem is more difficult, because many known methods will break down if the fraction of outliers is larger than 1/(p+1), where p is the dimension of the data (Donoho, 1982; Maronna, 1976; Stahel, 1981). This means that in high dimensions, a small amount of outliers can result in arbitrarily bad estimates.

2.1. Affine Equivariant Estimators

We are particularly interested in affine equivariant estimators of the data. A location estimator $\mathbf{y}_n \in \mathbb{R}^p$ is affine equivariant if and only if for any vector $\mathbf{v} \in \mathbb{R}^p$ and any nonsingular $p \times p$ matrix $\mathbf{M}$,

$$\mathbf{y}_n(\mathbf{M}x + \mathbf{v}) = \mathbf{M}\mathbf{y}_n(x) + \mathbf{v}.$$

A shape estimator $\mathbf{S}_n \in PDS(p)$ (the set of $p \times p$ positive definite symmetric matrices) is affine equivariant if and only if for any vector $\mathbf{v} \in \mathbb{R}^p$ and any nonsingular $p \times p$ matrix $\mathbf{M}$,

$$\mathbf{S}_n(\mathbf{M}x + \mathbf{v}) = \mathbf{M}\mathbf{S}_n(x)\mathbf{M}'.$$

Stretching or rotating the data will not change an affine estimate of the data. If the location and shape estimates are affine equivariant, the Mahalanobis Squared Distances are an affine invariant, which means the shape of the data determines the distances between the points. The only other real alternative is to make a prior assumption about the correct distance measure. It is important to have affine equivariant estimators so that the measurement scale, location, and orientation can be ignored. Since MSD's are affine invariant, the properties and procedures that use the MSD can be calculated without loss of generality for standardized distributions. For the properties under normality, we can use N(0,I).

2.2. Minimum Covariance Determinant

The Minimum Covariance Determinant (MCD) location and shape estimates are resistant to outliers because the outliers will not be involved in the location and shape calculations. From the MCD sample, the sample mean and covariance matrix, which are robust estimates of the location and shape of the underlying population, can be computed.

Finding the exact MCD sample can be time consuming and difficult. The only known method for finding the exact MCD is to search every half sample and calculate the determinant of the covariance matrix of that sample. For n=20, the search would require computing about 184,756 determinants; for n=100, the search would require computing about 10^{29} determinants. With any conceivable computer, it is clear that finding the exact MCD is intractable by enumeration.

2.3. Estimating the MCD

Since the exact MCD is often impossible to find, the algorithm used to estimate of the MCD is, in some sense, the estimator. Various algorithms have been suggested for estimating the MCD.

Hawkins proposed a method based on swapping points in and out of a sample of size h. The basic algorithm is as follows.

- Start with a subsample of size h, H_1.
- Swap points $x_{i'} \in H_1$ and $x_{j'} \notin H_1$ and call the new subsample H_2 if:

$$d_i = \det(\text{cov}(H_1)) - \det(\text{cov}(H_2)) > 0$$

  AND the above difference, d_i, is maximized over swapping all possible pairs of points $x_i \in H_1$ and $x_j \notin H_1$
- Let H_2 be the new subsample of size h
- Repeat the process until no swap lowers the $\det(\text{cov}(H_i))$ (or equivalently, until no swap gives $d_i > 0$) (Hawkins, 1994; Hawkins, 1999).

A faster algorithm was found independently by Hawkins (1999) and Rousseeuw and Van Driessen (1999). The core of the algorithm is based on what Rousseeuw calls the C-step. Instead of swapping one pair of points in and out, the C-step allows for many points to be interchanged at each step. Again, we start with a subset of the data of size h, H_1. We can compute $\bar{X}_{H_1}, S_{H_1}$, and $d^2_{S_{H_1}}(x_i, \bar{X}_{H_1}) = d^2_{H_1}(i)$ for each point $i = 1, \ldots n$ based on the sample H_1. We can then sort the distances based on a permutation π so that:

$$d^2_{H_1}(\pi(1)) \leq d^2_{H_1}(\pi(2)) \leq \ldots \leq d^2_{H_1}(\pi(n))$$

We will then assign $\{\pi(1), \pi(2), \ldots, \pi(h)\}$ to H_2. Using H_2 we can calculate $\bar{X}_{H_2}, S_{H_2}$, and $d^2_{S_{H_2}}(x_i, \bar{X}_{H_2})$ and repeat the process until the permutation, π, does not change. Rousseeuw and VanDriessen (1999) showed that the process will converge.

The question remains for both algorithms, where does the initial H_1 come from? Previously, Hawkins used a random subset of size h from the data. If the data have large amounts of contamination, a random subset of size h

will almost never look like the true underlying (uncontaminated) population (it will look most like the contaminated sample), so it will be hard for either swapping algorithm to converge to the true uncontaminated shape of the data. For contaminated data, Rousseeuw proposed starting with a random subset of size p+1 (the minimum number of points needed to define a nonsingular covariance matrix) and adding points until a subset of h points is constructed (Rousseeuw and Van Driessen, 1999). Points are added to the initial random subset based on their distances to the mean of the initial subset. The algorithm is as follows.

- Let H_0 be a random subset of $p + 1$ points
- Find $\bar{X}_{H_0}$ and S_{H_0} (If $\det(S_{H_0}) = 0$ then add random points until $\det(S_{H_0}) > 0$)
- Compute the distances $d^2_{S_{H_0}}(x_i, \bar{X}_{H_0}) = d^2_{H_0}(i)$ and sort them for some permutation π such that,

$$d^2_{H_0}(\pi(1)) \leq d^2_{H_0}(\pi(2)) \leq \ldots \leq d^2_{H_0}(\pi(n))$$

- $H_1 := \{\pi(1), \pi(2), \ldots, \pi(h)\}$

A random subset of p+1 points has a better chance of by chance looking like the uncontaminated data, and so H_1 here will be sometimes closer to the true data than a random subset of h points (and we only need this to happen on one of the trials). In this paper we are interested in finding quantiles for distances based on MCD estimates. Our simulations are all done with pure multivariate normal data, and therefore, we are able to find quantiles using data that was uncontaminated. Since our data was not contaminated, it was more effective to start with random subsets of size h, since diversity of starting points is of no value when there is no contamination.

The algorithm we used to estimate the MCD begins with a series of random starts, each of which is a randomly chosen half sample (or sample of size h) of the data points. We then use the algorithm referred to above as the C-step. For each random start, the procedure for calculating the MCD sample is as follows.

1. Compute the mean and covariance of the current half sample.
2. Calculate the MSD, based on the mean and covariance from step 1, for each point in the entire data set.
3. Choose a half sample of those points with the smallest MSDs from step 2.
4. Repeat steps 1-3 until the half sample no longer changes.

MCD sample will then be the half sample (in 3) with the minimum covariance determinant of all the random starts. A robust estimator like $d^2_{S^*}(x_i, \bar{X}^*)$, where S^* and $\bar{X}^*$ are the MCD shape and location estimates, is likely to detect outliers because outlying points will not affect the MCD shape and location estimates. For points x_i that are extreme, $d^2_{S^*}(x_i, \bar{X}^*)$ will be large, and for points x_i that are not extreme, $d^2_{S^*}(x_i, \bar{X}^*)$ will not be large. Here we are not subject to problems of masking and swamping.

3. Distance Distributions

Mahalanobis squared distances give a one-dimensional measure of how far a point is from a location with respect to a shape. Using MSD we can find points that are unusually far away from a location and call those points outlying. We have discussed the importance of using robust affine equivariant estimates for the location and shape of the data. Unfortunately, using robust estimates gives MSDs with unknown distributional properties. Consider n multivariate data points in dimension p, $x_i \sim N(\mu, \Sigma)$. Let S be an estimate of Σ such that, $mS \sim \text{Wishart}_p(\Sigma, m)$. Below are three distributional results for distances based on the above type of multivariate normal data.

1. The first distance distribution is based on the true parameters μ and Σ. We know that if the data are normal, these distances have an exact χ^2_p distribution (Mardia et al., 1979).

$$d^2_\Sigma(x_i, \mu) \quad \sim \quad \chi^2_p$$

Which gives:

$$\begin{aligned} \mathrm{E}[d^2_\Sigma(x_i, \mu)] &= p \\ \mathrm{Var}[d^2_\Sigma(x_i, \mu)] &= 2p \end{aligned}$$

2. The second distance distribution is based on the ordinary mean and covariance estimates. These distances have an exact Beta distribution (Gnanadesikan and Kettenring, 1972; Wilks, 1962). It is interesting to note that the unbiased estimator has a smaller variance than the estimator which takes μ and Σ as parameters. This is because fitting the mean and covariance allows the distances to be made smaller because the estimates accommodate random fluctuations in the data.

Given,

$$\begin{aligned} \bar{X} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ S &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})(x_i - \bar{X})^t \end{aligned}$$

then,

$$\frac{(n-1)^2}{n} d^2_S(x_i, \bar{X}) \sim \mathrm{Beta}\left(\frac{p}{2}, \frac{(n-p-1)}{2}\right)$$

Which gives:

$$\mathrm{E}\left[\frac{n d^2_S(x_i, \bar{X})}{(n-1)}\right] \quad = \quad p$$

$$\mathrm{Var}\left[\frac{nd_S^2(x_i,\bar{X})}{(n-1)}\right] = 2p\frac{(n-p-1)}{(n+1)}$$

3. The third distance distribution is based on an estimate of S that is independent of the x_i. These distances have an exact F distribution when μ is the location parameter (Mardia et al., 1979), and an approximate F distribution when $\bar{X}$ is the location parameter (using a Slutsky type argument (Serfling, 1980)). It is interesting to note here that the unbiased estimator has a larger variance than the estimator which takes μ and Σ as its parameters. This is because the independent variation in S adds to the variability of the distances which are in part functions of S.

   Given S and x_i independent,

$$\frac{np}{n-p}d_S^2(x_i,\mu) \sim F_{p,n-p}$$

   Using a variant of Slutsky's Theorem,

$$\frac{np}{n-p}d_S^2(x_i,\bar{X}) \dot{\sim} F_{p,n-p}$$

   Which gives:

$$\begin{aligned}\mathrm{E}\left[\frac{(n-p-2)}{n}d_S^2(x_i,\bar{X})\right] &\doteq p\\ \mathrm{Var}\left[\frac{(n-p-2)}{n}d_S^2(x_i,\bar{X})\right] &\doteq 2p\frac{(n-2)}{(n-p-4)}\end{aligned}$$

We will refer to the standard location and shape estimates ($\bar{X}$ and S) as within sample estimates and the MCD location and shape estimates ($\bar{X}^*$ and S^*) as out of sample estimates because extreme observations will not be used to calculate the MCD (with high probability). Our interest is in the extreme points which enter into the within sample calculations but not the out of sample calculations.

Since $\bar{X}$ and $\bar{X}^*$ are consistent estimators for μ, and S and $c^{-1}S^*$ (for some constant c) are consistent estimators for Σ, we know that the within sample and out of sample MSD are both asymptotically χ^2_p statistics (Mardia et al., 1979; Serfling, 1980). χ^2_p quantiles are often used for identifying MSD extrema (often inappropriately, as pointed out by (Rousseeuw and Van Zomeren, 1991)).

The main insight behind this paper is that distances based on MCD estimates of location and shape will behave like case 1 or 2 above for points that were used to calculate the MCD (equivalently, that have MSD's in the lower half of the empirical distribution of distances), and will behave more like case 3 for more extreme points. The theorem below confirms the intuition that extreme points have a distribution that is conditionally independent of the distribution of the MCD location and shape. Since the MCD shape estimate does not have a Wishart distribution, the case 3 formulae do not hold without alteration, but we use this as a framework for providing an approximating distribution for robust MSD's that is a large improvement on the χ^2_p approximation.

The elbow pattern in robust MSD's described by (Rousseeuw and Van Zomeren, 1991) can be seen in Figures 4 and 5, which show the mean ordered MSD's from the MCD in two different situations plotted against the χ^2_p quantiles. The distances that are in the smallest half of distances (coming from points that are included in the MCD subset) appear to follow a chi^2_p distribution since they lie on the line $y = x$, while the larger distances diverge in a systematic pattern.

Theorem 3.1 *Given n points, X_i, independently and identically distributed (iid) $N_p(\mu, \Sigma)$, find the MCD sample based on a fraction $\epsilon = \frac{h}{n}$ of the sample, and choose δ such that $\epsilon < \delta < 1$. Then points X_i such that $(X_i - \mu)'\Sigma^{-1}(X_i - \mu) > \chi^2_{p,\delta}$, X_i will be asymptotically independent of the MCD sample.*

Proof. We will show that points satisfying the condition have, for large n, a vanishingly small chance of being included in the MCD subset, regardless of any other property that they may have, which will imply the asymptotic independence that we seek. The proof will be given in steps.

1. We can think of the iid sample as coming from 3 truncated Normal distributions. We can generate an iid sample from normal distribution as follows:

   - Let n_1, n_2, n_3 come from a Multinomial $(n, \epsilon, \delta - \epsilon, 1 - \delta)$ distribution.
   - Let $X_1, X_2, \ldots, X_{n_1}$ come from a truncated normal distribution. The truncated normal distribution will be $N_p(\mu, \Sigma)$ with a truncation so that each of the points have $(X - \mu)'\Sigma^{-1}(X - \mu) \leq \chi^2_{p,\epsilon}$.
   - Let $X_{n_1+1}, X_{n_1+2}, \ldots, X_{n_1+n_2}$ come from a truncated normal distribution. The truncated normal distribution will be $N_p(\mu, \Sigma)$ with a truncation so that each of the points have $\chi^2_{p,\epsilon} < (X - \mu)'\Sigma^{-1}(X - \mu) \leq \chi^2_{p,\delta}$.
   - Let $X_{n_2+1}, X_{n_2+2}, \ldots, X_{n_1+n_2+n_3=n}$ come from a truncated normal distribution. The truncated normal distribution will be $N_p(\mu, \Sigma)$ with a truncation so that each of the points have $\chi^2_{p,\delta} < (X - \mu)'\Sigma^{-1}(X - \mu)$.

   Then the sample, $X_1, \ldots, X_{n_1}, X_{n_1+1}, \ldots, X_{n_1+n_2}, X_{n_1+n_2+1} \ldots, X_{n_1+n_2+n_3=n}$, is an iid sample of size n from $N_p(\mu, \Sigma)$.

   We can define the ellipsoidal regions R_1, R_2, R_3, such that

$$\begin{aligned}
\text{if} \quad (X-\mu)'\Sigma^{-1}(X-\mu) \leq \chi^2_{p,\epsilon} \quad &\Rightarrow \quad X \in R_1 \\
\text{if} \quad \chi^2_{p,\epsilon} < (X-\mu)'\Sigma^{-1}(X-\mu) \leq \chi^2_{p,\delta} \quad &\Rightarrow \quad X \in R_2 \\
\text{if} \quad (X-\mu)'\Sigma^{-1}(X-\mu) > \chi^2_{p,\delta} \quad &\Rightarrow \quad X \in R_3
\end{aligned}$$

   Note that these regions are based on the true ellipsoidal contours of the generating distribution and are not data dependent.

2. Letting the MCD location and shape matrix be denoted by $\bar{X}^*$ and S^*, we know,

$$\begin{aligned} \bar{X}^* &\rightarrow \mu \\ \frac{1}{c}S^* &\rightarrow \Sigma \quad \text{for some } c \ \ \text{(Tyler, 1983)} \end{aligned}$$

which gives,

$$c(X-\bar{X}^*)'S^{*-1}(X-\bar{X}^*) \rightarrow (X-\mu)'\Sigma^{-1}(X-\mu) \qquad \forall X.$$

If $X \in R_1$,

$$c(X-\bar{X}^*)'S^{*-1}(X-\bar{X}^*) \leq \chi^2_{p,\epsilon} + O_p(n^{-1/2})$$

If $X \in R_3$,

$$c(X-\bar{X}^*)'S^{*-1}(X-\bar{X}^*) \geq \chi^2_{p,\delta} - O_p(n^{-1/2})$$

3. Let R_1^* be the ellipsoid generated by the MCD sample and that just contains all of the MCD points. The radius of R_1^* will be $\chi^2_{p,\epsilon} + O_p(n^{-1/2})$. We will show that the points in R_3 will almost never, for large n, have MCD distances that will place them in R_1^*; i.e., will almost never be in the MCD subsample.

   Let $y \in R_3$, so that $(y-\mu)'\Sigma^{-1}(y-\mu) \geq \chi^2_{p,\delta}$. Then, $(y-\mu)'\Sigma^{-1}(y-\mu) = \chi^2_{p,\delta} + W$, where $W \geq 0$. Also, $c(y-\bar{X}^*)'S^{*-1}(y-\bar{X}^*) = \chi^2_{p,\delta} + W + O_p(n^{-1/2})$.

   Now,

$$\begin{aligned} P(y \in \text{ MCD sample}) &= P(c(y-\bar{X}^*)'S^{*-1}(y-\bar{X}^*) \leq \chi^2_{p,\epsilon} + O_p(n^{-1/2})) \\ &= P(\chi^2_{p,\delta} + W + O_p(n^{-1/2}) \leq \chi^2_{p,\epsilon} + O_p(n^{-1/2})) \\ &= P(\chi^2_{p,\delta} - \chi^2_{p,\epsilon} + W \leq O_p(n^{-1/2})) \\ &\rightarrow 0. \end{aligned}$$

If it were the case that the MCD subset could never include a point $y \in R_3$, then there would be exact independence. Any failure of independence involves a point $y \in R_3$ being in the MCD sample, which we have shown to be extremely improbably in large samples. $\square$

The experimental independence of the extreme points and the MCD sample can also be seen in Figure 6. The picture shows average distances of two sets of independently simulated data sets whose distances were computed using the same MCD estimates. The first set contains the MCD sample, the second set was generated completely independently of the first sample and the MCD estimates. The MCD estimates are approximately independent of the extreme points, and so the extrema behave like the completely independently generated data.

The only remaining step in approximating the distribution of the extreme distances from the MCD is to approximate the distribution of the MCD shape by a Wishart, so that we can apply the F-distribution result cited above.

If X_i is multivariate normal data, and $\bar{X}^*$ and S_Y^* are the MCD mean and covariance, then

1. $X_1, \ldots, X_n \sim N_p(\mu, \Sigma)$,
2. the distribution of S_X^* can be approximated by,

   $$mc^{-1} S_X^* \dot{\sim} \text{Wishart}_p(m, \Sigma), \tag{3.1}$$

   where m is an unknown degrees of freedom, and c is a constant satisfying $\mathrm{E}[S_X^*] = c\Sigma$ (which holds for some c because S_X^* is an affine equivariant shape estimator of Σ (Tyler, 1983)), and
3. the tail elements of X_i are approximately independent of S_X^*. (See Theorem 3.1 and Figure 6)

Then, using $\bar{X}^* \to \mu$,

$$\frac{c(m-p+1)}{pm} d^2_{S^*_X}(X_i, \bar{X}^*) \dot{\sim} F_{p,m-p+1}. \tag{3.2}$$

Using the above F distribution to calculate cutoff values for distances based on the MCD sample is a robust way of identifying outliers. The only remaining problem, then, is to estimate c and m correctly.

3.1. Finding the Degrees of Freedom for the F Distribution

Using a method of moments identification by the coefficient of variation (CV), Welch and Satterthwaite (Welch, 1937; Welch, 1947; Satterthwaite, 1946) estimated the degrees of freedom (df) for the well-known test $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 \neq \mu_2$ (when $\sigma_1 \neq \sigma_2$) which has a test statistic,

$$\frac{(x_1 - x_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \dot{\sim} t_{df}$$

$$df = \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1 - 1} + \frac{V_2^2}{n_2 - 1}}$$

$$\text{Where} \qquad V_1 = \frac{S_1^2}{n_1} \text{ and } V_2 = \frac{S_2^2}{n_2}$$

Using a similar method of moments idea, we can estimate the degrees of freedom associated with the F distribution of $\frac{c(m-p+1)}{pm} d^2_{S^*_Y}(Y_i, \bar{Y}^*)$. If for some m, S^*_Y had a distribution that was a multiple of a Wishart, then it would be the case that

$$mc^{-1} s^*_{ii} \dot{\sim} \chi^2_m \sigma_{ii}, \tag{3.3}$$

where s^*_{ii} are the diagonal elements of S^*_Y. Since the estimators are affine equivariant, we perform all calculations without loss of generality on N(0,I) data, in which case $\sigma_{ii} = 1$ and the diagonal elements are identically distributed (Grübel and Rocke, 1990). The best estimates of c and m in the $\mathrm{N}(\mu, \Sigma)$ case will be the same as the estimates of c and m in the N(0,I) case because of this affine equivariance.

From (3),

$$\mathrm{E}[mc^{-1}s^*_{ii}] = m \Rightarrow \mathrm{E}[s^*_{ii}] = c$$

and

$$\mathrm{Var}[mc^{-1}s^*_{ii}] = 2m \Rightarrow \mathrm{Var}[s^*_{ii}] = \frac{2c^2}{m}$$

which gives:

$$\mathrm{CV} = \frac{\sqrt{\mathrm{Var}[s^*_{ii}]}}{\mathrm{E}[s^*_{ii}]} = \frac{c\sqrt{2/m}}{c} = \sqrt{\frac{2}{m}}.$$

So we can estimate m and c by

$$\hat{m} = \frac{2}{\hat{\mathrm{CV}}^2} \qquad\qquad \hat{c} = \frac{1}{h}\sum_{i=1}^{h} s^*_{ii}$$

where CV ($\hat{\mathrm{CV}}$) is the (estimated) coefficient of variation of the diagonal elements of the MCD shape estimator. This can be done either from the asymptotics of the MCD shape matrix or by simulation. Note that the simulation will be used only to compute the mean and variance of the diagonal elements of the covariance matrix, and not the distribution of the distance order statistics, which greatly simplifies the task. Since the diagonal elements are identically distributed and uncorrelated, we can simulate N copies of the $p \times p$ MCD shape matrix from the n data points in each sample, and then estimate c and m from the mean and coefficient of variation of the Np diagonal elements.

Alternatively, an asymptotic expression for c exists that works well even for small samples.

$$c = \frac{P(\chi^2_{p+2} < \chi^2_{(p,h/n)})}{h/n}$$

where χ^2_ν is a Chi-Square random variable with ν degrees of freedom, and $\chi^2_{\nu,\epsilon}$ is the ϵ cutoff point for a χ^2_ν random variable. This formula is easily derived and is apparently well known (also see (Croux and Haesbroeck, 1999)).

For m there exists an asymptotic expression that is good in large samples and only moderately accurate in small samples (Croux and Haesbroeck, 1999). For small samples, simulation may be necessary to estimate m accurately. Croux and Haesbroeck used influence functions to determine an asymptotic expression for the variance elements of the MCD sample. Details are given in the Appendix.

4. Results

A common and reasonable method for identifying clusters of outliers is to find robust distances and then compute distributional quantiles to determine cutoffs. Any point with a distance larger than the cutoff point will be an outlier. Three distributional cutoff choices have been described,

1. χ^2_p (which is known to reject too few points),
2. F (from (2)) with degrees of freedom calculated from the asymptotic formulas, and
3. F (from (2)) with degrees of freedom calculated from simulations.

We examined the performance of these methods in the null case by a Monte Carlo study with p=5, 10, 20 and n=50, 100, 500, 1000. First, simulations of the MCD shape estimators with 1000 trials were undertaken to obtain values for m and c for each pair of n and p. Then the cutoff values for 5%, 1%, and 0.1% rejection for each of the three distribution choices were calculated.

Next, 1000 sets of independent data for each pair of dimension and size were simulated, and the number of points the cutoffs identified as outlying was counted. For the 5% nominal test, the percent identified as outliers is shown in Table 1. As expected, the Chi-Square cutoff points are far too liberal. The problem is worse in higher dimension, but gets better in larger samples. The asymptotic cutoff points are an improvement on the Chi-Square cutoffs, but are too conservative, especially in small sample sizes. The performance at sample sizes of 500 and 1000 is not bad. The simulated cutoff points are still conservative, but are closer to the nominal values than the asymptotic values.

Results for 1% and 0.1% nominal tests are in Tables 2 and 3. Again, the Chi-Square cutoffs are too liberal, the asymptotic cutoffs are too conservative, and the simulated cutoffs are conservative but closer to the nominal than the asymptotic cutoffs. Cutoffs from either the asymptotic or simulated methods can be used for rejecting outliers in multivariate data without fear that more than the nominal proportion of good data will be rejected (on the average). Further improvements in the small-sample cutoffs will increase the sensitivity of the detection procedure, but the method as it now stands is a large improvement on the previously available methods.

From the tables, we can see that the asymptotic accuracy depends primarily on n and not on p. As expected, the asymptotic cutoff becomes more accurate as n increases. These results lead to the following recommendations:

1. For large values of n (at least 1000 observations), asymptotic formulas should be used for cutoff values of outlying MCD distances.

2. For smaller values of n (less than 500 observations), the asymptotic formula for c can be used, but simulation will be necessary to find m more accurately. For very small values of n, the simulation cutoffs are still supe-

rior to the currently used Chi-Square cutoff values; the simulation cutoffs are somewhat conservative for small samples. The simulation programs, in Fortran, are available from either author.

5. Conclusion

A new method for determining outlying points in a multivariate normal sample has been derived. The methods presented here are superior to the commonly used Chi-Square cutoff. Asymptotic values for the cutoffs work well in samples of 1000 or larger, while a somewhat more computationally intensive simulation method can be used for smaller samples sizes.

Because this work concerns clusters of outliers, there are implications for clustering as well as outlier identification. It is possible that robust distances may be able to identify outlying points in populations that are made up of two or more different clusters.

Also, the only robust method discussed in depth here is the MCD. The above methods can be extended to other robust methods like the Rousseeuw's Minimum Volume Ellipsoid (Hampel et al., 1986; Rousseeuw, 1984; Rousseeuw and Leroy, 1987), S-estimation, and M-estimation.

Acknowledgments

The research reported in this paper was supported by grants from the National Science Foundation (DMS 95-10511, DMS 96-26843, ACI 96-19020, and DMS 98-70172) and the National Institute of Environmental Health Sciences, National Institutes of Health (P42 ES04699).

A. Appendix

In this appendix we provide for completeness the formulas due to Croux and Haesbroeck (1999) needed to estimate the degrees of freedom parameter m of the Wishart approximation.

$$\begin{align}
\alpha &= \frac{n-h}{n} \tag{A.1}\\
&\ \text{where } n \text{ is the sample size and } h = \left\lfloor \frac{(n+p+1)}{2} \right\rfloor. \notag\\
q_\alpha &\ \text{is}\ \ \text{such that: } 1-\alpha = P(\chi^2_p \leq q_\alpha) \tag{A.2}\\
c_\alpha &= \frac{1-\alpha}{P(\chi^2_{p+2} \leq q_\alpha)} \tag{A.3}\\
c_2 &= \frac{-P(\chi^2_{p+2} \leq q_\alpha)}{2} \tag{A.4}\\
c_3 &= \frac{-P(\chi^2_{p+4} \leq q_\alpha)}{2} \tag{A.5}\\
c_4 &= 3 \cdot c_3 \tag{A.6}\\
b_1 &= \frac{c_\alpha(c_3-c_4)}{1-\alpha} \tag{A.7}\\
b_2 &= 0.5 + \frac{c_\alpha}{(1-\alpha)}\left(c_3 - \frac{q_\alpha}{p}\left(c_2 + \frac{(1-\alpha)}{2}\right)\right) \tag{A.8}\\
v_1 &= (1-\alpha)b_1^2(\alpha(\frac{c_\alpha q_\alpha}{p}-1)^2-1) - 2c_3c_\alpha^2(3(b_1-pb_2)^2 \tag{A.9}\\
&\ + (p+2)b_2(2b_1-pb_2)) \notag\\
v_2 &= n(b_1(b_1-pb_2)(1-\alpha))^2c_\alpha^2 \tag{A.10}\\
v &= \frac{v_1}{v_2} \tag{A.11}\\
\hat{m} &= \frac{2}{c_\alpha^2 v} \tag{A.12}
\end{align}$$

References

Andrews, D., Bickel, P., Hampel, F., Huber, P., Rogers, W., and Tukey, J. (1972). *Robust Estimates of Location: Survey and Advances.* Princeton University Press.

Atkinson, A. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*, 89:1329–1339.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data.* John Wiley.

Campell, N. (1980). Robust procedures in multivariate analysis I: Robust canonical variate analysis. *Applied Statistics*, 29:1–8.

Campell, N. (1982). Robust procedures in multivariate analysis II: Robust canonical variate analysis. *Applied Statistics*, 31:1–8.

Croux, C. and Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis.* in press.

Daudin, J., Duby, C., and Trecourt, P. (1988). Stability of principal component analysis studied by the bootstrap method. *Statistics*, 19:241–258.

Davies, P. (1987). Asymptotic behavior of s-estimators of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15:1269–1292.

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

Donoho, D. (1982). *Breakdown Properties of Multivariate Location Estimators.* PhD thesis, Harvard University, Department of Statistics.

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

Grübel, R. and Rocke, D. (1990). On the cumulants of affine equivariant estimators in elliptical families. *Journal of Multivariate Analysis*, 35:203–222.

Hadi, A. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society B*, 54:761–771.

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986). *Robust Statistics: The Approach Based on Influence Functions.* John Wiley.

Hawkins, D. (1980). *Identification of Outliers*. Chapman and Hall.

Hawkins, D. (1994). The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. *Computational Statistics and Data Analysis*, 17:197–210.

Hawkins, D. (1999). Improved feasible solution algorithms for high breakdown estimation. *Computational Statistics and Data Analysis*, 30:1–11.

Hawkins, D., Bradu, D., and Kass, G. (1984). Location of several outliers in multiple-regression data using elemental sets. *Technometrics*, 26:197–208.

Huber, P. (1981). *Robust Statistics*. John Wiley.

Kent, J. and Tyler, D. (1991). Redescending m-estimates of multivariate location and scatter. *The Annals of Statistics*, 19:2102–2119.

Lopuhaä, H. (1989). On the relation between s-estimators and m-estimators of multivariate location and covariance. *The Annals of Statistics*, 17:1662–1683.

Lopuhaä, H. (1992). Highly efficient estimators of multivariate location with high breakdown point. *The Annals of Statistics*, 20:398–413.

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

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis*. Academic Press.

Maronna, R. (1976). Robust m-estimators of multivariate location and scatter. *The Annals of Statistics*, 4:51–67.

Maronna, R. and Yohai, V. (1995). The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90:330–341.

Pearson, E. and Chandra Sekar, C. (1936). The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika*, 28:308–320.

Penny, K. (1995). Appropriate critical values when testing for a single multivariate outlier by using the mahalanobis distance. *Applied Statistics*, 45:73–81.

Rocke, D. (1996). Robustness properties of s-estimators of multivariate location and shape in high dimension. *The Annals of Statistics*, 24:1327–1345.

Rocke, D. and Woodruff, D. (1993). Computation of robust estimates of multivariate location and shape. *Statistica Neerlandica*, 47:27–42.

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

Rousseeuw, P. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In Grossmann, W., Pflug, G., Vincze, I., and Werz, W., editors, *Mathematical Statistics and Applications, Volume B*. Dordrecht:Reidel.

Rousseeuw, P. and Leroy, A. (1987). *Robust Regression and Outlier Detection*. John Wiley.

Rousseeuw, P. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223.

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

Rousseeuw, P. and Van Zomeren, B. (1991). Robust distances: Simulations and cutoff values. In Stahel, W. and Weisberg, S., editors, *Directions in Robust Statistics and Diagnostics Part 2*, pages 195–203. Springer-Verlag, New York.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2:110–114.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley.

Stahel, W. (1981). *Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen*. PhD thesis, ETH Zurich.

Tyler, D. (1983). Robust and efficiency properties of scatter matrices. *Biometrika*, 70:411–420.

Tyler, D. (1988). Some results on the existence, uniqueness, and computation of the m-estimates of multivariate location and scatter. *SIAM Journal on Scientific and Statistical Computing*, 9:354–362.

Tyler, D. (1991). Some issues in the robust estimation of multivariate location and scatter. In *Directions in Robust Statistics and Diagnositcs Part III*. Springer-Verlag.

Welch, B. (1937). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29:350–362.

Welch, B. (1947). The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34:28–35.

Wilks, S. (1962). *Mathematical Statistics*. John Wiley.

Chi-Square(p) cutoff values

		n			
		50	100	500	1000
p	5	19.8	13.5	7.1	6.1
	10	29.9	21.7	8.5	6.8
	20	26.8	32.5	12.3	8.5

Asymptotic cutoff values

		n			
		50	100	500	1000
p	5	0.14	1.4	4.4	4.8
	10	0.06	0.8	4.2	4.7
	20	0.01	0.4	3.6	4.4

Monte Carlo cutoff values

		n			
		50	100	500	1000
p	5	3.3	3.8	4.9	4.9
	10	1.9	3.2	4.8	4.9
	20	1.8	2.6	4.5	4.8

Table 1. Each entry represents the percent of simulated data that were above a specific 5% cutoff value. (Ideally, an entry in a cell would be 5.) The cutoff values were determined by dimension, size, and method of analysis. We can see that the Chi-Square cutoffs consistently reject too many points as outlying. The asymptotic method is quite conservative, but it appears to become more accurate as n increases. The simulation method is very good for medium to large samples, and it has the best performance of the three for small samples.

Chi-Square(p) cutoff values					
	n				
		50	100	500	1000
p	5	10.3	5.5	1.8	1.4
	10	21.0	10.3	2.4	1.6
	20	23.6	24.8	4.0	2.3

Asymptotic cutoff values					
	n				
		50	100	500	1000
p	5	0.0	0.1	0.8	0.9
	10	0.0	0.1	0.8	0.9
	20	0.0	0.0	0.6	0.8

Monte Carlo cutoff values					
	n				
		50	100	500	1000
p	5	0.4	0.6	0.9	1.0
	10	0.3	0.5	1.0	1.0
	20	0.4	0.4	0.8	1.0

Table 2. Each entry represents the percent of simulated data that were above a specific 1% cutoff value. (Ideally, an entry in a cell would be 1.) The cutoff values were determined by dimension, size, and method of analysis. Again, we see the same results, the Chi-Square cutoffs consistently reject too many points as outlying. The asymptotic method is quite conservative, but it appears to become quite accurate as n increases. The simulation method is very good for medium to large samples, and it has the best performance of the three for small samples.

Chi-Square(p) cutoff values					
	n				
p		50	100	500	1000
	5	4.48	1.69	0.29	0.19
	10	12.10	3.86	0.41	0.22
	20	19.82	15.17	0.78	0.35

Asymptotic cutoff values					
	n				
p		50	100	500	1000
	5	0.00	0.00	0.06	0.08
	10	0.00	0.00	0.07	0.09
	20	0.00	0.00	0.05	0.08

Monte Carlo cutoff values					
	n				
p		50	100	500	1000
	5	0.02	0.04	0.09	0.10
	10	0.02	0.03	0.10	0.10
	20	0.01	0.04	0.08	0.10

Table 3. Each entry represents the percent of simulated data that were above a specific 0.1% cutoff value. (Ideally, an entry in a cell would be 0.1.) The cutoff values were determined by dimension, size, and method of analysis. Again, we see the same results, the Chi-Square cutoffs consistently reject too many points as outlying. The asymptotic method is quite conservative, but it appears to become quite accurate as n increases. The simulation method is very good for medium to large samples, and it has the best performance of the three for small samples.

Figure Captions

1. Figure 1. Mahalanobis squared distances for the HBK data plotted against the χ_3^2 expected order statistics using the ordinary mean and covariance matrix. There are by construction 14 introduced outliers; these are masked when the mean and covariance are used to determine distances.

2. Figure 2. Mahalanobis squared distances for the HBK data plotted against the χ_3^2 expected order statistics using the MCD mean and covariance matrix. All 14 outlying points are clearly visible as outlying.

3. Figure 3. Mahalanobis squared distances for the Milk data plotted against the χ_8^2 expected order statistics using the MCD mean and covariance matrix. One outlier is apparent, but how many outlying points are there? One? Five? Six?

4. Figure 4. Mean Mahalanobis squared distances for simulated $(n = 100, p = 5)$ data plotted against the χ_5^2 expected order statistics using the MCD mean and covariance matrix. The points that are in the MCD sample appear to have a χ_5^2 distribution, but the points not included are definitely not distributed χ_5^2.

5. Figure 5. Mean Mahalanobis squared distances for simulated $(n = 500, p = 5)$ data plotted against the χ_5^2 expected order statistics using the MCD mean and covariance matrix. Again, the points that are in the MCD sample appear to have a χ_5^2 distribution, but the points not included, and especially the furthest outlying points, are not distributed χ_5^2. Even in large samples, there is still an elbow effect.

6. Figure 1. This figure illustrates the lack of dependence of extreme points on the MCD estimates. The distances for the dependent data set, the "o"'s, are calculated using the MCD estimates from the "o" data. Independent data are then simulated, the "+"'s, and the distances are calculated using the MCD estimates from the "o" data. For both sets of data, the points are averages of the ordered distances for 1000 repetitions of dimension 5 size 100 data. It is apparent that the extreme distances are not affected by whether the MCD was calculated using the same sample or a different one.

Figure 1: MSD using ordinary estimates for HBK Data

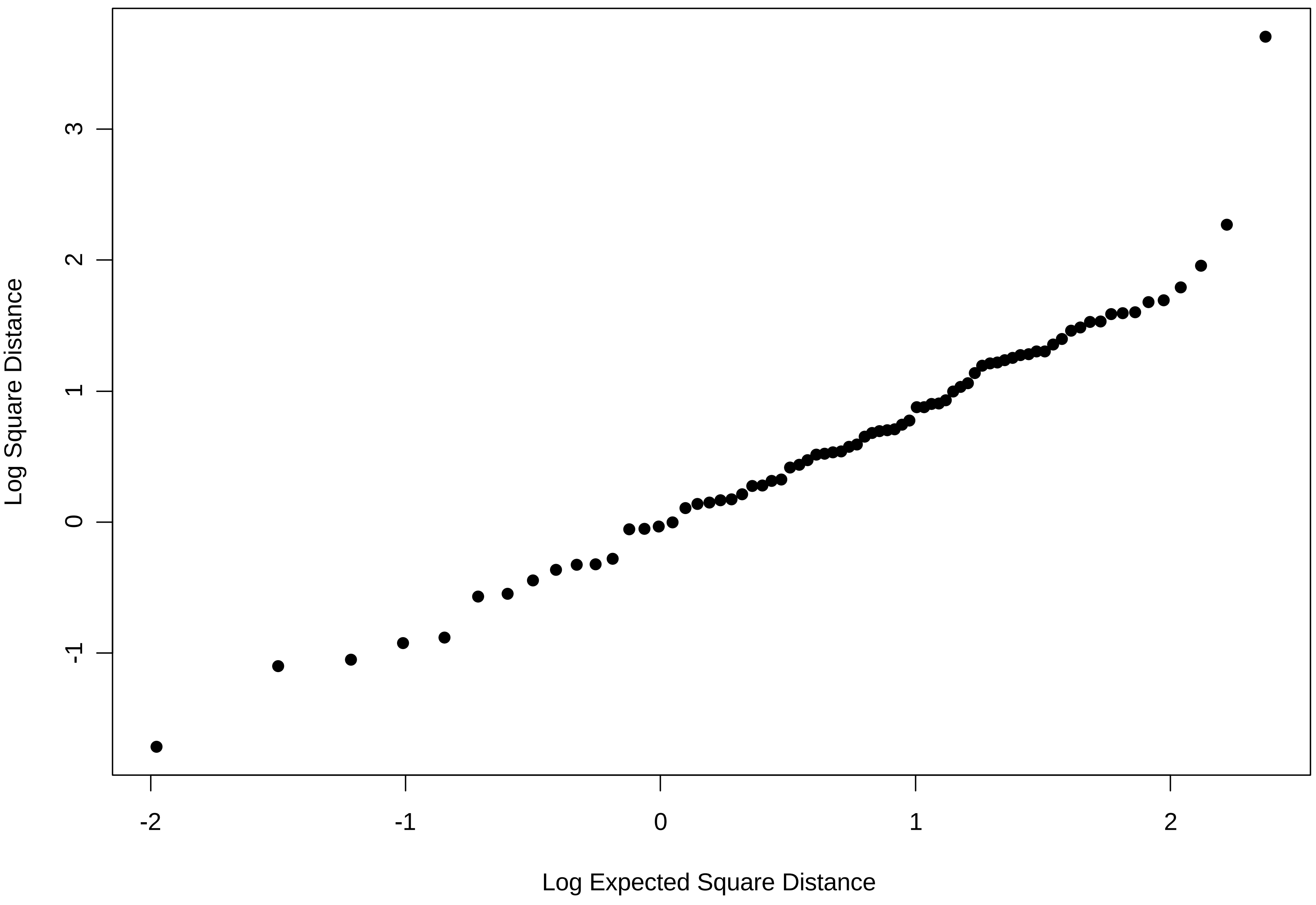


Figure 2: MSD using MCD for HBK Data

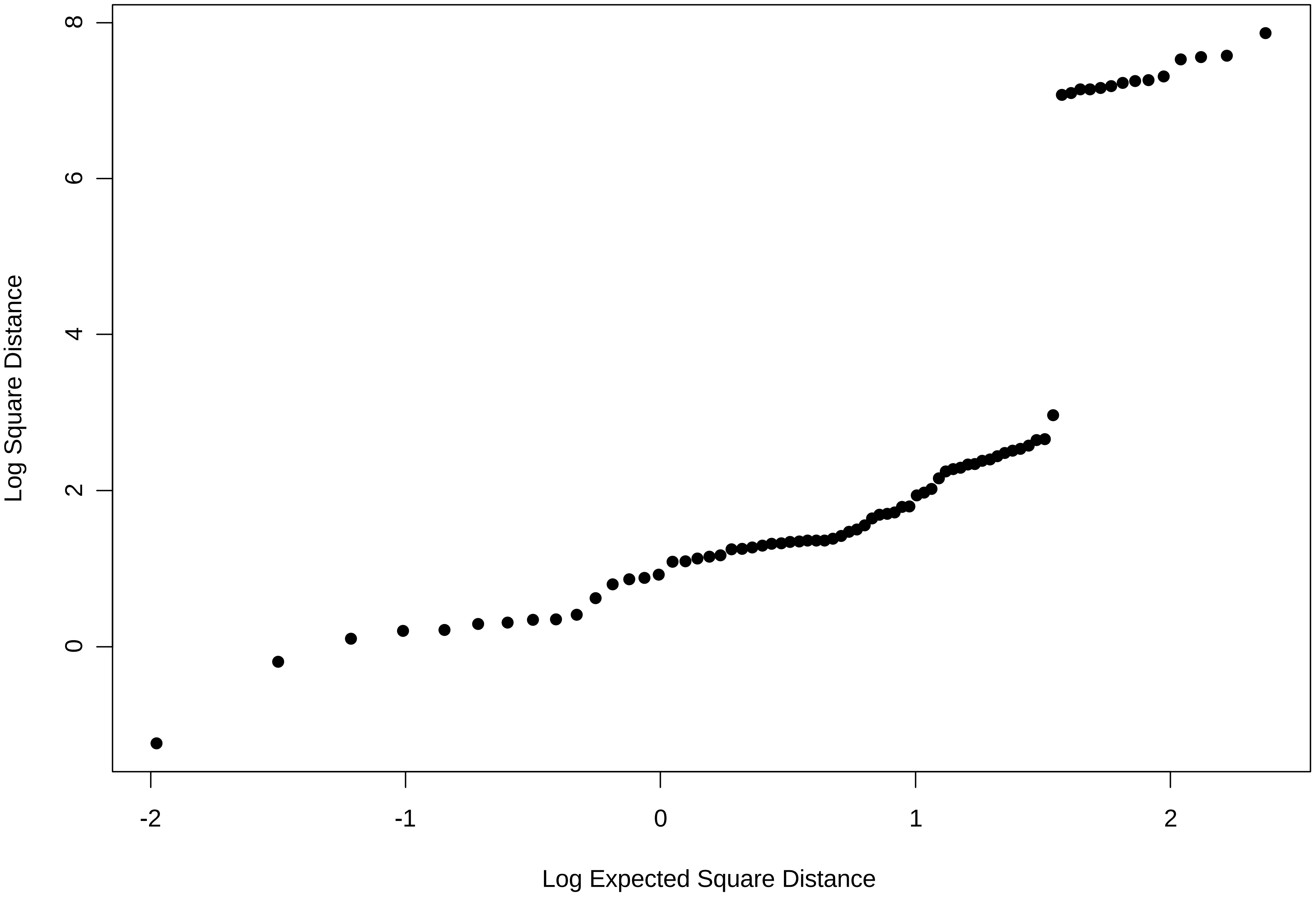


Figure 3: MSD using MCD for Milk Data

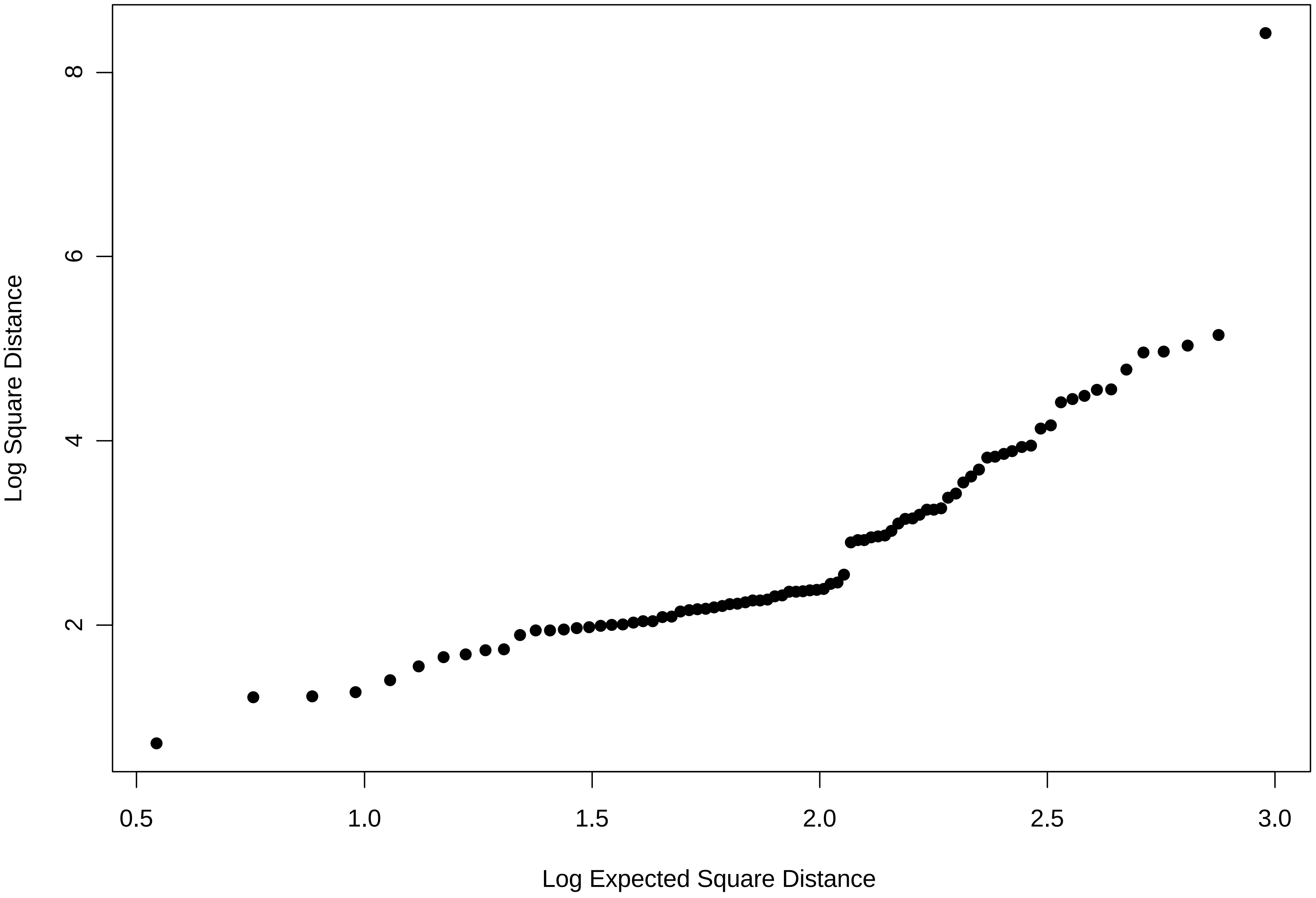


Figure 4: Elbow Effect for p=5 n=100

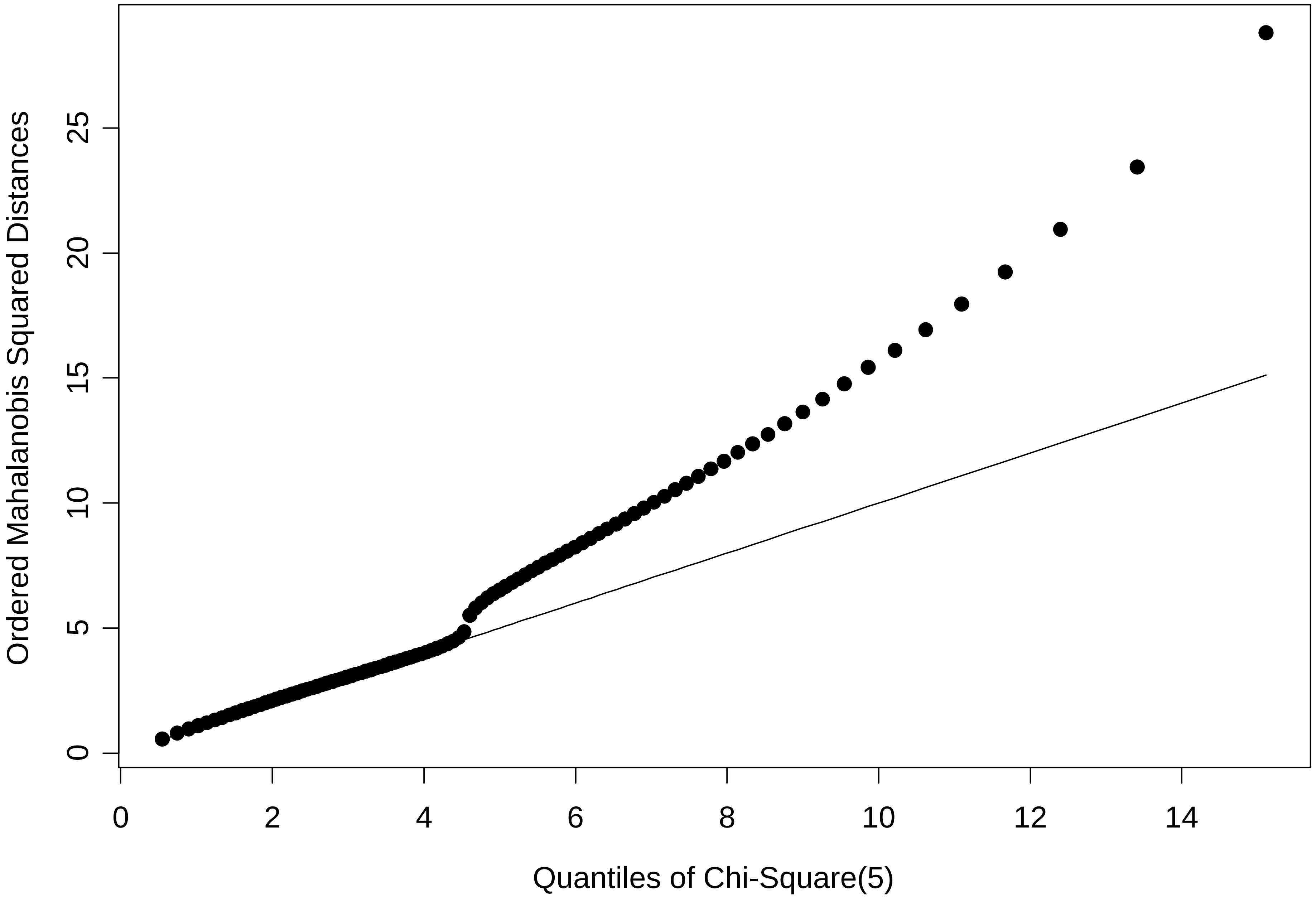


Figure 5: Elbow Effect for p=5 n=500

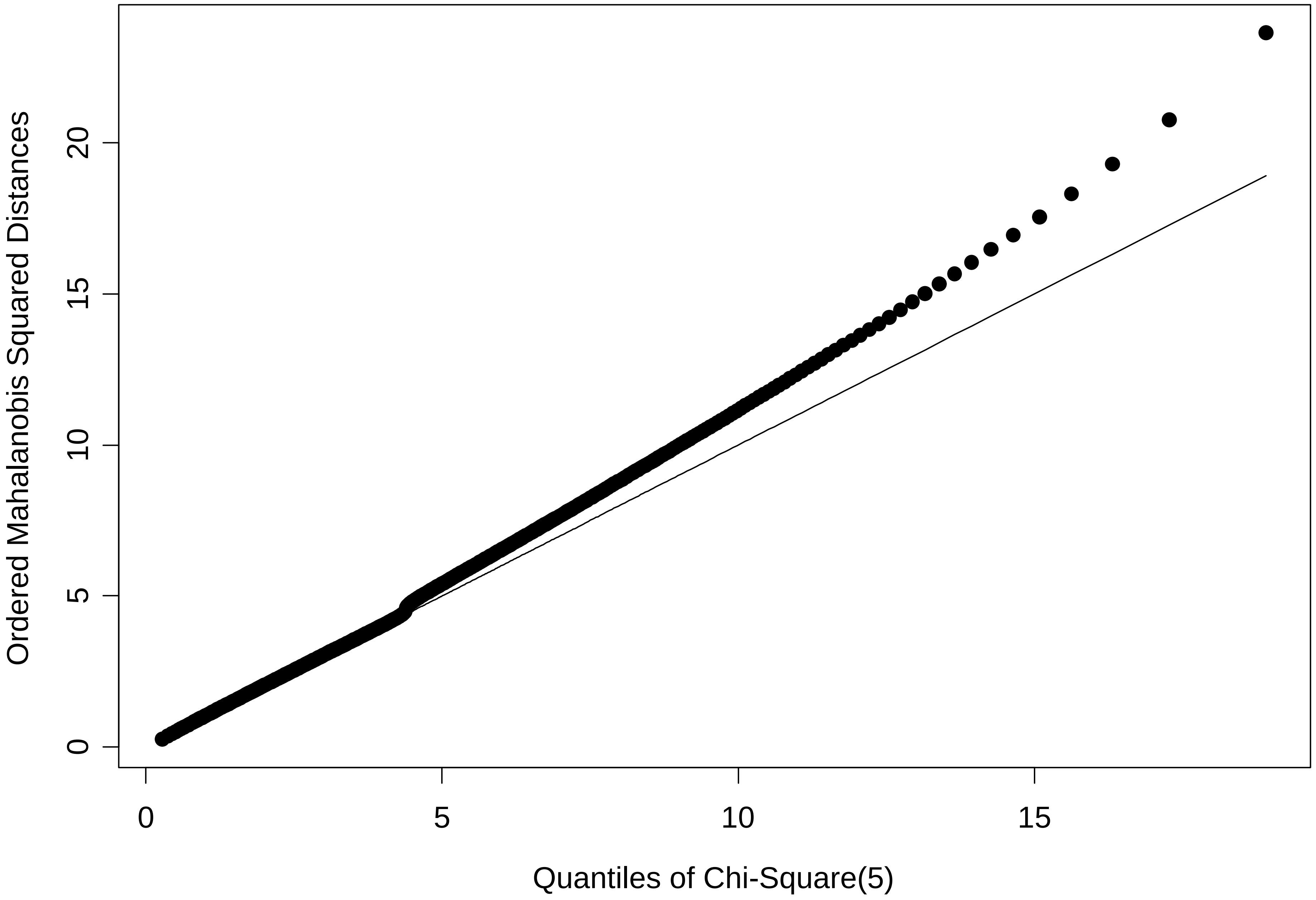
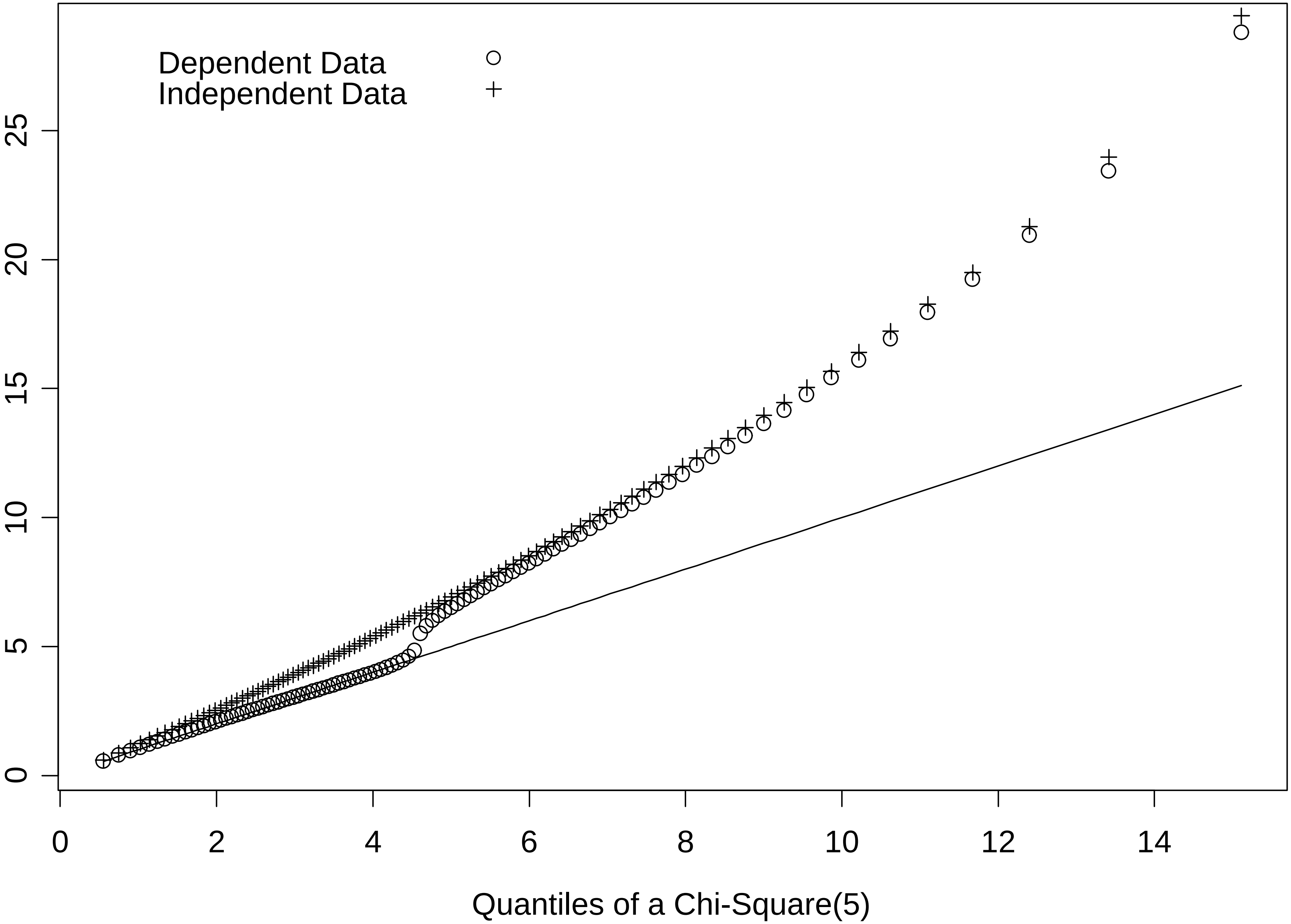


Figure 6: Independence of Distant Points and MCD sample, p=5 n=100


Dependent Data ○
Independent Data +
Quantiles of a Chi-Square(5)