ON THE BIASEDNESS OF THE KURTOSIS TEST FOR MULTIVARIATE NORMAL OUTLIERS*

by

Steven J. Schwager
Biometrics Unit, Cornell University, Ithaca, NY 14853

BU-818-M** May, 1983

Abstract

The general outlier problem is considered for a multivariate normal random sample with mean slippage. The locally best invariant test of the null hypothesis, that there are no outliers, versus the alternative hypothesis, that some outliers are present, is known to be: reject the null hypothesis whenever Mardia's multivariate sample kurtosis is sufficiently large. Conditions are given under which this test is biased, for fixed sample size n, asymptotically in the scalar slippage magnitude parameter. However, as n increases, these conditions cannot be met unless the fraction of outliers in the data approaches 21.13...%. This greatly diminishes the importance of the test's biasedness from an applied viewpoint. The bias of the kurtosis test is compared to other outlier tests.

---

* Paper presented in the Statistical Education Section of the American Statistical Association Summer Meetings, Toronto, Canada, August 1983.

** In the Biometrics Unit Series, Cornell University, Ithaca, NY 14853.

1. Introduction

Outliers are observations that depart so greatly from the pattern of the rest of the data that special treatment of the outliers should be considered. Model deficiencies, occasional gross measurement errors, a contaminated error distribution, and numerous other conditions can produce outliers in statistical data. An outlier, or aberrant observation, can be thought of as an observation so discrepant from the remaining data as to suggest that it may have been generated by a different underlying mechanism. Barnett and Lewis (1978), Beckman and Cook (1983), and Hawkins (1980) provide extensive surveys of the outlier literature.

Outliers in multivariate data constitute a more difficult problem than the univariate case, because of the great variety of ways in which an observation can be discordant (Gnanadesikan, 1977, page 271). For instance, an outlier can have a huge departure from the rest of the data in one component, or systematic small departures in all components, or a structure lying anywhere between these extremes. The covariance structure of the data must be utilized in assessing these departures, which may occur in any direction in p-dimensional space.

There is an important distinction between outlier detection and outlier identification. Outlier detection is the determination of whether outliers are present in the data; outlier identification is the specification of which observations are the discrepant ones.

The problem considered in this paper is outlier detection for data that, if outliers are absent, constitute a random sample from a multivariate normal $N(\mu,\Sigma)$ distribution. Any observation with a different distribution is an outlier. Under the mean slippage model treated here, all observations are independent normal random variables with common covariance matrix Σ; however, k of

the means, corresponding to k outliers, differ from the common mean μ of the rest, and perhaps from each other. In this situation, the outlier problem can be formulated as a test of the null hypothesis, that there are no outliers, versus the alternative hypothesis, that some outliers are present in the data. The locally best invariant test is (Schwager and Margolin, 1982): reject the null hypothesis whenever Mardia's multivariate sample kurtosis $b_{2,p}$ is sufficiently large.

The main result of this paper is that the test based on Mardia's multivariate sample kurtosis is biased. For fixed sample size n, the test remains biased as the scalar parameter reflecting the magnitude of departure, or slippage, from the null hypothesis increases without bound. This asymptotic bias is similar to, but much less serious than, the asymptotic bias caused by the masking effect in studentized residual-based outlier procedures. This is because the asymptotic bias of the kurtosis test occurs only when the proportion of outliers in the data is large. For instance, with 50 observations, the test is asymptotically unbiased whenever there are eight or less outliers in any spatial configuration; with 100 observations, whenever there are 18 or less outliers; and so on.

The larger n is, the closer to 21.13...% the proportion of outliers must be to make the kurtosis test biased asymptotically in the slippage magnitude parameter. This greatly diminishes the importance of the test's biasedness from an applied viewpoint.

2. The General Outlier Problem

This section is drawn from Schwager and Margolin (1982), where more details are provided.

A multivariate normal random sample can be specified by the matrix equation $Y = e\mu + U$, where the $n \times p$ data matrix Y has i.i.d. rows $Y_1, \cdots, Y_n$, e is

an $n \times 1$ vector of 1's, μ is the unknown $1 \times p$ mean vector, and the rows of the $n \times p$ error matrix U are i.i.d. $N(0,\Sigma)$ with unknown covariance matrix Σ. It is assumed that $n \geq p+1$ to insure that μ and Σ are estimable.

For any matrix $A = (a_{ij})$, define $\|A\| = (\Sigma_{ij} a_{ij}^2)^{\frac{1}{2}}$. Embed the multivariate normal random sample model in the multivariate mean model with mean slippage,

$$Y = e\mu + \Delta A + U \quad . \tag{2.1}$$

The terms e, μ, and U are as above, and $n \geq p+1$; in addition, Δ is a nonnegative scalar and A is an arbitrary $n \times p$ matrix with rows $a_1, \cdots, a_n$ such that: (c1) $\|A\| = 1$ unless $\Delta = 0$, in which case $A = 0$; and (c2) more than half of the rows of A are 0. In this model, the observation Y_i is an outlier if the ith row of A is nonzero. This model extends a model for univariate outliers due to Ferguson (1961).

There are no outliers present if and only if $\Delta = 0$. By (c2), more than half of the observations are drawn from the $N(\mu,\Sigma)$ population; by (c1), (c2), and the nonnegativity of Δ, the parametrization is unique. The general outlier problem consists of model (2.1); hypotheses $H_0 : \Delta = 0$ and $H_1 : \Delta > 0$; action space $\mathcal{A} = \{D_0, D_1\}$, where D_i is the decision to act as if hypothesis H_i is true, $i = 0,1$; state space $\Theta = \{(\Delta, A, \mu, \Sigma) : \Sigma > 0,\ \Delta \geq 0;\ (c1),(c2)\ \text{hold}\}$; and loss function L with $L(\theta, D_i) = i$ if $\Delta = 0$, $L(\theta, D_i) = 1 - i$ if $\Delta > 0$.

The general outlier problem is invariant under the group G of transformations of the form $g(Y) = \Gamma YC + ec$ where Γ is an $n \times n$ permutation matrix, C is a $p \times p$ nonsingular matrix, and c is a $1 \times p$ vector. Because of this invariance, only decision rules invariant under G will be considered.

Mardia's multivariate sample kurtosis of an $n \times p$ matrix Y is defined (1974) as

$$b_{2,p}(Y) \equiv b_{2,p} = n\Sigma_{i=1}^{n}[(Y_i - \bar{Y})S^{-1}(Y_i - \bar{Y})']^2 \quad ,$$

where $\bar{Y} = (1/n)\Sigma_{i=1}^{n} Y_i$ and $S = (Y - e\bar{Y})'(Y - e\bar{Y})$. The locally best invariant test for the general outlier problem is based on the test statistic $b_{2,p}$. Whenever the proportion of nonzero rows of A is at most $(3 - \sqrt{3})/6 \approx 21\%$, the locally best invariant test rejects H_0 if Mardia's multivariate sample kurtosis is too large.

Theorem 2.1: For the general outlier problem, let $a_i = 0$ for $i = m+1, \cdots, n$, where $m/n \leq (3 - \sqrt{3})/6 = .2113\cdots$. The locally best invariant test of $H_0 : \Delta = 0$ versus $H_1 : \Delta > 0$ is: reject H_0 whenever $b_{2,p} \geq K$, where the constant K is determined by the size α of the test, n, and p. This test is locally best invariant uniformly in $(a_1, \cdots, a_m)$.

The local behavior of a test at $\Delta = 0$ is fully specified by the derivatives of its power function $\beta(\Delta)$ at $\Delta = 0$. The locally best test of $\Delta = 0$ versus $\Delta > 0$ can be obtained by maximizing the kth derivative $\beta^{(k)}(0)$ over the class of α-level tests, where k is the smallest positive integer such that $\beta^{(k)}(0)$ is not identically zero for all α-level tests. For the general outlier problem, $k = 4$. This is even, so the locally best invariant test is also locally best unbiased invariant whenever it is unbiased.

The test of Theorem 2.1 clearly satisfies the local requirements for a locally best unbiased invariant test. In addition, the global issue of unbiasedness must be addressed. Somewhat surprisingly, it is resolved by the result that this test is biased.

The invariance structure in Ferguson's (1961) treatment of the univariate case is slightly different from the structure used here. Ferguson's invariance was under permutations of the observations $y_1, \cdots, y_n$ and under transformations of every y_i to $cy_i + d$ with constants $c > 0$ and d arbitrary; the univariate case of the invariance structure used here is the same except that any constant $c \neq 0$ is allowed. In spite of this difference, the results of later sections of this

paper apply to Ferguson's situation, as minor adjustments to the exposition easily establish. Thus, the claim that a locally best unbiased invariant test is given by rejecting H_0 when the (univariate) kurtosis $b_2 \geq K$ (Ferguson, 1961, Theorem 2.2) is incorrect. This test, as will be seen, is locally unbiased but not globally unbiased.

3. Bias of the Locally Best Invariant Test

Two lemmas give useful information about the subsets of the sample space on which $b_{2,p}$ assumes certain values.

Lemma 3.1: Let μ, Σ, n, p, and the size α of the test assume arbitrary values. The test of Theorem 2.1 rejects H_0 whenever $b_{2,p} \geq K$, where the critical value K is a function of α, n, and p. Assume that A satisfies $b_{2,p}(A) < K$, where $b_{2,p}(A)$ denotes the multivariate sample kurtosis computed from the matrix A. Fix all of these parameters, so $\Pr[b_{2,p}(Y) \geq K]$ is a function only of Δ. Then as $\Delta \to \infty$,

$$\Pr[b_{2,p}(Y) \geq K] \to 0 \quad .$$

Proof: The sample $Y = e\mu + \Delta A + U$ is a function only of U and Δ. To examine the behavior of $b_{2,p}(Y)$, viewed as a function of U and Δ, choose ρ such that

$$b_{2,p}(A) < \rho < K \quad .$$

Observing the identity

$$b_{2,p}(e\mu + \Delta A + U) = b_{2,p}[A + (1/\Delta)U] \quad ,$$

we denote the common value by $b(\Delta, U)$.

The space $\mathcal{U}$ of all possible values of U is np-dimensional Euclidean space, a linear metric space. Take as the metric, denoted by d, any norm, for example Euclidean distance on np-space. Sets of measure zero, consisting of points where $b_{2,p}$ is not defined, will be understood to be omitted as necessary throughout this proof.

For $\Delta > 0$, define the sets $N(\Delta)$ and $M(\Delta)$ by

$$N(\Delta) = \{U: b(\Delta,U) < \rho\} ,$$
$$M(\Delta) = \{U: b(\beta,U) < \rho \text{ for all } \beta \geq \Delta\} = \bigcap_{\beta \geq \Delta} N(\beta) . \tag{3.1}$$

It is clear from the continuity of $b_{2,p}$ that $N(\Delta)$ is an open set for every positive Δ. The family of sets $N(\Delta)$ converges to $\mathcal{U}$ as $\Delta \to \infty$, for

$$\mathcal{U} = \bigcap_{\Delta>0} \bigcup_{\beta \geq \Delta} N(\beta) = \bigcup_{\Delta>0} \bigcap_{\beta \geq \Delta} N(\beta) .$$

To show this, it suffices to demonstrate that $\bigcup \bigcap N(\beta) = \mathcal{U}$, since it is immediate that

$$\bigcup \bigcap N(\beta) \subset \bigcap \bigcup N(\beta) \subset \mathcal{U} .$$

For any $U \in \mathcal{U}$, $b(\beta,U)$ approaches $b(\infty,U)$, i.e. $b_{2,p}(A)$, as β increases without bound. Thus, for any U, there is a value of Δ such that

$$b(\beta,U) < \rho \text{ for all } \beta \geq \Delta .$$

Consequently,

$$\bigcup_{\Delta>0} \bigcap_{\beta \geq \Delta} N(\beta) = \bigcup_{\Delta>0} \{U: b(\beta,U) < \rho \text{ for all } \beta \geq \Delta\} = \mathcal{U} .$$

This also shows that $\mathcal{U} = \bigcup_{\Delta>0} M(\Delta)$.

It is proven below, in Lemma 3.2, that $M(\Delta)$ is an open set for every positive Δ, so the family $\{M(\Delta): \Delta > 0\}$ is an open covering of $\mathcal{U}$. This family is an expanding collection of sets, since $\Delta < \beta$ implies that $M(\Delta) \subset M(\beta)$, so the family converges to $\mathcal{U}$ as Δ approaches ∞. The Heine-Borel Theorem guarantees that any closed, bounded subset of $\mathcal{U}$ has a finite subcover, which can here be reduced to a single set $M(\Delta)$.

For any $\delta > 0$, there exists a closed, bounded set $F \subset \mathcal{U}$ with $Pr(F) > 1 - \delta$. The covering argument shows the existence of a Δ for which $F \subset M(\Delta)$. So

$$1 - \delta < Pr(F) \leq Pr[M(\Delta)] = Pr[\{U: b(\beta,U) < \rho \text{ for all } \beta \geq \Delta\}] .$$

Taking complements,

$$\delta > Pr[\{U: b(\beta,U) \geq \rho \text{ for some } \beta \geq \Delta\}]$$
$$\geq Pr[\{U: b_{2,p}(e\mu + \beta A + U) \geq K \text{ for some } \beta \geq \Delta\}] .$$

It follows that, since $b_{2,p}(A)$ is less than K, given any positive δ, however small, there exists a corresponding Δ_δ for which

$$Pr[b_{2,p}(Y) \geq K | \Delta \geq \Delta_\delta] < \delta . \quad \text{QED}$$

Lemma 3.2: For any $\Delta > 0$, the set $M(\Delta)$ defined in (3.1) is open. Furthermore, the following relations hold for all Δ:

$$N(c\Delta) = cN(\Delta) = \{cU: U \in N(\Delta)\} \quad \text{for any } c > 0 ;$$
$$M(\Delta) = \{U: cU \in N(\Delta) \text{ for all } c \text{ in } [0,1]\} . \tag{3.2}$$

Proof: The first part of (3.2) is immediate upon noting that $b(\Delta,U) = b(c\Delta,cU)$ for any positive c. For the second part, $U \in M(\Delta)$ if and only if $U \in N(c'\Delta) = c'N(\Delta)$ for every $c' \geq 1$. Equivalently, $(1/c')U \in N(\Delta)$ for all $c' \geq 1$. And $0(n \times p) = 0 \times U \in N(\Delta)$ for all Δ and U. So $U \in M(\Delta)$ if and only if $cU \in N(\Delta)$ for all $c \in [0,1]$.

Take $U \in M(\Delta)$. We will construct a neighborhood $H(U) \subset M(\Delta)$. The set $N(\Delta)$ is open, and $cU \in N(\Delta)$ for c in [0,1], so there corresponds to each c a neighborhood of cU, denoted by G(cU), which lies with $N(\Delta)$. Define $L = \{cU: c \in [0,1]\}$, and observe that it is a closed, bounded

subset of $\mathcal{U}$. The family $\{G(cU): c \in [0,1]\}$ is an open cover of L, so the Heine-Borel theorem shows the existence of a finite subcover $\{G(c_iU): i = 1, \ldots, q\}$. Let $\hat{G}$ denote the set $\bigcup_{i=1}^{q} G(c_iU)$, and $Bd(\hat{G})$ the boundary of $\hat{G}$. Since $\hat{G}$ is an open set, the distance between its boundary and L must be greater than zero. Calling this distance r, we see

$$r = d[L, Bd(\hat{G})] > 0 \quad .$$

It is an exercise in analysis to verify that: for all $V, W \in \mathcal{U}$, if $W \in L$ and $d(V,W) < r$, then $V \in \hat{G} \subset N(\Delta)$.

Define the neighborhood H(U) to be the open ball of radius r about U,

$$H(U) = \{V: d(U,V) < r\} \quad .$$

Choose any member V of H(U) . The metric d is a norm, so for any $c \in [0,1]$,

$$d(cV, cU) = cd(V,U) < cr \leq r, \quad \text{and} \quad cU \in L \ ,$$

so $cV \in \hat{G} \subset N(\Delta)$. In other words, $cV \in N(\Delta)$ for all $c \in [0,1]$, so $V \in M(\Delta)$ by the second part of (3.2). QED

Theorem 3.1: For the multivariate mean model, mean slippage general outlier problem, assume that the fraction of outliers is at most $(3 - \sqrt{3})/6$. The locally best invariant test of $H_0: \Delta = 0$ versus $H_1: \Delta > 0$, rejecting H_0 whenever $b_{2,p}(Y) \geq K$, is biased.

Proof: The test is unbiased if and only if for all α, n, p, and $\theta = (\Delta, A, \mu, \Sigma) \in \Theta$,

$$Pr[b_{2,p}(Y) \geq K|\theta] \geq Pr[b_{2,p}(Y) \geq K|\Delta = 0] = \alpha \quad .$$

To construct an example for which this inequality does not hold, take $\alpha = .05$, $n = 200$, and $p = 1$, so $K = 3.57$ (Pearson and Hartley, 1966, p. 208). Choose $A(200 \times 1)$ to consist of 40 entries equal to 1, the rest equal to 0; then $b_{2,1}(A) = 3.25 < K$.

Now Lemma 3.1 shows that, as all parameters except Δ remain fixed, and $\Delta \to \infty$, $\Pr[b_{2,p}(Y) \geq K] \to 0$. So there exists $\theta \in \Theta$ for which $\Pr[b_{2,p}(Y) \geq K|\theta] < \alpha$. QED

Lemma 3.1 allows a much stronger conclusion than the biasedness of the locally best invariant test. Whenever $b_{2,p}(A) < K$, the power of the test goes to zero as $\Delta \to \infty$. This is quite plausible, since for any sample Y, the term ΔA dominates the other terms in $Y = e\mu + \Delta A + U$ when $\Delta \to \infty$. As a result, Y approaches a multiple of A, and $b_{2,p}(Y)$ approaches $b_{2,p}(A)$. It seems natural that $\Pr[b_{2,p}(Y) \geq K] \to 0$ under these conditions.

4. Asymptotic Bias of the Locally Best Invariant Test

A situation of special interest occurs as $n \to \infty$ while $b_{2,p}(A)$ remains fixed. This would happen, for example, if observations were taken in equal-sized groups or blocks, with all groups corresponding to sub-matrices of A that are identical up to a row permutation. The results to be given are for the case $p = 1$, so the standard symbol b_2 will be used for $b_{2,1}$. It will be shown that $b_2(A) > 3$ whenever the fraction of nonzero rows of A is less than $(3 - \sqrt{3})/6$.

Theorem 4.1: Restrict the collection of real numbers $x_1, \cdots, x_n$ by the condition $x_{k+1} = \cdots = x_n = 0$, so $\bar{x} = (1/n) \sum_{i=1}^{k} x_i$. The kurtosis of this collection is defined by

$$b_2(x_1, \cdots, x_k) = n \frac{\Sigma_1^k (x_i - \bar{x})^4 + (n-k)\bar{x}^4}{[\Sigma_1^k (x_i - \bar{x})^2 + (n-k)\bar{x}^2]^2} \quad . \tag{4.1}$$

Assume that $x_i \neq 0$ for some i, and that $f = k/n < \frac{1}{3}$. Then b_2 takes on its minimum value when $x_1 = x_2 = \cdots = x_k$, and this value is

$$\min_x b_2(x_1, \cdots, x_k) = 1/f(1-f) - 3 \quad .$$

Proof: Define for any positive integer j the quantity

$$S_j = \sum_{i=1}^{n} (x_i - \bar{x})^j \quad .$$

Then $S_1 = 0$, and the skewness and kurtosis of the collection $x_1, \cdots, x_n$ can be expressed as

$$\sqrt{b_1} = n^{1/2} S_3 / S_2^{3/2}; \quad b_2 = n S_4 / S_2^2 \quad . \tag{4.2}$$

The partial derivative of b_2 with respect to x_i, for $i = 1, \cdots, k$, is

$$\frac{\partial b_2}{\partial x_i} = 4n S_2^{-2}[(x_i - \bar{x})^3 - (S_4/S_2)(x_i - \bar{x}) - S_3/n] \quad .$$

Thus, at an extremum, the equation

$$y^3 - (S_4/S_2)y - S_3/n = 0 \tag{4.3}$$

must be satisfied by $y = x_1 - \bar{x}, \cdots, x_k - \bar{x}$. Substituting each of these roots into (4.3) and summing the k resulting equations gives

$$[S_3 + (n-k)\bar{x}^3] - (S_4/S_2)(n-k)\bar{x} - kS_3/n = 0 \quad ,$$

which simplifies to

$$(-\bar{x})^3 - (S_4/S_2)(-\bar{x}) - S_3/n = 0 \quad .$$

We see from this that $-\bar{x}$ is also a root of (4.3), and recall that, for $i = k+1, \cdots, n$, $x_i - \bar{x} = -\bar{x}$.

The cubic equation (4.3) has three unequal real roots whose sum is

zero. To establish this claim, we compute the discriminant and use (4.2) to obtain

$$D = \left(-\tfrac{1}{3}\,\frac{S_4}{S_2}\right)^3 + \left(\tfrac{1}{2}\,\frac{S_3}{n}\right)^2 = (S_2/n)^3\,[b_1/4 - b_2^3/27] \quad .$$

It is an identity (see Kendall and Stuart, 1969, p. 92, ex. 3.19) that $b_1 < b_2 - 1$, so

$$b_1/4 - b_2^3/27 < g(b_2) \quad \text{where} \quad g(u) = -u^3/27 + u/4 - 1/4 \quad .$$

It is easily verified by elementary calculus that $g(u) \leq 0$ for all $u \geq 0$, which implies that

$$b_1/4 - b_2^3/27 < 0 \quad .$$

Therefore D is negative, so the three roots of (4.3) are distinct real numbers. They must sum to zero because the coefficient of y^2 in (4.3) is 0 .

We now investigate the possible configurations of the three roots of (4.3), and proceed to express b_2 as a function of the fractions of the collection $\{x_i - \bar{x}\}$ taking each value. There are only three patterns the roots may follow:

(i) $-a - d,\ d,\ a,$ where $0 < d < a$,

(ii) $-a,\ 0,\ a,$ where $0 < a$,

(iii) $-a,\ -d,\ a + d,$ where $0 < d < a$.

Assume case (ii). Let k_1, k_2, and k_3 be integers such that, as i ranges from 1 to $k = k_1 + k_2 + k_3$,

$$x_i - \bar{x} = \begin{Bmatrix} -a \\ 0 \\ a \end{Bmatrix} \quad \text{with frequency} \quad \begin{Bmatrix} k_1 \\ k_2 \\ k_3 \end{Bmatrix} \quad .$$

Then

$$\sum_{i=1}^{k} (x_i - \bar{x}) = (n-k)\bar{x} = (k_3 - k_1)a \quad .$$

So $-\bar{x} = (k_1 - k_3)a/(n-k)$. Since $-\bar{x}$ is a root of (4.3), it must be -a, 0, or a . In fact, it must be 0, and thus $k_1 = k_3$, for

$$|k_1 - k_3| \leq k < n - k \quad .$$

Letting f denote the fraction $(k_1 + k_3)/n$, we can compute b_2 from (4.1) as

$$b_2 = n \frac{\Sigma_1^k x_i^4}{[\Sigma_1^k x_i^2]^2} = 1/f \quad .$$

We conclude that, if $x_1, \ldots, x_k$ gives root pattern (ii) for (4.3), or equivalently if $S_3 = 0$, then $b_2 = 1/f$, where nf is the number of non-zero x_i's .

Now assume pattern (i). The analysis of this case also applies to pattern (iii) after an obvious sign change, so it suffices to treat (i). Let k_1, k_2, k_3 be integers summing to k, such that, as i ranges from 1 to k,

$$(4.4) \qquad x_i - \bar{x} = \begin{Bmatrix} a \\ d \\ -a-d \end{Bmatrix} \text{ with frequency } \begin{Bmatrix} k_1 \\ k_2 \\ k_3 \end{Bmatrix} \quad .$$

Then

$$(n-k)\bar{x} = \sum_{i=1}^{k} (x_i - \bar{x}) = k_1 a + k_2 d + k_3(-a-d) \quad ,$$

so

$$(4.5) \qquad -\bar{x} = \frac{k_3 - k_1}{n-k} a + \frac{k_3 - k_2}{n-k} d \quad .$$

Being a root of (4.3), this must equal a, d, or -a - d . The following paragraph shows that a and -a - d cannot be equal to $-\bar{x}$, proving that

$-\bar{x} = d$.

If $-\bar{x} = a$, (4.5) implies that

$$(n - k_2 - 2k_3)a + (k_2 - k_3)d = 0 \quad ;$$

but this is not possible, for $|d| < |a|$ and

$$|k_2 - k_3| \leq k < n - 2k \leq |n - k_2 - 2k_3| \quad .$$

Similarly, if $-\bar{x} = -a - d$, it follows that

$$(n - 2k_1 - k_2)a + (n - k_1 - 2k_2)d = 0 \quad .$$

But the left-hand side is a sum of products of positive terms, so this too is a contradiction, and $-\bar{x} = d$ must hold.

Several useful facts emerge from this conclusion. By (4.5),

$$0 < d/a = (k_3 - k_1)/(n - k_1 - 2k_3) < 1 \quad , \tag{4.6}$$

so $k_3 > k_1$. It is not surprising that the parameter k_2 fails to appear, for $x_i - \bar{x} = d$ if and only if $x_i = 0$. From the definition (4.1) of b_2, (4.4), and our knowledge that $\bar{x} = -d$,

$$b_2 = n \frac{k_1 a^4 + k_3(a + d)^4 + (n - k_1 - k_3)d^4}{[k_1 a^2 + k_3(a + d)^2 + (n - k_1 - k_3)d^2]^2} \quad .$$

Define

$$f = (k_1 + k_3)/n, \quad t = k_3/n, \quad \text{and} \quad g = 1 - f \quad ,$$

so f is the fraction of x's which are non-zero (taking the values $a - d$ and $-a - 2d$), and t the fraction which are negative (taking the value $-a - 2d$) . Using the expression for d/a given in (4.6), b_2 can be expressed as a function of f and t:

$$b_2 = \frac{(f - t)(1 - f - t)^4 + t(1 - 2f + t)^4 + (1 - f)(2t - f)^4}{[(f - t)(1 - f - t)^2 + t(1 - 2f + t)^2 + (1 - f)(2t - f)^2]^2} \quad .$$

The polynomials in f and t are of distressing complexity. Fortunately, the numerator turns out to be divisible by the expression in square brackets in the denominator, and after much tedious algebraic simplification we arrive at

$$b_2 = \frac{3t^2 - 3ft + (g^2 - fg + f^2)}{(8g - f)(t^2 - ft) + fg}$$

$$= \frac{3}{8g - f} + \frac{8g^3 - 12fg^2 + 6f^2g - f^3}{8g - f} \cdot \frac{1}{(8g - f)(t^2 - ft) + fg} \quad .$$

For any fixed proportion f of non-zero x's, we wish to find t on $(f/2, f]$ which minimizes $b_2(f,t)$. The condition $t > f/2$ is a consequence of $k_3 > k_1$. It is easy to show from $f < \frac{1}{3}$ that $8g - f$ and $8g^3 - 12fg^2 + 6f^2g - f^3$ are positive. Consider the quadratic polynomial in t, $t^2 - ft + fg/(8g - f)$. Its discriminant is

$$D = f^2 - 4fg/(8g - f) = -f(2g - f)^2/(8g - f) < 0 ,$$

so it has no real roots, and thus is positive for all real t . Its minimum occurs at $t = f/2$, so it is obviously increasing on $(f/2, f]$. Therefore, it is maximized at $t = f$, and

$$\min_t b_2(f,t) = b_2(f,f) = (1 - 3f + 3f^2)/f(1 - f)$$
$$= 1/f(1 - f) - 3 .$$

This is a decreasing function of f on $(0, \frac{1}{3})$; so the higher the fraction of non-zero x's, the lower the minimum kurtosis, which is obtained (at $t = f$) when all of the non-zero x's are equal. The proof is completed by noting that $1/f(1 - f) - 3$ is less than $1/f$, the kurtosis found with root pattern (i), on $(0, \frac{1}{3})$. QED

Corollary 4.1: If the fraction f of non-zero rows of $A(n \times 1)$ is less than $(3-\sqrt{3})/6$, then $b_2(A) > 3$.

Proof: From Theorem 4.1, for any value of $f = k/n$,

$$\min_x b_2(x_1, \cdots, x_k) = 1/f(1-f) - 3 \quad .$$

This function is decreasing on the interval $(0, \frac{1}{3})$, and its value at $f = (3-\sqrt{3})/6$ is 3 . Since A is a set of n numbers, at most k of which are non-zero, for $f < (3-\sqrt{3})/6$ we have

$$b_2(A) \geq 1/f(1-f) - 3 > 3 \quad . \qquad \text{QED}$$

As mentioned earlier, the critical value K of the test of Theorem 2.1 is a function of the size α, n, and p . As n increases with p and α fixed, $K \to p(p+2)$, for

$$[b_{2,p} - p(p+2)]/[8p(p+2)/n]^{\frac{1}{2}}$$

is asymptotically N(0,1) (Mardia, 1970). For $p = 1$, K approaches 3 as n increases without bound. Consequently, if the fraction of non-zero rows of $A(n \times 1)$ is bounded below $(3-\sqrt{3})/6$, $b_2(A)$ will be bounded above 3, and for sufficiently large n, $K < b_2(A)$. Thus, for large enough n, the situation that caused the test to be biased as $\Delta \to \infty$, namely $b_2(A) < K$, cannot arise.

In Table 1, values of K are given for various sample sizes and $p = 1, 2$, with $\alpha = .05$. The entries for $p = 1$ are taken from Pearson and Hartley (1966, p. 208), those for $p = 2$ from Mardia (1974).

		Sample Size (n)				
		50	100	200	1000	5000
Dimension	1	3.99	3.77	3.57	3.26	3.12
(p)	2	9.453	9.210	8.919	8.419	8.186

Table 1: Critical Values K for which $Pr[b_{2,p} \geq K|H_0] = \alpha = .05$.

5. A Comparison of Outlier Rules

A commonly used test for outliers in univariate normal data is based on the extreme studentized residual. It rejects H_0 if $ESR \equiv \max_i |x_i - \bar{x}|/s \geq K$, where $s^2 = \Sigma(x_j - \bar{x})^2/(n-1)$. This test can be generalized easily to the multivariate case (see, for example, Karlin and Truax, 1960). It can be, and often is, applied sequentially if multiple outliers are suspected.

Consider the case in which the fraction f of the n observations are distributed $N(\mu + \Delta, \sigma^2)$, while the remaining observations are distributed $N(\mu, \sigma^2)$. The univariate case is treated to take advantage of existing tables. As Theorem 4.1 shows, this arrangement of the $m = nf$ outliers is least favorable to the outlier detection test based on $b_{2,p}$, since it results in a lower sample kurtosis as $\Delta \to \infty$ than any other arrangement of m outliers. It will be shown that the $b_{2,p}$ outlier detection test is asymptotically unbiased as $\Delta \to \infty$ over a greater range of f than the ESR test, where in fact $Pr[\text{reject } H_0|\Delta] \to 1$ as $\Delta \to \infty$. The $b_{2,p}$ test is able to detect values of f approaching 21.13% as n becomes larger, while the maximum f that the ESR test can detect decreases to much lower values as n becomes larger.

Let t denote the $1 - \alpha/2n$ quantile of a t distribution with $n - 2$ degrees of freedom, and define

$$C = t(n-1)/[n(n-2+t^2)]^{\frac{1}{2}} \quad .$$

Then C approximates the $1-\alpha$ quantile of ESR (Hawkins, 1980, p. 136). It is routine to show that as $\Delta \to \infty$,

$$\lim_{\Delta \to \infty} \text{ESR} = \left(\frac{n-1}{n} \cdot \frac{1-f}{f}\right)^{\frac{1}{2}} \quad .$$

The ESR test detects the presence of outliers iff this limit equals or exceeds C, that is, iff

$$f \leq [1 + nC^2/(n-1)]^{-1} \quad .$$

By Theorem 4.1, as $\Delta \to \infty$ the sample kurtosis approaches $1/f(1-f) - 3$, so the kurtosis test detects outliers whenever this quantity equals or exceeds the $1-\alpha$ quantile B of kurtosis under normal sampling, that is, when

$$f \leq \tfrac{1}{2}(1 - [1 - 4/(3+B)]^{\frac{1}{2}}) \quad .$$

For $\alpha = .05$, values of C, B, and the maximum number m and fraction f of outliers that can be detected by $b_{2,p}$ and ESR are shown for $n = 20, 30, 50, 100$ in Table 2. For $n = 20$, B is given by Ferguson (1961). Clearly, as n increases, the kurtosis test is able to detect a greater fraction of outliers, while the ESR test can detect a steadily decreasing fraction of outliers.

				ESD max		$b_{2,p}$ max	
n	t(1-.025/n;n-2)	C	B	f	m	f	m
20	3.5101	2.708	4.15	.1147	2	.1681	3
30	3.4786	2.908	*	.1026	3	*	5
50	3.5051	3.128	3.99	.0910	4	.1730	8
100	3.6008	3.384	3.77	.0799	7	.1802	18

Table 2: Maximum Outlier Detection Levels of ESD and Kurtosis Tests.

* Tabled value of B for $n = 30$ not available; $m = 5$ follows from assumption that $B \leq 4.15$ for $n = 30$.

6. Conclusions

The properties of the test of Theorem 2.1, rejecting H_0 whenever $b_{2,p} \geq K$, may be summarized briefly as follows. Let the fraction f of nonzero rows of A be less than $(3-\sqrt{3})/6$. The test is locally best invariant, uniformly in A. It is biased, asymptotically in Δ, when $K < b_{2,p}(A)$. However, as n increases, it appears that f must approach $(3-\sqrt{3})/6$ for this to happen, so for reasonably large samples and f somewhat less than 21.13...%, the test may be expected to behave better than its biasedness would suggest. The power $\Pr[b_{2,p}(Y) \geq K]$ is a highly complex function of α, n, p, Δ, and A.

Bibliography

Barnett, V. and Lewis, T. (1978). Outliers in Statistical Data. Wiley, New York.

Beckman, R.J. and Cook, R.D. (1983). Outlier·····s. Technometrics 25, 119-149.

Ferguson, T.S. (1961). On the rejection of outliers. Proc. Fourth Berkeley Symp. Math. Statist. Prob. I, 253-287.

Gnanadesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations. Wiley, New York.

Hawkins, D.M. (1980). Identification of Outliers. Chapman and Hall, New York.

Karlin, S. and Truax, D.R. (1960). Slippage problems. Ann. Math. Statist. 31, 296-324.

Kendall, M.G. and Stuart, A. (1969). The Advanced Theory of Statistics, Volume I, 3rd edition. Hafner, New York.

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

Mardia, K.V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. Sankhyā B 36, 115-128.

Pearson, E.S. and Hartley, H.O. (1966). Biometrika Tables for Statisticians, Volume I, 3rd edition. Cambridge University Press, Cambridge.

Schwager, S.J. and Margolin, B.H. (1982). Detection of multivariate normal outliers. Ann. Statist. 10, 943-954.