UNIVERSITY OF SOUTHAMPTON

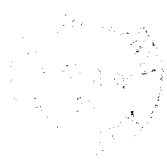
FACULTY OF MATHEMATICAL STUDIES

MULTIVARIATE STATISTICAL OUTLIERS

BY

CHRYSSEIS CARONI-RICHARDSON

A thesis submitted for the degree of Doctor of Philosophy



To my parents, Clive and Mark

ACKNOWLEDGEMENTS

I would like to record my deep gratitude to my supervisor, Dr. P. Prescott, for the contribution he has made to this work. I also thank the National Technical University, Athens, Greece for support in many ways, and especially the staff of the computing centre.

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF MATHEMATICAL STUDIES

Doctor of Philosophy

MULTIVARIATE STATISTICAL OUTLIERS

by Chrysseis Caroni-Richardson

Most of the extensive literature on outliers refers to the univariate case. This thesis takes up the topic of outliers in multivariate data, examining the performance of existing tests, and developing tests using other procedures and tests for specific data structures.

Chapter 1 introduces key concepts and provides examples to illustrate some of the main themes. Chapter 2 comprises a full review of previous work on outliers in multivariate data. Wilks' test is examined in the third chapter. It is confirmed by simulation that the Bonferroni approximation used to provide percentage points is accurate for testing for one outlier, but not for two or more. Simulated percentage points are constructed for up to four outliers, for sample sizes up to 100 and up to 5 dimensions. Chapter 4 presents sequential application of Wilks' statistic based on Rosner's procedures for univariate statistics which control the error level of the test for different numbers of outliers.

Chapter 5 examines Rohlf's test using distances in the minimum spanning tree. This is shown not to give a test with good properties. In Chapter 6, a two-outlier test is constructed by union-intersection methodology. This is sometimes more powerful than Wilks' test, but much less powerful under other data configurations. In Chapter 7, tests are derived for outliers in normal data with structured covariance matrices, specifically block structure and equicorrelation. It is shown by simulation that these tests are substantially more powerful than Wilks'. The final chapter examines outliers in the context of the multivariate linear model. Residuals are defined and the related topic of influence on estimates of regression coefficients is considered.

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 The content of this thesis

The subject of this thesis is the study of outliers in data and the related topic of the influence of particular observations on the outcome of an analysis. Although the history of the ideas can be traced back a long way, it is in recent years that they have received a lot of attention in the statistical literature, partly because their implementation as a routine part of statistical analysis needs modern computing facilities. This is especially true when one considers applications in multivariate problems, which are the particular theme of this thesis.

1.2 General ideas

Because very substantial reviews of the topic of outliers already exist (Barnett and Lewis, 1978 and 1984) Hawkins, 1980a; Beckman and Cook, 1983), a full general review will not be undertaken here. The present chapter will introduce some of the main points, including a brief discussion of some of the major univariate outlier detection methods and examples of outlier problems in real data. Chapter 2 will provide a full review of the multivariate outlier problem, which will be seen to be a relatively undeveloped aspect of outlier research despite the general practical importance of multivariate data. Subsequent chapters will then consider specific multivariate outlier procedures. Chapter 3 investigates the main existing test, due to Wilks, and Chapter 4 develops sequential test procedures based on statistics of Wilks' type. The following two chapters look at different methodologies for the same general problem: Chapter 5 examines Rohlf's gap test and Chapter 6 develops an outlier test based on union-intersection test construction. Attention then turns to structured data

problems, with the development of tests for outliers in data with particular patterns in the covariance matrix in Chapter 7 and the extension of univariate analyses of residuals and influence in the general linear model to the multivariate case in Chapter 8.

We must first consider what are 'outliers', what is 'influence' and why do they matter? The definition of outliers is discussed in detail in the major reviews. Suggestions include,

> "An outlying observation is one that appears different from the rest of the sample."
> (Kendall and Buckland, 1960)

> "Outliers are values which are either too large or too small compared with the rest of the observations."
> (Gumbel, 1960)

> "An outlier is an observation whose value is not in the pattern of values produced by the rest of the data."
> (Daniel, 1960)

> "An outlying observation or outlier is one that appears to deviate markedly from other members of the sample in which it occurs."
> (Grubbs, 1969)

The general idea of an outlier as a point which appears to be substantially different from the remainder of the sample is clear. Probably no more formal definition than this is necessary, if it is even possible, but in fact two distinct senses are often distinguished. To explain these, it is best to think of statistical analysis as consisting, in most situations, as the definition of a model describing the population from which a sample is available, followed by fitting the model and drawing inferences about the population on the basis of the fit. The central role of the model is clear, although in many circumstances the actual model may not really be made explicit and in fact may not be crucial to the validity of the analysis. (For example, t-tests on means require the normal distribution - the model - for the theory to be exact, but are very acceptable approximations under a wide

range of departures from this.) An outlier is an observation which is, in this framework, different from the rest of the sample. This could mean either that it is generated by a different model or that it appears to be different, for example as seen in a graphical representation or in possessing an extreme value of some statistic measuring some concept of distance between members of the sample. An outlier in the former sense, a point generated by a different model, need not be an outlier in appearing different - though if it is not, it probably cannot be detected. An outlier in the latter sense is either an outlier in the former sense or is a statistically improbable value arising because of an unusually large random component. In this thesis, 'outlier' is generally taken to mean a point which has the appearance of being different - the 'discordant observation' of Barnett and Lewis. Sometimes it is necessary to use the sense of an observation generated by a different model, especially when such points are generated in simulation studies of the power of test statistics, and then the change in meaning will be made clear in the text. Observations generated by a different model from the remainder of the sample are sometimes called 'model outliers' or 'contaminants' (Hawkins, 1980a).

The presence of outliers often indicates that something is wrong. The model may be wrong; the population to which the model is applied may be wrongly specified, so that heterogeneity in the sample correctly reflects an unrecognized heterogeneity in the population; the measurement of the outlying observation may have been incorrectly carried out or incorrectly recorded. Recognition of this leads to allowing for the outlier in some way - perhaps simply by discarding it - and this may affect the final conclusions and hence any action to be taken as a result of the statistical analysis. This is not the only way of looking at the matter, as occasional

examples can be found where the sole purpose of the analysis was to pick out the interesting outliers from the uninteresting mass of other points (Beckman and Cook's example is of counts of radiation levels over an area of central Canada in which a satellite had come to earth; outlying counts from the general background radiation indicated possible locations of satellite debris), but serves for most situations. Thus outliers matter because the outcome of the analysis may differ according to whether or not they are recognized and, if recognized, what is done about them.

At this point, it may be appropriate to comment on the action that may be taken after identification of outliers. This will vary, depending particularly on the purpose of the analysis and the framework in which it is conducted. It goes without saying that the first action should always be a simple check that the data provided were correct. Thereafter, the simplest - and possibly the commonest - action is to reject outliers and carry out analysis on the remaining data. This would be the more justified the more firmly established was the basic, uncontaminated model. It might also be justified in cases where a known mechanism existed for contamination. For example, this might arise in laboratory determinations of concentrations, of micro-organisms in sea water; it is usually necessary to dilute the original sample, and any error in executing or recording the dilution will result in a multiplicative error in the final value - which turns into an additive slippage since analysis of such data is often on the log scale. The other simple action is the opposite of the first, namely to keep the outliers and reject the rest of the data, which is what happens in the relatively uncommon examples where the purpose of the analysis is to identify these unusual points. If neither form of rejection applies, then there must be some kind of accommodation of the outliers. If the model is to be retained, the outliers may be handled by a robust

estimation procedure. Otherwise, the model will be adapted in some way, so that these points cease to be discordant - adopting a mixture model, for example.

However, the presence of outlying data values need not have any practical effect on the outcome of an analysis. Whether it does or not depends on various factors including sample size and the way in which the outliers differ from the other observations. Also, there may be effects on some aspects of the analysis but not on others. At this point, we are turning to the notion of influence (Cook and Weisberg, 1982). This is an idea that can be given mathematical expression for some purposes, as a function describing the stability of estimates in relation to changes in sample values. Such numerical expressions of influence will be used at various points of this thesis. Outliers and influence are closely related, although by no means the same thing. An outlier need not have much influence in the above sense (though it probably should in the wider sense of aiding recognition of shortcomings in the model), nor need an influential observation be an outlier, at least in appearing as an outlier in the way that outliers are usually investigated. For example, Figure 1.2.1 is a commonly used illustration of possible effects in regression. In the regression of y on x, point B has very high influence since its distance from the other points in x-space will tend to force the regression line to pass close to B, whereas it would be nearly horizontal were B not there. Point A does not have such influence on the regression coefficients, though it may well have the effect of influencing the residual mean square, which might affect the results of tests of significance.

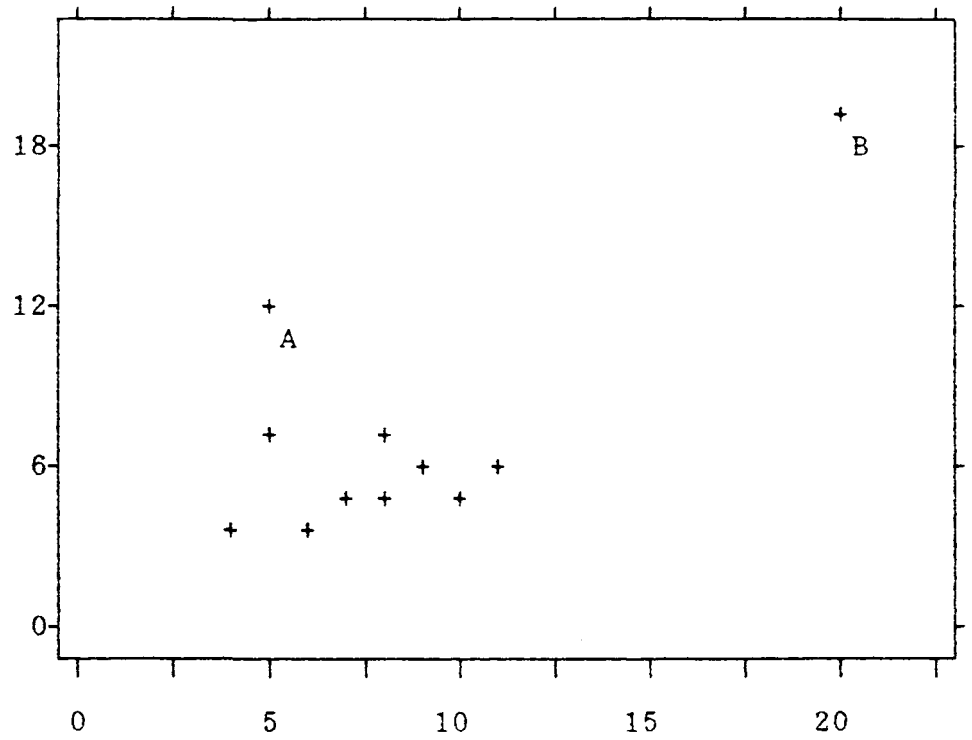


Figure 1.2.1 Illustration of outliers and influential points in a bivariate sample.

A conventional examination of the data for potential outliers would look at standardized residuals from the regression line: this would not indicate anything odd about B, though it would for A. Widely available computing packages now include influence measures (for example, program P9R of BMDP and the regression procedure of the microcomputer package SPSS/PC+), but older programs offered only residuals, which would not have revealed anything about point B.

The above illustration depends on our being interested in the regression of y on x. If the simple correlation between y and x were to be investigated, the approach would be different. Outliers would now be assessed in relation to the bivariate distribution, not to the regression line, and the anomalous position of B would be recognized instantly. The measure of influence on the

sample correlation coefficient could be used, although the standard packages do not provide this. Again we see the importance of the model. Not only do our concepts of outlier and influence involve comparison to a model, but the way in which we look for outliers and examine influence depends on the model assumed. The purpose of a part of this thesis will be to develop methods suitable for some particular models in multivariate problems.

After these general remarks, a number of specific methods of univariate outlier declaration will now be presented, the purpose being especially to illustrate those ideas which will be found relevant in the multivariate problem.

1.3 Some univariate outlier tests

In most of the literature, outlier detection is approached as an outlier testing problem. In this framework, the null hypothesis is that the sample $x_1,\ldots,x_n$ was generated as n independent realizations of a random variable following an assumed model, the distribution F:

$$H_0 : x_i \sim F \qquad i=1,\ldots,n$$

A reasonable form of alternative hypothesis is one which implies that most of the sample conforms to the null, but a small number of points (perhaps only one) have been generated by another model which tends to give values much different from those obtained under the null. The most popular choice is the slippage alternative. If F has location and scale parameters, then there may be slippage in the mean or slippage in the variance. Expressed for the case of one possible outlier and a normal model,

$$\text{A:}\quad H_0 : x_i \sim N(\mu,\sigma^2) \qquad i=1,\ldots,n$$
$$H_1 : x_i \sim N(\mu,\sigma^2) \qquad i\neq j$$
$$x_j \sim N(\mu+a,\sigma^2)$$

or

$$\text{B:}\quad H_0 : x_i \sim N(\mu,\sigma^2) \qquad i=1,\ldots,n$$
$$H_1 : x_i \sim N(\mu,\sigma^2) \qquad i\neq j$$
$$x_j \sim N(\mu,b\sigma^2)$$

with b>1. The labelling recalls Ferguson's (1961) definition of these alternatives as Models A and B. It should be noted that the index j is unknown in nearly all circumstances. While it is possible that there could be prior grounds for suspecting that one particular sample member has arisen from a different distribution, it is much more usual for the testing to be either a routine screening or a procedure carried out because an observation looked suspiciously out of line with the rest. In either case the formal framework has to allow for the possibility that any j=1,...,n could give rise to an outlier. The specific one under investigation will presumably be the most extreme in relation to some statistic, so that the test statistic will often be the maximum or minimum of a set of (probably correlated) values, which will give rise to difficulties in finding distributions of test statistics. One way of coping with the lack of specification of j is to use the two-stage maximum likelihood method to construct a test. The idea of this is that the likelihood ratio test of H_1 against H_0 is set up in the usual way for a specified j. This is the first stage. The second is to take the extreme value of the resulting test statistic over all choices of j. It is clear that the applications of this method are many and include the multivariate problem.

The above is not the only way of approaching the detection of outliers (there is, for example, some

Bayesian analysis described in Chapter 12 of Barnett and Lewis, 1984, in which the work of I. Guttman is prominent), nor are the above specifications the only alternatives possible within this framework (Barnett and Lewis, 1984, § 2.3). However, with very few exceptions, it is the way that can be found in the multivariate literature. Nor is it necessary, of course, to restrict attention to the normal distribution but - as will be seen in Chapter 2 - there is very little on any other multivariate distribution apart from the normal, so this restriction will be kept for the univariate case.

Before looking at examples of univariate test statistics, it may be observed that the idea that an outlier appears different from the rest of the sample implies that outliers will occur as extreme order statistics. Hence most statistics are expressed in terms of ordered $x_{(1)} \leq \ldots \leq x_{(n)}$ rather than the original sample $x_1, \ldots, x_n$.

Barnett and Lewis (1984) distinguish six basic types of test statistics, as follows.

1) Excess/spread statistics

The outlier is characterized by being unusually distinct from its neighbour, in relation to overall spread of the sample. One example is

$$\left(x_{(n)} - x_{(n-1)}\right) / \left(x_{(n)} - x_{(1)}\right)$$

which tests for an upper outlier. This statistic and several others of the same generic form are due to Dixon (1950, 1951).

2) Range/spread statistics

An example is

$$\left(x_{(n)} - x_{(1)}\right) / s$$

due to David, Hartley and Pearson (1954), where s is the

usual sample standard deviation. The curious feature of such statistics, which do not seem to be widely used, is that no particular point is being tested - is the outlier $x_{(1)}$, $x_{(n)}$ or both?

3) Deviation/spread statistics

The statistic testing for an upper outlier here is

$$\left(x_{(n)}-\bar{x}\right)/s$$

which was probably used for many years before Thompson (1935) produced exact results.

4) Sums of squares statistics

The simplest example here is S_n^2/S^2 where S_n^2 denotes the sum of squares in the reduced sample obtained by omitting $x_{(n)}$ and S^2 is the sum of squares in the full sample. This is actually equivalent to the statistic just given; in fact, reduced sums of squares can always be expressed in terms of deviations from the full sample mean $\bar{x}$. Grubbs (1950) gave various statistics of this kind.

5) High-order moment statistics

These are the statistics of sample skewness and sample kurtosis. They are shown by Ferguson (1961) to have some optimal properties. These statistics also do not test specific points as outliers.

6) Extreme/location statistics

These statistics are relevant to distributions with a fixed origin, such as the gamma which is confined to $x \geq 0$, since under these circumstances no shift in location is possible. For a distribution such as the normal, a statistic such as $x_{(n)}/\bar{x}$ is not invariant to arbitrary shifts and appears to be useless.

A few remarks will now be made on the questions of choice of statistic, distributional results and extension to testing for two or more outliers. On the first of

these issues, it can be seen from the above selection of upper outlier test statistics that there are likely to be several choices in any situation. The choice will seldom be clear, for there are different grounds for judging the performance of tests and the performance may anyway depend on the type of outlier. Even where optimality can be shown, it is of limited help. Thus, skewness and kurtosis are only locally optimal for small shifts, which are circumstances under which no test can have much power. For greater shifts in the mean, tests based on the studentized residuals do better.

Distributional results for test statistics are usually not available explicitly because of the complexity introduced by order statistics. However, in many cases, a recurrence relation applies which enables the density function of the outlier statistic in a sample of size n to be derived in terms of the density in a sample of size n-1. This method is illustrated in detail by Barnett and Lewis (1984, p. 178). If it does not apply, then a general way of obtaining percentage points for tests is to obtain conservative points by using the Bonferroni inequality. Because of its importance, this well known method will be repeated here. Let T be the outlier statistic

$$T=\max_i T_i$$

maximizing over choice of sample member i. Then

$$P(T>t)=P(\bigcup T_i>t)$$

$$=P(\bigcup E_i)$$

where E_i is the event $T_i>t$. Now

$$P(\bigcup E_i)=\sum P(E_i)-\sum\sum P(E_i\cap E_j)+\sum\sum\sum P(E_i\cap E_j\cap E_k)-\ldots. \tag{1.3.1}$$

and Bonferroni's inequality in its most general form states the fact that the partial sums formed by taking

more and more terms on the right hand side are alternately above and below the left hand side, with absolute differences becoming smaller. The usefulness of this result arises when it is not possible to calculate the distribution of T because the joint distribution of all the E_i is intractable. In this case, it is probably impossible even to obtain the distribution of pairs E_i, E_j so that not even the second term above can be found. What remains is the first Bonferroni inequality

$$P(\cup E_i) \leq \sum P(E_i) = nP(E_1).$$

Using this, the percentage points of T can be approximated. If $P(T_1 > t_{\alpha/n}) = \alpha/n$, then

$$P(T > t_{\alpha/n}) = P(\cup E_i) \leq \alpha.$$

This use of the first Bonferroni bound is what is usually meant when a test is called simply a 'Bonferroni test'. It is a conservative test, so that when a null hypothesis is rejected at nominal level of significance α using such a procedure, the true significance level is even lower.

The value of this approximation is that it can be applied in many problems, since all that is required is the distribution of an ordinary, unoptimized statistic. Furthermore, it is often a very good approximation, at least for applications to one outlier.

For problems involving two or more outliers, the first Bonferroni bound cannot be expected to be as good. The reason is that (1.3.1) should now be rewritten so that the event E_i is labelled E_{ij} in the two-outlier case meaning that the statistic computed as if points i and j were outliers has an extreme value (some multiple outlier statistics will be discussed below). However, the second term $P(E_{ij} \cap E_{kl})$ will now include contributions that are not small, as $P(E_{ij} \cap E_{ik})$ where i,j are genuine outliers and k is any other sample point. The test remains a

conservative test, but may be very much so. The difficulty of deriving exact distributions is greater for multiple-outlier statistics than for the single outlier, so the Bonferroni method, even if it is not very accurate, may be the only real means of analytical progress. The alternative will probably be simulation of percentage points: even this may not be a simple proposition, because for two outliers there will be $\binom{n}{2}$ statistics to evaluate in each sample, so a large-scale simulation will need very heavy computation.

The case of two or more outliers will now be discussed a little further before returning to the single outlier tests to see what effect the possible existence of more than one outlier may have on them. Some statistics already mentioned may apply immediately to the case of more than one outlier, such as the measures of skewness and kurtosis. Others can be extended. For example, the deviation/spread statistics for testing for two upper outliers would be

$$(x_{(n)}-\bar{x})/s$$

as before, and

$$(x_{(n-1)}-\bar{x})/s$$

A simultaneous test statistic for both $x_{(n)}$ and $x_{(n-1)}$ would have to combine these two values. The obvious way to do this is to take the sum of squared values

$$\{(x_{(n)}-\bar{x})^2+(x_{(n-1)}-\bar{x})^2\}/s^2$$

so that in effect the sum of squares statistic $S^2_{n,n-1}/S^2$ (Grubbs, 1950) has been obtained. An alternative retaining the deviation/spread form would be:

$$\left(x_{(n)}+x_{(n-1)}-2\bar{x}\right)/s$$

(Murphy, 1951). This can be seen to be testing the difference between the mean of the two hypothetical

outliers and the overall mean, and hence it is not surprising that it has optimal properties against the alternative that the outliers are generated from one distribution (that is, have the same slippage). However, McMillan (1971) showed that it was not robust against departures from this alternative, and that the sum of squares statistic would then be preferred. This reiterates the point that such optimality results as are known are of limited use.

A point concerning comparisons between tests that was not elaborated on earlier is the grounds of comparison. Barnett and Lewis (1984) select three useful performance measures for the single outlier case from possibilities proposed by David (1981). These are:

the power of the test in the usual sense of the probability of accepting the alternative hypothesis when it is true;

the probability of the contaminant (the point from the contaminating distribution) being the extreme value and being declared as an outlier by the test;

the probability of the contaminant being declared an outlier given that it is the extreme value.

The last two are relevant because the issue is not only whether or not a test declares an outlier to be present, but if this is actually from the contaminating distribution or is a point from the main distribution with an unusually large random component. The number of performance measures increases for the multiple outlier problem: Beckman and Cook (1983) listed six criteria, again incorporating the success of the method in identifying the correct points as outliers. However, published work in the multiple outlier problem largely concentrates on simply the number of outliers detected, an emphasis that will be shared by this thesis. This point leads to the final topic of this section, the question of how many outliers may be identified in one sample of data.

In the first place, consider a test for a specified number $t \geq 1$ of outliers. (The only restriction on t is that it is a small number relative to the sample size, for otherwise another analysis, such as fitting a mixture distribution, would be more suitable.) What happens if the actual number of outliers is different from t? There are two general phenomena which are relevant here, and the sensitivity of a test to each of them is another criterion relevant to test choice. These phenomena are masking and swamping. The former is relevant when the actual number of outliers exceeds the number being tested. This was expounded by Pearson and Chandra Sekar (1936) in relation to the extreme studentized deviate

$$(x_{(n)} - \bar{x})/s$$

for testing for a single upper outlier. They pointed out that if there was a second outlier, $x_{(1)}$ or $x_{(n-1)}$, then s could be so much further inflated over what would be expected in an uncontaminated sample, that the ratio would no longer be big enough to declare $x_{(n)}$ to be an outlier. Hence, the presence of a further, less extreme, outlier 'masks' the presence of the most extreme point. Swamping is the opposite effect: there are fewer outliers k than the $t \geq 2$ being tested for, but these are such extreme ones that the t-outlier statistic is sufficiently large to declare t outliers. The true outliers have carried along t-k other points with them and these are falsely declared to be outliers.

Since the commonest practical situation is not only that it is not known which points may be outliers, but not even known how many may be outliers, it must be regarded as a proper function of outlier testing in general to suggest the second decision as well as the first. Often, this is not considered in a formal way. If the outlier test is only being carried out because inspection of the data suggested that it is necessary, then probably the

same inspection suggests how many outliers to test for. In these circumstances, though, the whole formal hypothesis testing framework is dubious (see Collett and Lewis, 1976, and Example 1 of the following section). On the other hand, if outlier testing is treated as a routine screening of data, then choice of number of outliers ought to be allowed for formally. The most popular method of selecting the number of outliers is to apply tests for t=1,...,k outliers, where k is a chosen maximum, and to select the final value by comparing the results. If meaningful significance levels are to be obtained, this needs to be set up as a proper sequential (also called 'consecutive') procedure, with significance levels at each stage adjusted to allow for the other stages. This will be discussed in detail in Chapter 4, where a multivariate procedure of this type will be developed.

This overview of some of the main points of the study of outliers will be supplemented in the following section by a few examples, to be followed in Chapter 2 by a detailed review of the subject in the multivariate case.

1.4 Some examples of applications

The four examples presented here have different features as follows. The first is a fairly standard example of a univariate single outlier test, carried out because the sample appears heterogeneous. Example 2 shows a case where data appear to have been wrongly treated as if there were an outlier through failure to attempt a statistical evaluation of an apparent difference. Example 3 moves to bivariate data and illustrates both a test in this situation and the way in which the inclusion or exclusion of certain points may influence the results. Example 4 also is concerned with bivariate data and looks at an influence analysis in more detail.

Example 1

This example illustrates the simplest situation, of testing for an outlier in a univariate sample. The data (unpublished, provided by Prof. G.C.Lyketsos, University of Athens) are the scores of patients with alopecia on the "lack of self-confidence" subscale of the Personality Deviance Scale (Foulds, 1976) and form part of a series of studies of psychosomatic disorders. The scores for the 26 patients are:

6, 11, 13(2), 14(6), 15(9), 16(6), 18

where the number in brackets indicates how often the score was recorded, if more than once.

In these data, the value 6 catches the eye as much lower than the rest: perhaps some test should be conducted to see if this impression is justified, probably leading to omission of this patient if the test result is positive. One possibility is to use the lower outlier version of the maximum studentized deviate already introduced:

$$T_1=(\bar{x}-x_{(1)})/s$$

This takes the value 3.89, well beyond the critical value of the 1% level of statistical significance for a one-tailed test (Table VIIIa of Barnett and Lewis, 1984, extracted from Grubbs and Beck, 1972). However, the choice of a one-tailed test is based only on inspection of the data, so it is more correct to use a two-tailed test since an extreme upper order statistic would have led to the same investigation. The statistic is

$$\max\{(x_{(n)}-\bar{x})/s,(\bar{x}-x_{(1)})/s\}$$

It too is significant beyond the 1% level (Barnett and

Lewis, 1984, Table VIIIb, from Pearson and Hartley, 1972).

Another possibility is to use a Dixon-type statistic,such as

$$\left(x_{(2)}-x_{(1)}\right)/\left(x_{(n)}-x_{(1)}\right)$$

for testing a lower outlier, or the two-sided version

$$\max\left\{\frac{\left(x_{(n)}-x_{(n-1)}\right)}{\left(x_{(n)}-x_{(1)}\right)},\frac{\left(x_{(2)}-x_{(1)}\right)}{\left(x_{(n)}-x_{(1)}\right)}\right\}$$

The value of either statistic is 5/12=0.417, which is again statistically significant at 1% (two-tailed; Table XIVb of Barnett and Lewis, 1984). Thus there appear to be quite strong grounds for marking this observation as an outlier. It may be repeated that, one-tailed or two-tailed, the interpretation of any significance level in outlier testing is somewhat dubious because the test procedure is usually two-stage, the actual test following upon the decision that (in the analyst's judgment) some aspect of the data is surprising. Collett and Lewis (1976) make this point and go on to investigate, in a designed experiment, some of the factors affecting perception of possible outliers.

Example 2

Besides indicating a discordant observation, outlier tests may serve to indicate that an observation should not be treated as different from the rest. In the following example, attention seems to have been mistakenly concentrated on a value that was rather lower than the others. The data (Table 1.4.1) are the percentages of employees with serological evidence of past infection with hepatitis B, in six hospitals (Snydman et al, 1984).

Table 1.4.1 Positive tests for evidence of hepatitis B infection in screening hospital employees (Snydman et al, 1984).

Hospital	Number Screened	Positive	
		n	%
1	283	37	13.1
2	619	90	14.5
3	405	63	15.6
4	275	43	15.6
5	281	37	13.2
6	246	22	8.9
1-5 combined	1863	270	14.5

(Many epidemiological studies of this kind have been published, because medical staff are at high risk of infection with this dangerous disease but the vaccine is so expensive that it is more economical to screen to identify the susceptible than simply to vaccinate everybody.) Although the overall test of homogeneity in this table gave a non-significant result ($X^2_5 = 7.19$, $p = 0.2$) the authors selected hospital 6, observed that it differed substantially from the other five combined (with $X^2_1 = 5.16$, $p = 0.023$ for this comparison) and hence included a dummy variable to represent hospital 6 as one of the explanatory factors in their logistic regression model for infection rates. There were no grounds for this beyond their inspection of the data.

Is the rate in hospital 6 really excessively low, judged as the extreme of a set of six samples? There is not much ready-made theory for this problem. Barnett and Lewis (1984, p.200) give exact binomial theory for the case of equal sample sizes - but here they are not nearly equal. Instead (p.239) there is a conservative test, using tail probabilities of the hypergeometric distribution, which can easily be worked out on a microcomputer. The tail probability for hospital 6 is 0.0090. This may be

compared to 0.05/6 = 0.0083 for a Bonferroni test at the 5% level of the hypothesis of equal probabilities of past infection in the six populations against the alternative of downward slippage in one. Thus it is a one-tailed test and, as there seems to be nothing to justify one-tailed testing, it may be more appropriate to compare against 0.025/6 = 0.0042. So the calculated p-value appears to be well above the level at which it is justified to talk of hospital 6 as differing from the rest.

Example 3

This example is chosen partly because it is a multivariate one and partly because it illustrates that it is not sufficient to think only about a single outlier. The data, consisting of the responses (in terms of level of the hormone prolactin) of ten patients to electroconvulsive therapy (ECT) and to thyrotropin-releasing hormone (TRH), were published by Papakostas et al (1986). The importance of the experiment is as a contribution to the understanding of the mechanism of ECT, which has been widely used in therapy for many years without anyone really knowing why it works. The data used in the following discussion were read from a graph in the original publication, so do not correspond exactly to the true values:

Patient	1	2	3	4	5	6	7	8	9	10
ECT response	11	13	10	12	39	19	16	29	24	69
TRH response	10	19	27	28	44	49	50	59	80	98

The product-moment correlation between ECT and TRH responses is 0.794 (published value 0.824), statistically significant beyond the 1% level. The simple scattergram (Figure 1.4.1) suggests however that observation 10 is in some way quite distinct from the rest.

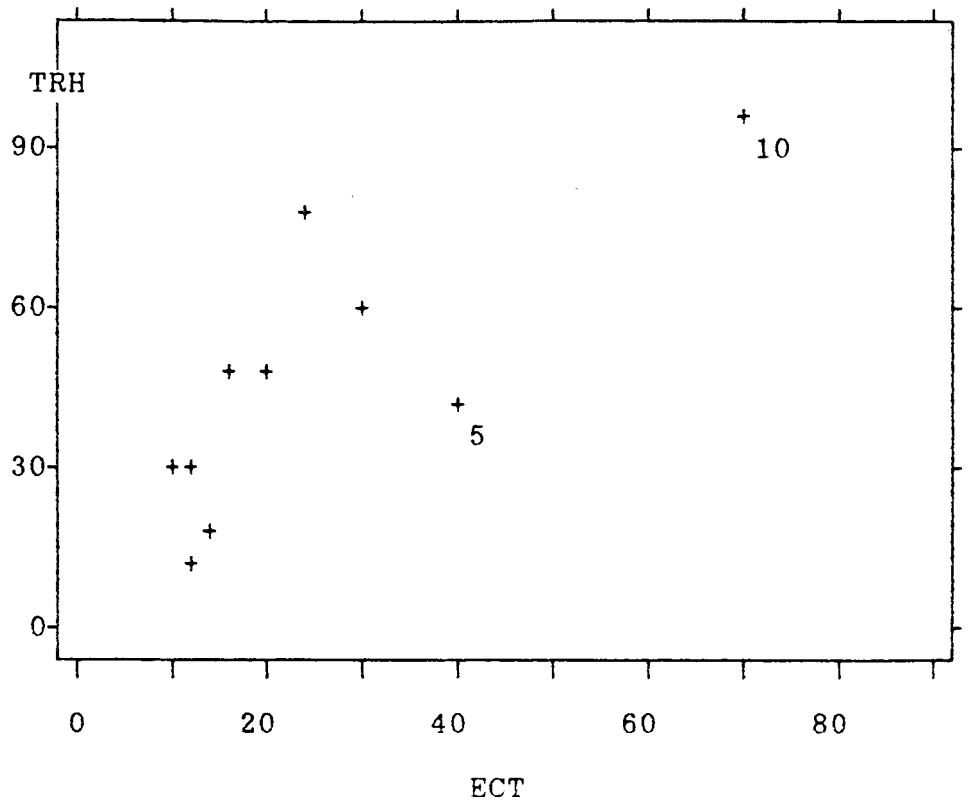


Figure 1.4.1 Prolactin in response to electroconvulsive therapy (ECT) and thyrotropin-releasing hormone (TRH). Source: Papacostas et al. (1986)

If this point is omitted, the correlation falls to 0.590 with p=0.095. A result like this, outside the usual levels of statistical significance, might well have meant that the report would not have been published, so the question of whether or not observation 10 "belongs" with the rest is an important one.

After omitting observation 10, observation 5 might now catch the eye as substantially different from the others and if this too is omitted the original high correlation is restored, the value being 0.841 (p = 0.009). The details of the relationship between the two responses are changed, however: the regression slope for ECT on TRH falls from 0.53 (standard error 0.14) in the original sample to 0.25 (s.e. 0.07) on omission of points 5 and 10.

Clearly, there is a need for objective procedures to guide this discarding of points, when so many alternative results are available. The basic test to be applied here is Wilks' (1963) test for an outlier in a multivariate normal sample, which will be discussed in detail in later chapters. The minimum value of the statistic is 0.254, on omission of observation 10, which is significant only at the 10% level (conservative test using Bonferroni bounds). This test extends easily to considering two or more outliers, although the Bonferroni approximation appears to be much less adequate in this case (Barnett and Lewis, 1984; Hawkins, 1980a). For two outliers, the test statistic is a minimum on omitting observations 5 and 10, with the value of 0.0481 falling below the 5% level (0.0585) but not quite the 2.5% level (0.0460).

There seem to be reasonable grounds therefore to suspect the homogeneity of the sample, with two observations appearing not to be from the same distribution as the other eight. Notice that the one-outlier test did not give a statistically significant result, probably because the inclusion of the other apparent outlier was inflating the variances and so causing "masking" - the failure to identify extreme values because of the presence of other extreme values. Notice also that there was no adjustment of significance levels of the two successive tests (for one or two outliers) to allow for multiple testing.

Another way of looking at these data is from the point of view of the idea of influence. Observation 10 is initially singled out because it gives the impression of strongly affecting the estimate of the correlation between ECT and TRH, as indeed it does. Devlin et al (1975; see § 2.7 of this thesis) give graphical methods of illustrating the influence of individual observations on the sample correlation coefficient, one of which is illustrated in Figure 1.4.2.

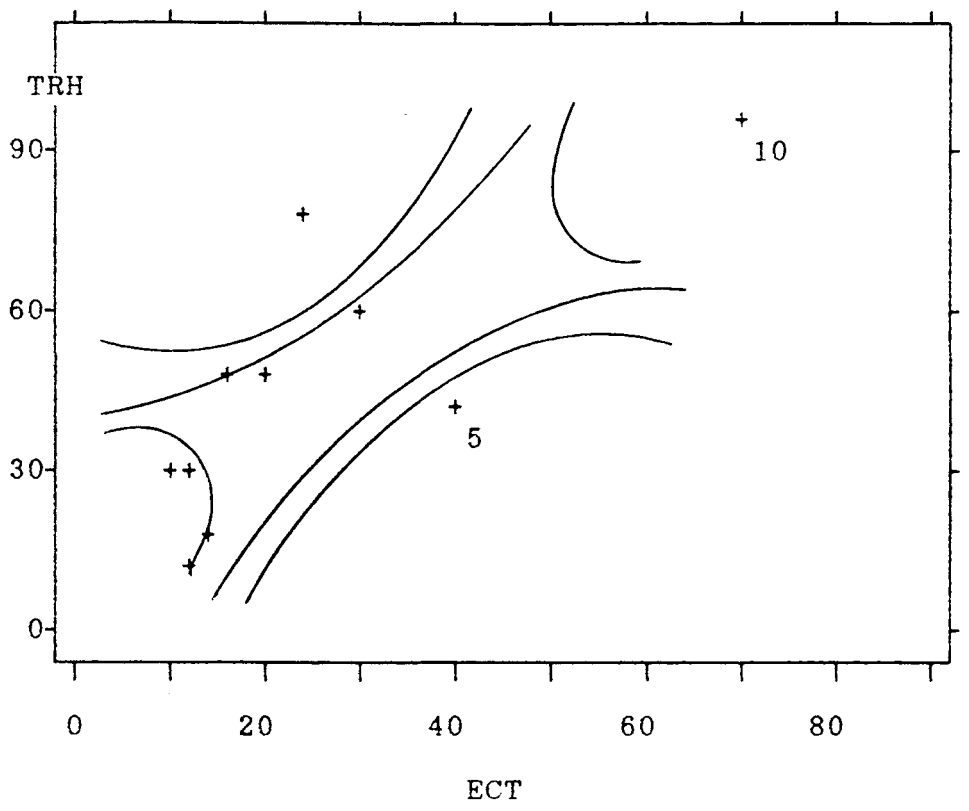


Figure 1.4.2 Prolactin data with contours of sample influence function for correlation coefficient.

In this diagram, the scatterplot is augmented by contours of an approximation to the sample influence function. These indicate by how much the correlation coefficient would change were a point on the contour to be deleted. The approximation does not seem to be very good in this problem, possibly because the sample is small and the calculation of the contours includes the effect of the probable outliers.

Example 4

The final example is concerned more directly with influence rather than outliers. It involves a linear

regression and is thus connected with the problem considered in Chapter 8. The context is a study of the improvement in the condition of schizophrenics after a course of treatment with the drug haloperidol (Smith et al, 1984). The original analysis regressed improvement (measured as percentage improvement on the psychosis factor of the Brief Psychiatric Rating Scale) on the level of haloperidol in the blood and found a need for a quadratic term. This indicates that response falls away at higher levels of haloperidol and therefore the medication must aim to get the level into a certain range, the 'therapeutic window'. This was the major conclusion of the paper and depended entirely on the statistical analysis.

Some correspondents were unhappy with this conclusion (Van Putten et al, 1985; Kirch et al, 1985). In particular, they looked at the graph of improvement against haloperidol level (Figure 1.4.3) and saw one extreme point which appeared to have a lot of influence in determining the curvature. Moreover, the regression was a weighted one (due to variable accuracy of measurement) and it was this point which carried the highest weight.

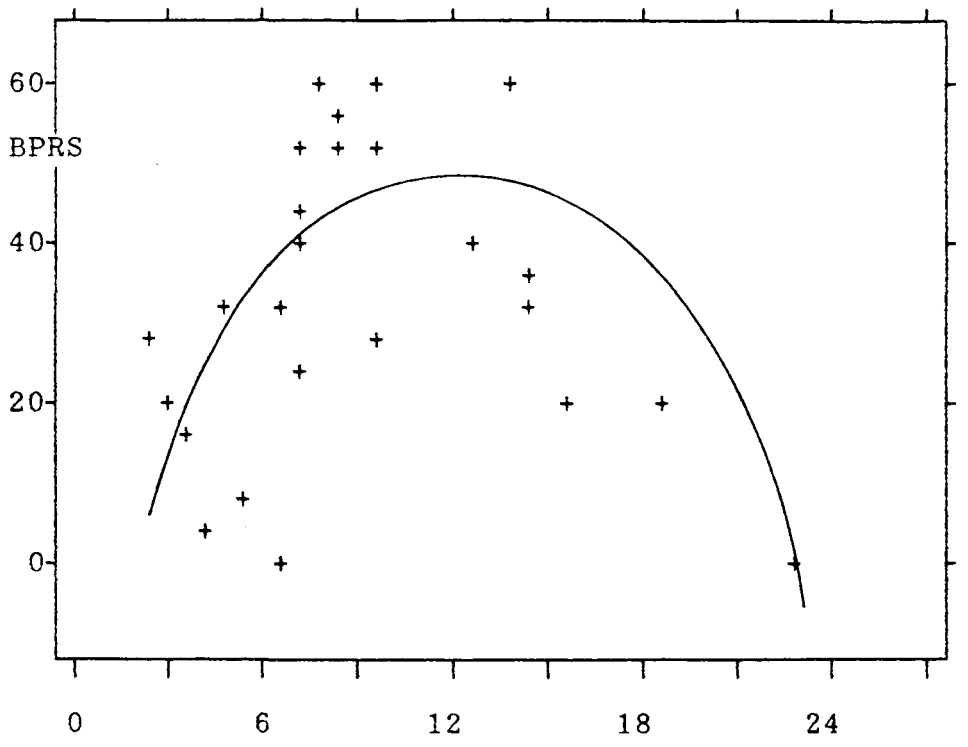


Figure 1.4.3 Percentage improvement in score on BPRS psychosis factor against steady-state plasma haloperidol level.
Source: Smith et al. (1984); data corrected in Smith (1985)

This correspondence would have been unnecessary if the original analysis had reported the influence analysis which is now routinely available in the major statistical packages. Fitting the regression using the program P9R in the BMDP package, it is found that Cook's distance measure of influence on the estimated regression coefficients (Cook, 1977) is indeed the largest for the point called into question by the correspondents, but is numerically quite small. With this point included, the regression equation for improvement is

$$-6.38 + 8.70\ (\text{level}) - 0.38\ (\text{level})^2$$

and without it

$$-12.13 + 10.31\ (\text{level}) - 0.47\ (\text{level})^2$$

The 95% confidence interval for the "acceptable range" of blood level of haloperidol (as defined by Smith et al) changes only from 6.9 - 17.6 to 6.7 - 17.2.

This example illustrates the need for examining and measuring influence in fitting a model. In higher-dimensional problems, such as the multivariate regression considered in Chapter 8, the need is all the greater since simple plotting is not available. In such cases, influence examination is particularly helpful in indicating the points with undesirably high influence instead of, as here, providing reassurance that a visually suspect point does not affect matters unduly. The example also illustrates that 'influence' can have many meanings. Cook's distance looks at the change in the vector of regression coefficients, but here the feature of more direct interest is the location of the 95% confidence interval indicating the therapeutic window: the influence on this piece of output from the analysis is required.

CHAPTER 2

MULTIVARIATE OUTLIER DETECTION : A REVIEW

2.1 Introduction

Relatively little has been written on the problem of detecting outliers in multivariate data, in comparison to the large literature for the univariate case. Barnett and Lewis (1984) devote only 26 out of 288 pages of text to the multivariate problem, and Hawkins (1980a) only 11 out of 127. The main reason for this must be the greater difficulty - both analytical and computational - of the multivariate case. For the same reasons of difficulty, the bulk of the literature on all aspects of multivariate analysis is limited to the normal distribution and thus there has been no cataloguing of outlier tests for different distributions in the multivariate case as there has been in the univariate.

Although an outlier in multivariate data might also appear as an outlier on one or more of the univariate marginal distributions, it does not necessarily do so. The purely multivariate concept of correlation may be involved, so that the outlier differs from the rest of the data set in violating the pattern of relationships between variables (as in the quotation from Daniel in the previous chapter). Therefore new methods are needed for handling the multivariate problem. There is also scope for new methodology in considering an outlier as a point which "appears" different from the rest of the sample. Taking this to mean, literally, its appearance in a graphical representation of the data, connects the problem of detecting multivariate outliers to the problem of obtaining a low-dimensional display of high-dimensional data. There are thus two main themes to be found in the literature on multivariate outliers: formal methodology related to hypothesis testing and informal methodology

linked to graphical displays.

At this point, it is appropriate to refer to the idea of generalized distance, which appears in both formal and informal methods, as will be seen subsequently. It arises as a partial solution to the problem of ordering multivariate data. Ordering is a basic part of outlier testing in the univariate case: it is important because the concept of an outlier as an observation noticeably different from the rest implies that an outlier has to appear at one or the other extreme of the list of ordered sample values. In particular, Dixon's tests use only these order statistics. In the multivariate problem, there is no direct equivalent of univariate order statistics, so tests of this type cannot be applied. However, other univariate test statistics order points in respect of distance from the mean of the sample, as in the maximum studentized range

$$\max_i \frac{|x_i - \bar{x}|}{s} \qquad (2.1.1)$$

and this idea can be extended to the multivariate case, because a sub-ordering (Barnett, 1976) of observations in relation to the sample mean is provided by generalized distances

$$(x_i - \bar{x})' S^{-1} (x_i - \bar{x}) \qquad (2.1.2)$$

where S is the sample covariance matrix.

The first candidate for an outlier in a multivariate sample is that observation x_j which maximizes (2.1.2). This is equivalent to Wilks' statistic for testing for outliers in multinormal data, which is to be discussed in detail in the following chapter. Here, we note that this choice of statistic can be justified in three ways:

1) (2.1.2) is the direct multivariate equivalent of (2.1.1);

2) the multivariate normal density $N_p(\mu,\Sigma)$ is given by

$$(2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\{-(1/2)(x-\mu)'\Sigma^{-1}(x-\mu)\}$$

so that (2.1.2) provides estimates of probability densities and the sample points are ordered in relation to the contours of the p.d.f;

3) as will be shown in Chapter 3 a likelihood ratio approach to a hypothesis testing problem leads to (2.1.2).

The second and third of these ways suggest general methods of developing tests for use with multivariate distributions. Furthermore, likelihood ratio is not the only general method of test construction in common use in multivariate analysis. There is also the union-intersection method, and an application of this to constructing a multivariate outlier test will be explored in Chapter 6 of this thesis.

The following sections review existing formal methods for testing for outliers in multivariate data and a later section of this chapter looks at the available informal methods.

2.2 Tests for a single outlier in multivariate normal data

With the exception of some work by Barnett to be mentioned in the following section, the multinormal case discussed here covers all the multivariate outlier literature. To examine the outlier-testing problem requires the setting up of null and alternative hypotheses. The null will be

$$H_0 : x_i \sim N_p(\mu,\Sigma) \qquad i=1,\ldots,n$$

- that is, there are n independent observations from the same multivariate normal distribution. In problems of practical interest, both μ and Σ are usually unknown.

Hawkins (1980a) lists three possible alternative hypotheses for contamination of the data by k outliers. Writing them in general form for the k-outlier problem, they are

$$\text{Model 1 - } H_1 : \begin{array}{ll} x_{j(i)} \sim N_p(\mu,\Sigma) & i=k+1,\ldots,n \\ x_{j(i)} \sim N_p(\mu_i,\Sigma) & i=1,\ldots,k \end{array}$$

where $j(i)$ is an unknown permutation of the integers $1,2,\ldots,n$;

$$\text{Model 2 - } H_2 : \begin{array}{ll} x_{j(i)} \sim N_p(\mu,\Sigma) & i=k+1,\ldots,n \\ x_{j(i)} \sim N_p(\mu,a_i\Sigma) & i=1,\ldots,k \end{array}$$

where a_i is a scalar; and

$$\text{Model 3 - } H_3 : \begin{array}{ll} x_{j(i)} \sim N_p(\mu,\Sigma) & i=k+1,\ldots,n \\ x_{j(i)} \sim N_p(\mu,\Sigma_i) & i=1,\ldots,k \end{array}$$

Model 1 is a slippage of the mean, while models 2 and 3 both represent changes in the covariance matrix so are analogous to a univariate slippage of the variance model. Barnett and Lewis (1984) discuss models 1 and 2 for the single-outlier case. Model 3 has not been investigated.

It will be shown in Chapter 3 that the likelihood ratio test of model 1, for a single outlier and for a specified outlier candidate j, leads to the statistic

$$\Lambda_j = \frac{|A_j|}{|A|} \qquad (2.2.1)$$

where A is the sum of squares and products matrix of the entire sample

$$A = \sum_i (x_i-\bar{x})(x_i-\bar{x})'$$

and A_j is the equivalent quantity recalculated after omitting observation x_j. A reasonable choice of outlier test statistic when it is not known beforehand which point is the potential outlier is therefore to find that x_i giving the extreme value of (2.2.1); thus

$$\min_j \frac{|A_j|}{|A|} \qquad (2.2.2)$$

is the outlier test statistic. This is the two-stage maximum likelihood method.

The statistic (2.2.2) was introduced by Wilks (1963) and is the only multivariate outlier statistic in common use. Wilks' paper will be discussed in more detail in the following chapter: here it will simply be noted that Wilks motivated this choice of statistic by an interpretation of $|A|$ in terms of volumes of simplexes formed by points from the samples, so that the outlier is the point whose removal most reduces this value and hence leaves as compact a set of remaining points as possible. If the problem is in fact one-dimensional, then $|A_j|/|A|$ reduces to a ratio of two ordinary sums of squares S_j^2/S^2 and hence gives Grubbs' statistic. It is well known that this is equivalent to testing with studentized deviations from the mean: for example, for testing for an upper outlier

$$\frac{S_n^2}{S^2} = 1 - \frac{n(x_{(n)}-\bar{x})^2}{(n-1)S^2}$$

A similar result holds for the multivariate case. The reduced sum of squares and products matrix is, (assuming without loss of generality that the nth point has been omitted)

$$A_n = \sum_{i=1}^{n-1} (x_i-\bar{x}_n)(x_i-\bar{x}_n)'$$

where

$$\bar{x}_n = \sum_{i=1}^{n-1} \frac{x_i}{(n-1)}$$

Now because

$$n\bar{x} = (n-1)\bar{x}_n + x_n \qquad (2.2.3)$$

we can substitute for $\bar{x}_n$ in A_n

$$A_n = \frac{1}{(n-1)^2} \sum_{i=1}^{n-1} \{(n-1)x_i - n\bar{x} + x_n\}\{(n-1)x_i - n\bar{x} + x_n\}'$$

$$= \frac{1}{(n-1)^2} \sum_{i=1}^{n-1} \{(n-1)(x_i - \bar{x}) - \bar{x} + x_n\}\{(n-1)(x_i - \bar{x}) - \bar{x} + x_n\}'$$

$$= \sum_{i=1}^{n-1} (x_i - \bar{x})(x_i - \bar{x})' - \frac{1}{n-1}(x_n - \bar{x})(x_n - \bar{x})'$$

$$= A - (x_n - \bar{x})(x_n - \bar{x})' - \frac{1}{n-1}(x_n - \bar{x})(x_n - \bar{x})'$$

$$= A - \frac{n}{n-1}(x_n - \bar{x})(x_n - \bar{x})' \qquad (2.2.4)$$

This well-known updating formula now permits calculation of an alternative form for Λ. Let B be the partitioned matrix

$$B = \begin{pmatrix} A & \sqrt{\frac{n}{n-1}}(x_n - \bar{x}) \\ \sqrt{\frac{n}{n-1}}(x_n - \bar{x})' & 1 \end{pmatrix}.$$

Then its determinant can be expressed in two alternative ways:

$$|B| = |A| \cdot \{1 - \frac{n}{n-1}(x_n - \bar{x})' A^{-1} (x_n - \bar{x})\}$$

$$= \left| A - \frac{n}{n-1}(x_n - \bar{x})(x_n - \bar{x})' \right|$$

(Morrison, 1976, p.68)

Hence

$$\Lambda_n = \frac{\left| A - n/(n-1)(x_n - \bar{x})(x_n - \bar{x})' \right|}{|A|}$$

$$= 1 - \frac{n}{n-1}(x_n - \bar{x})' A^{-1} (x_n - \bar{x}) \qquad (2.2.5)$$

Consequently, Wilks' Λ ratios are monotonic functions of the generalized distances. This offers ease of computing all the n ratios Λ_i in a sample, since just one matrix inversion is needed.

Another alternative form can be obtained by observing that (2.2.3) can be rewritten as

$$x_n - \bar{x} = \frac{(n-1)}{n} (x_n - \bar{x}_n)$$

so that (2.2.4) is the same as

$$A = A_n + \frac{(n-1)}{n} (x_n - \bar{x}_n)(x_n - \bar{x}_n)'$$

The same device of partitioning a matrix as led to (2.2.5) then gives

$$\Lambda_n^{-1} = 1 + \frac{n-1}{n} (x_n - \bar{x}_n)' A_n^{-1} (x_n - \bar{x}_n)$$

$$= 1 + \frac{T_n^2}{n-2} \qquad (2.2.6)$$

where T_n^2 is Hotelling's T^2 statistic for testing the hypotheses

$$H_0 : x_i \sim N_p(\mu, \Sigma) \qquad i=1,\ldots,n$$

$$H_1 : x_i \sim N_p(\mu, \Sigma) \qquad i=1,\ldots,n-1$$
$$x_n \sim N_p(\mu_n, \Sigma).$$

Thus the outlier testing problem is equivalent to a two-group comparison.

One special variation on model 1 will be mentioned before model 2 is discussed. This arises when the covariance matrix V is either known or estimated independently from the sample which is being investigated for outliers. Tests have been suggested analogous to the version (2.2.5) of Wilks' statistic, using distances expressed in the general form

$$R(x; x_0, \Gamma) = (x - x_0)' \Gamma^{-1} (x - x_0) \qquad (2.2.7)$$

in the notation of Barnett and Lewis (1984). In this

notation, the generalized distances equivalent to Wilks' test are $R(x;\mu,A)$, or $R(x;\mu,S)$ where S is the sample covariance matrix. Siotani (1959) investigated $R(x;x_0,\Sigma)$ for Σ known and $x_0=0$, μ (known) or $\bar{x}$. Knowing μ gives a very simple case, although entirely unrealistic, for the $R(x;\mu,\Sigma)$ are then independent X_p^2 variates and the outlier problem requires only the order statistics of a chi-square distribution. Barnett and Lewis (1984) give appropriate tables, corrected from Gupta (1960). They also reproduce Siotani's tabulations for the case $R(x;\bar{x},\Sigma)$ and for $R(x;\bar{x},V)$ where V is an independent estimate of Σ. None of these cases seem to be of sufficient practical importance to be worth pursuing any further: however, distances of the above form (2.2.7) will be seen again in the subsequent section on graphical methods.

The analysis of model 2, for unknown μ, Σ and a_i, was investigated for a single outlier by Ferguson (1961). He defined the problem as the search for the optimal decision rule, in the sense of maximizing the probability $p_i(D_i)$ of declaring that x_i is the contaminant when this is in fact true. Within the class of decision rules which are invariant under shifts of location and under rotation, have size α (probability of correctly declaring that no value is an outlier is $1-\alpha$) and for which $p_i(D_i)$ is independent of i, Ferguson found that the optimal rule again uses distances $R(x_i;\bar{x},S)$. If j is that observation with the greatest value of this distance in the sample, then this observation is declared to be the contaminant if

$$R(x_j;\bar{x},S)>k$$

where k is chosen so that the rule has the desired size. Hence, because of (2.2.5), Ferguson's decision rule for model 2 is just the same as Wilks' test for model 1. A further point from Ferguson's analysis is that his decision rule is the uniformly best procedure over all values of the parameter a.

Very little exists in the testing literature apart from Wilks' statistic. Rousseeuw's robust version of the statistic will be mentioned in § 2.7. A test based on kurtosis will be presented in the following section, since it is not specifically a test for one outlier. The only remaining analytical method of investigating the presence of a single outlier in multivariate normal data is Guttman's (1973) Bayesian analysis. This appears to be the only extension of the Bayesian methodology to the multivariate case, which is not surprising since there will often be formidable problems of evaluating integrals in the univariate case which become excessive in the multivariate case.

Guttman adopts the slippage in the mean model and writes the likelihood as

$$\sum_{j=1}^{n} (1/n) \prod_{i \neq j} \phi(x_i;\mu,\Sigma).\phi(x_j;\mu+a,\Sigma)$$

where ϕ is the usual multivariate normal density. This form is adopted because the prior probability that any specified observation x_j is the outlier is 1/n. The technique is to impose a non-informative prior joint distribution for (μ,σ^2,a) which is simply proportional to σ^{-2}. If this is combined with the likelihood to yield the posterior joint distribution of μ, σ and a, then μ and σ^2 can be integrated out to give the posterior marginal distribution of a. Its form is a weighted combination of multivariate t distributions (one for each observation), so that it is easy to find the posterior mean and variance of the marginal distribution of a. It does not seem possible to carry out a simultaneous assessment of all components of a, but each component a_i can be looked at just as in the univariate case, by finding the posterior odds

$$\gamma_i = \frac{P(a_i>0)}{P(a_i<0)}$$

High values of such odds can be taken to indicate that mean shift has occurred and in which component. The values of the weights for each observation in the posterior distribution of a indicate which observation may be an outlier. These weights are

$$c_j = \frac{|A_j|^{-(n-2)/2}}{\sum_{i=1}^{n} |A_i|^{-(n-2)/2}}$$

where A_j denotes as usual the sum of squares and products matrix of the reduced sample obtained by omitting observation x_j. In other words, the relative values of the quantity used to assess which observation may be an outlier are simply the $|A_j|$, exactly as in Wilks' statistic.

2.3 Tests for two or more outliers in multivariate normal data

The remarks on the derivation of Wilks' one-outlier statistic suggest how to extend to the case of 2 or more outliers. Wilks' volume argument applies equally well to omitting a set T=(ijk...) of points from the original sample as to omitting a single point. His statistic therefore becomes in general

$$\max_T \frac{|A_T|}{|A|} \quad .$$

This may also be derived by two-stage maximum likelihood starting with the alternative hypothesis H_1 of the previous section and is equivalent to a one-way multivariate analysis of variance between k+1 groups (namely, k groups each consisting of a single point - the outlier candidates - and one group consisting of the remaining n-k points hypothesized to conform to the main uncontaminated distribution). Wilks (1963) provided Bonferroni percentage points for the two-outlier

statistic, as for the single outlier, and discussed the general case, giving distributions for some particular cases with 3 or 4 outliers. Simple use of Wilks' statistics for up to 4 outliers is discussed in detail in Chapter 3 and consecutive application of these statistics for different numbers of outliers in the same sample is discussed in Chapter 4.

Bacon-Shone and Fung's (1987) graphical method for detecting one or more outliers, based on Wilks' statistic, will be discussed in the section on informal methods.

The decision rule approach of Ferguson appears not to have been investigated for more than one outlier. In fact, model 2 in general seems to have been considered only for the single outlier case.

Another general multivariate outlier test statistic will be mentioned at this point because it is a test for any number of outliers. This is Schwager and Margolin's (1982) test using the sample kurtosis proportional to

$$\sum_{j=1}^{n} \{ (x_j - \bar{x})' S^{-1} (x_j - \bar{x}) \}^2$$

defined by Mardia (see Mardia, Kent and Bibby, 1979). Exactly as in the univariate case, this has certain optimality properties with normal data, although again this is of limited practical importance. Schwager and Margolin recommend their test as an overall test for the presence of any outliers in the same way that an F test in the analysis of variance serves to confirm the presence of some differences which must then be identified in detail by other means. This provides an overall significance level irrespective of what is done in a subsequent sequential procedure.

It may be noted that there is no multivariate equivalent to the use of sample skewness to test for outliers, which provides an optimal one-sided test in the

univariate case. This is because the unidirectional hypothesis has no multivariate equivalent.

2.4 Other multivariate distributions

Although a large number of multivariate distributions have been defined (for example, Johnson and Kotz, 1972), nothing apart from the multivariate normal appears to be used very much. Only Barnett (1979) has investigated any of these other distributions from the point of view of outlier detection. He considered three distributions - uniform, exponential and Pareto - all in the bivariate case only. Test statistics are considered in relation to two general principles for detection of multivariate outliers, which he elsewhere (Barnett and Lewis, 1984) labelled as principles A and B. Principle B is just the statement of the two-stage maximum likelihood method for testing the null hypothesis specifying some model F against an alternative hypothesis specifying the contaminated model F', where it is not known which observation may be the contaminant. His principle A is similar, but refers to no particular alternative hypothesis:

"The most extreme observation is that one, x_i whose omission from the sample $x_1, x_2, \ldots, x_n$ yields the largest incremental increase in the maximized likelihood under F for the remaining data. If this increase is surprisingly large, declare x_i to be an outlier." (Barnett and Lewis, 1984, p.246).

Barnett's first example is the case of two independent uniformly distributed random variables with known ranges. This may seem too simple to be useful, but he suggests an application in a cancer diagnosis problem. An area is being estimated, so that the relevant quantity is a product of random variables: Barnett provides a short table for testing a test statistic suitable for this

particular problem, as well as tables for test statistics based on distance criteria. However, despite this practical illustration, the usefulness of this distribution is probably very limited. Of much more general interest are skew distributions, as shown by a few examples of bivariate data extracted from the literature by Barnett. In order to look at representations of such data, he studied the two other distributions in his paper, the exponential and the Pareto.

Quite a lot has been written on outlier detection in the univariate exponential distribution, which is a distribution with many practical applications (in lifetime distributions, for example, and in connection with Poisson processes). Being a long-tailed distribution, it is also interesting from the point of view of studying outliers since it naturally produces observations which may appear to the eye to be extreme. The first difficulty in extending the study to the multivariate problem is that there is no one "multivariate exponential distribution". Johnson and Kotz (1972) list 6 bivariate alternatives; the one selected by Barnett is due to Gumbel (1960) and has density

$$f(x_1,x_2)=\{(1+\theta x_1)(1+\theta x_2)-\theta\}\exp(-x_1-x_2-\theta x_1 x_2) \qquad (2.4.1)$$

for $x_1>0$, $x_2>0$ and $0<\theta<1$. The marginal distributions of X_1 and X_2 are both standard exponentials (that is, with parameter 1). The product-moment correlation between X_1 and X_2 is a function of θ, and varies from 0 to approximately -0.40 as θ increases from 0 to 1. If θ is zero, so that X_1 and X_2 are independent, then $2(X_1+X_2)$ follows the X^2_4 distribution, so tables of gamma order statistics (Gupta, 1960; Barnett and Lewis, 1984) can be applied to this problem.

Now suppose that θ is non-zero and its value is known. Applying Principle A, a suitable test statistic is the maximum (or minimum, but the former is probably usually

the more interesting, representing an "upper" outlier) of $U=X_1+X_2+\theta X_1X_2$ The distribution function of $T=1+\theta U$ is

$$H(t)=1-\{(t\ln t)/\theta+1\}\exp\{-(t-1)/\theta\} \qquad (2.4.2)$$

The distribution function of the sample maximum is

$$[H(t)]^n$$

and hence simple iterative methods can produce exact percentage points for the maximum of U. Barnett (1979) gives tables of 5% and 1% points for a range of values of θ and fresh tables can be found in Barnett and Lewis (1984).

The question now is how to proceed if θ is unknown, since this is the more realistic problem. Barnett does not offer a complete solution. No analytical progress appears possible with a statistic of the form $x_1+x_2+\hat{\theta}x_1x_2$ for an estimator $\hat{\theta}$. Observing from simulations that critical values do not seem to depend strongly on θ, Barnett suggests that a conservative test might be carried out - presumably by taking as critical value the most extreme of all critical values for different θ. Another suggestion is to use $x_1+x_2+kx_1x_2$ for a selected constant k, so that θ is ignored: this test statistic has not been investigated.

A feature of the bivariate exponential distribution considered above is that it only admits negative correlations. On the other hand the bivariate Pareto distribution (of the first kind; Mardia, 1962) has only positive correlations. Its density is

$$f(x_1,x_2)=a(a+1)(\theta_1\theta_2)^{a+1}(\theta_2x_1+\theta_1x_2-\theta_1\theta_2)^{-(a+2)} \qquad (2.4.3)$$

with $x_1\geq\theta_1\geq0$, $x_2\geq\theta_2\geq0$ and $a>0$, or $a>2$ for the existence of second-order moments. The product-moment correlation is a^{-1}, so that $0<\rho<0.5$. If, as before, the parameters are assumed known and Principle A is applied, the test

statistic obtained is

$$R=(X_1/\theta_1)+(X_2/\theta_2)-1 \qquad (2.4.4)$$

which has distribution function

$$G(r)=1-r^{-a}(1+a-(a/r))$$

so that simple iteration again produces exact percentage points for testing the sample maximum. (Note that this function is given correctly in Barnett's paper, but misprinted in Barnett and Lewis, 1984.) Barnett (1979, reproduced in Barnett and Lewis, 1984) provides a table of values. Again, this result is of limited interest in itself because the more realistic case is when the parameters of (2.4.3) are unknown. Some progress is possible for the case of unknown θ_1 and θ_2 but known a. Reasonable, although not maximum likelihood, estimators of θ_1 and θ_2 are the minima of the two marginal distributions, $X_{1(1)}$ and $X_{2(1)}$, so that it is obvious to try substituting these for θ_1 and θ_2 in (2.4.4). The distribution of the resulting quantity has not been found, but since

$$\frac{X_{1j}}{X_{1(1)}} + \frac{X_{2j}}{X_{2(1)}} - 1 \leq \frac{X_{1j}}{\theta_1} + \frac{X_{2j}}{\theta_2} - 1$$

for any sample observation (X_{1j},X_{2j}), it follows that the percentage points obtained for the case of known θ_1 and θ_2 provide conservative bounds for unknown θ_1 and θ_2, with a known in both situations. Barnett's simulated percentage points for unkown θ_1 and θ_2 appear quite close to the exact percentage points for known θ_1 and θ_2, so this test is a good approximation. However, its usefulness is limited by the assumption that a is known. Further work to lift this restriction has not been carried out.

The common feature of these applications is that only relatively uninteresting problems have been solved, because the problems posed by the need to estimate

parameters of the distributions have not been overcome. This appears to be an inevitable consequence of attempting to go beyond the multivariate normal distribution. The point will be met again in relation to Rohlf's gap test, introduced in the following section and studied further in Chapter 5.

The fact that the normal distribution is relatively straightforward to handle in comparison to other distributions suggests the possibility of carrying out transformations to multivariate normality to obtain tests for other cases. This idea has also been considered by Barnett (1983). The principal limitation of his method is that it is again necessary to take the distributions as having known parameter values. If this is accepted, then the method in the bivariate case is to transform the random variable (X_1,X_2) to the pair of independent N(0,1) random variables (U_1,U_2) by

$$F_{X_1}(x_1)=\Phi(u_1)$$

$$F_{X_2|x_1}(x_2)=\Phi(u_2) \qquad (2.4.5)$$

where F_{X_1} and $F_{X_2|x_1}$ denote marginal and conditional distribution functions and Φ is as usual the standard normal distribution function. The obvious outlier test statistic in the space of (U_1,U_2) is $U_1^2+U_2^2$; half the largest value of this is distributed as the largest order statistic in an exponential sample, giving critical value for a size α test

$$-2\ln\{1-(1-\alpha)^{(1/n)}\}.$$

Barnett looks at some properties of this test for the cases of bivariate exponential and Pareto distributions as before. One point that emerges is that the asymmetric treatment of the two original random variables X_1 and X_2 in (2.4.5) is of little consequence in these two

situations. However, these are relatively trivial matters beside the fact that again it is not possible to take any proper account of the need to estimate parameters of the original distributions in order to carry out transformations (2.4.5) as in most practical circumstances.

In conclusion, it has to be said that little has yet been achieved in the study of outliers from non-normal multivariate distributions. Barnett (1979) introduced his efforts as an "attempt to awaken interest" in the topic. There seems not to have been much response so far.

2.5 Rohlf's gap test

The method introduced by Rohlf (1975) is placed at this point because it links the formal methods of outlier detection with the informal. It is formal to the extent that a test of significance has been proposed, but informal in not explicitly specifying any underlying distribution and also in that it could be used simply as a graphical display. Rohlf's method forms the subject of detailed investigation in Chapter 5; it will only be summarised here.

The idea behind the method is similar to that behind Dixon's gap tests for univariate outliers. It will be recalled that Dixon's tests use in various ways differences $x_{(k)}-x_{(k-1)}$ between successive order statistics. If such a "gap" is unusually large, there is an indication that the point corresponding to $x_{(k)}$ is not from the same distribution as $x_{(1)},\ldots,x_{(k-1)}$ and, by implication, neither are $x_{(k+1)},\ldots,x_{(n)}$. These tests therefore have some appeal for detecting "clusters" of outliers, as well as single ones.

In the multivariate case, our inability to define order statistics means that such gap tests cannot be applied directly. However, the general idea still stands, that an

outlier or cluster of outliers must be separated from the main set of points by a distance which is relatively large compared to distances within the main set. The question is, how to identify these distances? Rohlf suggests looking at the minimum spanning tree (MST) of the data set, because if the largest distance in it is unusually big (in relation to the other distances), then there appears to be an outlier or cluster of outliers. The MST may be examined through a probability plot of its elements - a gamma plot is proposed on empirical grounds. On the same grounds, the largest distance can be tested approximately using tables of the gamma. The details of these proposals will be filled in later in Chapter 5 and the performance of Rohlf's test as a formal test will be studied.

2.6 Graphical methods

Given the notion of an outlier as a point which appears different from the rest, any of the many ways of producing graphical and pictorial representations of multivariate data is a potential aid in outlier detection. A specific emphasis on graphical means of outlier detection will be found in Gnanadesikan (1977), which draws especially on Gnanadesikan and Kettenring (1972) among earlier work.

Probability plots of a set of observed distances in the sample against expected order statistics of a theoretical distribution provide one basic means of looking at the homogeneity of a set of points. It seems that Healy (1968) first advanced this idea. He pointed out that squared generalized distances

$$D^2=(x-\mu)'\Sigma^{-1}(x-\mu) \qquad (2.6.1)$$

for known μ and Σ follow the X^2_p distribution. In the bivariate case, the expected order statistics of X^2_2 are (Cox and Lewis, 1966)

$$D^2_{(i)} = \sum_{j=1}^{i} \frac{2}{n-j+1} \qquad , \ i=1,\ldots,n$$

so a probability plot is easily carried out. To make it even simpler, Healy suggested a normal probability plot using $\sqrt{X^2}$ or $\sqrt[3]{X^2}$ In order to deal with the case of unknown μ and Σ, he proposes use of the familiar

$$(x-\bar{x})'S^{-1}(x-\bar{x}) \qquad (2.6.2)$$

but without comment on the effect of inserting these estimates.

The expression (2.6.2) has been denoted earlier by $R(x;\bar{x},S)$. Gnanadesikan and Kettenring (1972) consider similar graphical displays using the classes of measures

$$(x_j-\bar{x})'S^b(x_j-\bar{x})=R(x_j;\bar{x},S^{-b}) \qquad (2.6.3)$$

and

$$(x_j-\bar{x})'S^b(x_j-\bar{x})/(x_j-\bar{x})'(x_j-\bar{x})$$
$$=R(x_j;\bar{x},S^{-b})/R(x_j;\bar{x},I).$$

Different values of b may serve to highlight different outliers. These classes may be extended further by using the sample correlation matrix R in place of the sample covariance matrix S, and by using x_j-x_i $(j\neq i)$ in place of $x_j-\bar{x}$, so that inter-point distances are examined. For displaying these statistics, the basic tool is the gamma probability plot, based on the argument that the $R(x_j;\bar{x},\Gamma)$ for multivariate normally distributed data are approximately a set of independent gamma variates, whatever Γ is being used. (This was also employed by Rohlf in the gap test.) To carry out a gamma plot it is necessary to have an estimate of the parameter r in the distribution

$$f(x)=(\alpha/\Gamma(r))(\alpha x)^{r-1}e^{-\alpha x}, \qquad x>0.$$

(It is not necessary to estimate the scale parameter α,

because this only alters the slope of the entire plot, and has no effect on the departures from linearity which one is looking for in searching for outliers). Methods for estimating the shape parameter r are provided by Wilk, Gnanadesikan and Huyett (1962a, 1962b).

There is no closed-form estimator of r, so these plots are not particularly convenient to use in practice. It does not appear that much use is made of (2.6.3) for any choice other than b=-1, the usual generalized distance.

The only recent contribution to graphical detection of multivariate outliers is by Bacon-Shone and Fung (1987). They use Wilks' statistic directly, for specified numbers of outliers which can be greater than one (in which case, of course, Wilks' statistic is the same as the generalized distance). The methodology is as follows. To examine the sample for a given number of outliers, t, all $\binom{n}{t}$ ratios Λ_T are computed, where T is a set of t indices. If the standard asymptotic result from likelihood ratio testing is applied, the quantity

$$W_T=-\{n-(p+t+3)/2\}\ln \Lambda_T \qquad (2.6.4)$$

follows approximately the distribution X^2_{pt} for large n. Now Bacon-Shone and Fung observe that the interest for outlier detection lies in the largest values of W_T (smallest values of Λ_T). In particular, if there really are t>1 outliers, then there are (n-t+1) sets T containing the most extreme t-1 of these plus one other point, and these sets will contain the biggest values of W_T. Hence interest can be focussed on just the (n-t+1) largest W_T. Because most of these derive from sets with t-1 indices in common, Bacon-Shone and Fung say that the distribution of the W_T can be partitioned as

$$X^2_{p(t-1)}+X^2_p$$

where the first term is due to the common indices.

Consequently the largest n-t+1 expected quantiles of W_T can be approximated as a constant plus the quantiles of X^2_p corresponding to upper tail probabilities

$$k/(n-t+2), \qquad k=1,\ldots,n-t+1.$$

A plot of the (n-t+1) largest values of W_T against these quantiles should give an approximately linear plot; departures from linearity will suggest outliers. The above argument does not apply for t=1 outlier. In that case, the suggestion is to plot the values of Λ_j (j=1,...,n) against quantiles of the Beta distribution which describes such a ratio. The plots in the article, however, seem to use a plot for W again, not Λ.

The procedure is to produce a plot for each potential number of outliers up to a chosen maximum. If inspection of these does not clearly suggest how many and which points are outliers, Bacon-Shone and Fung suggest a sequential procedure, eliminating clear outliers and then looking at the reduced sample. A more formal sequential test procedure, based on Wilks' statistic, is the subject of Chapter 4 of the present thesis.

All methods mentioned so far provide displays of selected distances, rather than displays of the points themselves. Of methods for displaying multivariate data points, the most familiar is principal components analysis and this can be found in various forms in the multivariate outlier detection problem.

The customary use of principal components analysis (PCA) leads to a plot of the points in the space of the first two principal axes, if these account for a satisfactory percentage of the data. If one point appears to be well separated from the rest on this plot, it seems to be indicated as a possible outlier. However, outliers will not necessarily appear on the first few axes. The first few axes represent those linear transformations of

the original variables which have the largest variance. Hence they must tend to incorporate those original variables which have large variances or pairs with large covariances, if the analysis has been carried out on the covariance matrix, or to incorporate pairs of variables with large correlations if the correlation matrix was used for the analysis. Any outliers which affect variables other than those which would be strongly represented in the PCA of the uncontaminated data, are therefore unlikely to be seen in the space of the first few components. Outliers which are discernible there will tend to be those outliers whose effect is to increase the uncontaminated variances and covariances or correlations. It follows that other outliers must be sought elsewhere than in the first few components. These will be the outliers whose presence creates an apparent correlation, where none existed in the uncontaminated data, so add new dimensions to the principal components. The last few components in particular may also indicate another type of outlier, one which breaks a pattern that is so strong as to amount almost to collinearity. Since the last principal component is that linear combination of the original variables with minimum variance, any linear combination which is almost constant will be close to the last principal component. An outlier which breaks such patterns will be seen only if the last few components are examined. An application of this idea to provide a check on the accuracy of records being added to a data base is given by Hawkins (1974). The general topic of PCA and outliers is discussed in most detail by Gnanadesikan and Kettenring (1972). Note that the association of types of outliers with the first or last few principal components is given the wrong way round in Hawkins' (1980a) review.

The main drawback to use of PCA for outlier detection is that, for sensitivity to all kinds of outliers and outliers affecting all variables, it is necessary to

inspect a lot of components. On the other hand, the main advantage of PCA in most applications (with exceptions such as its use in regression: see Jolliffe, 1982) is that a large part of the information from a large number of variables is represented in just the first few components. Having lost this advantage, PCA is no longer an especially helpful method for multivariate outlier detection.

Besides plotting points in the space of selected components, Gnanadesikan and Kettenring also suggest probability plotting of scores on individual components. Even if the original data are not normally distributed, these scores may be reasonably close to normality and a normal probability plot can be carried out. Hawkins (1974, 1980a) looks at the possibility of more formal testing on the basis of scaled principal components residuals.

If $X \sim N(\mu,\Sigma)$ and the covariance (or correlation, in most applications) matrix Σ is diagonalized by the transformation C, so that $C\Sigma C'=\Lambda=\mathrm{diag}(\lambda_i)$, then the principal component residuals of a vector X_i are

$$Y_i=C(X_i-\mu) \sim N(0,\Lambda)$$

Hawkins rescales to

$$Z_i=\Lambda^{-(1/2)}Y_i \sim N(0,I)$$

and suggests statistics such as

$$\max_i |z_{ij}|, \qquad \sum_j z_{ij}^2,$$

based on z_i and then maximized over i. In practice, the above transformations will be carried out with estimates of μ and Σ rather than known values, so that the distribution of z_i is not normal. Hawkins states that the asymptotic normality result does not help for reasonable values of n, so that the scope for formal testing is in

fact very limited unless some distributional results applicable to small samples are discovered. The exception to this is the statistic

$$\max_i \left(\sum_j z_{ij}^2 \right)$$

since this is just Wilks' single-outlier statistic. PCA will be mentioned further in the following section.

In conclusion, it cannot be said that any one graphical method has emerged as being especially useful in the detection of multivariate outliers. Few of the ideas which have been suggested seem to be actually applied.

2.7 Robust estimation and influence

The inter-relationship of outliers and influence has already been mentioned. Since the emphasis here is on the topic of outliers, the relationship may be viewed here from that point of view, so that one can say that an outlier is usually influential in the sense that it has a much larger impact than other points do on the estimation of certain quantities. In particular, it is well known that estimates of correlations can easily be distorted substantially by the occurrence of outliers. It follows that methods of detecting influential points have a contribution to make to the detection of outliers, although of course influential points are not necessarily outliers.

One approach to the identification of influential points is to quantify the influence of each point on the statistic of interest. For a bivariate correlation coefficient r, Devlin, Gnanadesikan and Kettenring (1975) used as sample influence function simply

$$I(x_i;r)=(n-1)(r-r_i)$$

measuring the effect on the correlation of omitting x_i from the sample of size n, thus changing the correlation

from r in the full sample to r_i in the reduced sample. In the bivariate case, contours of constant I can then be superimposed on the scattergram, as in Example 3 of Chapter 1. Another suggestion is to look at the equivalent function for Fisher's transformation $z=\tanh^{-1}(r)$. This is approximately distributed as the product of two independent standard normals. A probability plot could be carried out to detect extreme values. If the greater familiarity of an ordinary normal probability plot helps, this can be achieved by transforming the ordered sample influence values $i_1 \leq \ldots \leq i_n$ to $\{v_j\}$ via

$$\Phi(v_j)=G(i_j), \qquad j=1,\ldots,n$$

where G is the distribution function of the product of standard normals and Φ is the standard normal distribution function (Gnanadesikan, 1977).

One difficulty in studying influence functions in multivariate problems is that one is often interested in a rather complex quantity, such as a largest eigenvalue in principal components analysis, so that it is hard to obtain distributional results. Nonetheless, there is some published work on influence in such contexts, for example Critchley (1985) on principal components analysis and Campbell (1978) for discriminant analysis.

As with the direct detection of outliers, so the detection of influential observations leaves the question of what to do with them once found. Robust estimation may provide the solution. A simple example of a robust estimator is a trimmed mean for estimating a univariate sample mean, in which a pre-selected number of the most extreme points at each end of the list of order statistics are discarded and only the remaining points are used. Outliers hence do not contribute to the estimation.

Rousseeuw (1989) has defined robust outlier detection

statistics which are simply the Mahalanobis distances (2.2.1) but with $\bar{x}$ and S replaced by robust estimates. Specifically, he recommends his minimum volume ellipsoid (MVE) esimators, which are based on the ellipsoid of smallest volume which contains at least half the points of the sample. These estimators give protection against a large proportion of contaminating points, so that masking is almost impossible. However, he recommends that, because of problems with collinearity, the MVE should not be used unless $n/p > 5$, so it is not a method applicable to small samples unless the dimensionality is also low.

Also of interest are methods which retain the full sample but may weight the points differently. If these weights are calculated from the sample, as opposed to being imposed, then a successful method downweights the more influential observations. Examination of the final weight for each point indicates which are the influential ones.

A method of this kind has been successfully applied by Campbell (1980) to the estimation of a covariance matrix, using M-estimators (Maronna, 1976). The problem had earlier been considered by Gnanadesikan and Kettenring (1972), who did not go so far as to specify what weight to use. Campbell takes as estimators of mean and covariances:

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

$$V = \frac{\sum w_i^2 (x_i - \bar{x})(x_i - \bar{x})'}{\sum w_i^2 - 1} \qquad (2.7.1)$$

where summations are over $i=1,\ldots,n$ and the weights w_i are obtained from a function

$$w_i = w(d_i) = \omega(d_i)/d_i$$

with

$$d_i^2=(x_i-\bar{x})'V^{-1}(x_i-\bar{x})$$

- an estimate of Mahalanobis distance. The equations (2.7.1) need to be iterated to a solution. The function ω controls the contribution of each point and Campbell uses the form:

$$\begin{aligned}\omega(d)&=d, && d\le d_0\\ &=d_0\exp\{(-1/2)(d-d_0)^2/b_2^2\}, && d>d_0.\end{aligned}$$

This means that influence increases linearly up to a certain point, but levels off ($b_2=\infty$) or begins to decline again to zero as distance from the mean increases further. Campbell recommends $(b_1,b_2)=(2,\infty)$ or $(2,1.25)$, where $d_0=\sqrt{p}+b_1/\sqrt{2}$.

Campbell goes on to consider a robust principal component analysis (RPCA). Although the obvious thing to do is carry out an ordinary PCA of the robust covariance matrix in (2.7.1), he rejects this for the following reason. A particular point's weight in (2.7.1) is a function of its distance from the robust $\bar{x}$. A given distance may be made up in various ways from contributions on different components. One way is for virtually all the distance to be in the direction of one component. If this is so, it is possible that a greater downweighting would be desirable to counteract this point's influence on this component. Consequently this method is used only to start off an iterative procedure, as follows.

First, V from (2.7.1) is used to provide initial estimates of the first eigenvector, giving associated first principal component scores y_i. M-estimation of the mean and variance of y is then carried out, giving a new set of weights w_i which are substituted into (2.7.1) to obtain a new $\bar{x}$ and V. This is repeated until a stable first principal component u_1 is obtained. After the first iteration, the weights w_i are taken as the minimum of the

current and previous weights, to avoid oscillation. The data are then transformed into values orthogonal to the space of u_1, the analysis repeated on the transformed data matrix and so on until all components have been derived. Ultimately, besides all the usual output of a PCA, the RPCA provides a list for each point of its weight in the estimation of each component.

Although this method looks fairly complex to carry out, it requires only standard matrix operations and is now easily available because Matthews (1984) has programmed it as a GENSTAT macro. Both Campbell and Matthews provide a full example of the application of the method. It seems that the detailed information it offers, together with its ease of use, may make RPCA the most valuable of all the informal methods of outlier detection. It could be used in all those situations where ordinary PCA is used for examination of the data. Campbell (1982) also developed a similar analysis for the more structured problem of canonical variates analysis.

Finally, one related point will be mentioned. Matthews' example used data published by Royston (1983) who had used them to illustrate his Ω test of multivariate normality, an extension of the univariate Shapiro-Wilk W test. Matthews discusses the relative merits of RPCA and Ω. Of course, RPCA as with most informal methods does not explicitly assume multivariate normality, but would not make much sense with data that were seriously non-normal. For this reason, and because the only practical general test statistic is for the normal case, it is reasonable to pay some attention to tests of multivariate normality as contributing to testing for outliers. However, Matthews suggests that these tests are not very powerful against the alternative of a normal distribution contaminated by a small number of outliers, and that the RPCA will be more informative about the nature of extreme points.

CHAPTER 3

WILKS' MULTIVARIATE OUTLIER TEST STATISTIC

3.1 A single outlier

As discussed in Chapter 2, the difficulties inherent in multivariate analysis mean that there has been no substantial addition to the literature on multivariate outlier testing since Wilks' (1963) basic contribution. His statistic can be expressed in various forms (see § 2.2) and its choice can be motivated in various ways, including Wilks' own volume argument and Ferguson's decision rule under the alternative hypothesis of slippage of variance. The derivation to be given here shows how the statistic is obtained by the two-stage maximum likelihood analysis of the popular slippage in the mean alternative, called model 1 in § 2.2. For a single outlier, the hypotheses are:

$$H_0: \quad x_i \sim N_p(\mu,\Sigma), \qquad i=1,\ldots,n$$

against

$$H_1: \quad x_i \sim N_p(\mu,\Sigma), \qquad i\neq j$$
$$\qquad x_j \sim N_p(\mu+a,\Sigma)$$

where j, a, μ and Σ are all unknown. The test statistic is found for a particular j and then its extreme over all choices of j is taken.

Under H_0, the likelihood is

$$\prod_{i=1}^{n} (|2\pi\Sigma|)^{-1/2}\exp\{(-1/2)(x_i-\mu)'\Sigma^{-1}(x_i-\mu)\}$$

so that the log-likelihood is

$$l(\mu,\Sigma)=(-np/2)\ln(2\pi)-(n/2)\ln|\Sigma|-(1/2)\sum_{i=1}^{n}(x_i-\mu)'\Sigma^{-1}(x_i-\mu) \qquad (3.1.1)$$

Now writing

$$x_i-\mu=x_i-\bar{x}+\bar{x}-\mu$$

the summation in third term becomes

$$\sum_{i=1}^{n}(x_i-\bar{x})'\Sigma^{-1}(x_i-\bar{x})+n(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu).$$

Furthermore, the first term is a scalar so equals its own trace,

$$\begin{aligned}
&\mathrm{tr}\{\sum_{i=1}^{n}(x_i-\bar{x})'\Sigma^{-1}(x_i-\bar{x})\}\\
&=\sum_{i=1}^{n}\mathrm{tr}(x_i-\bar{x})'\Sigma^{-1}(x_i-\bar{x})\\
&=\sum_{i=1}^{n}\mathrm{tr}\ \Sigma^{-1}(x_i-\bar{x})(x_i-\bar{x})'\\
&=\mathrm{tr}\ \Sigma^{-1}\{\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})'\}
\end{aligned}$$

where each step uses standard properties of traces. Hence, writing

$$nS=\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})'$$

for the sum of squares and products matrix, substituting in (3.1.1) gives

$$l(\mu,\Sigma)=-\frac{np}{2}\ln(2\pi)-\frac{n}{2}\ln|\Sigma|-\frac{n}{2}\mathrm{tr}(\Sigma^{-1}S)-\frac{n}{2}(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu) \qquad (3.1.2)$$

Now the last term involves a positive semi-definite quadratic form, which takes the value zero if and only if $\mu=\bar{x}$. No other term involves μ and therefore the maximum likelihood estimator of μ is $\hat{\mu}=\bar{x}$. The m.l.e. of Σ may be found by maximizing

$$l(\hat{\mu},\Sigma)=-\frac{np}{2}\ln(2\pi)-\frac{n}{2}\ln|\Sigma|-\frac{n}{2}\mathrm{tr}(\Sigma^{-1}S)$$

or

$$l(\hat{\mu},V)= - \frac{np}{2}\ln(2\pi)+\frac{n}{2}\ln|V| - \frac{n}{2}\mathrm{tr}(VS) \qquad (3.1.3)$$

where $V=\Sigma^{-1}$.

Standard matrix results (e.g. Mardia, Kent and Bibby, 1979, Appendix A) show that

$$\frac{\partial \ln|V|}{\partial V} = 2\Sigma-\mathrm{diag}\Sigma$$

and

$$\frac{\partial \mathrm{tr}(VS)}{\partial V} = 2S-\mathrm{diag}S$$

Hence, from (3.1.3),

$$\frac{\partial l}{\partial V} = 0 \quad \Rightarrow \quad 2M-\mathrm{diag}M=0$$

where $M=\Sigma-S$, and this can only be satisfied by $M=0$. Thus

$$\hat{\Sigma}=S.$$

It can be seen that with these m.l.e.'s, the maximized log-likelihood under the null hypothesis is

$$l_0= - \frac{np}{2}\ln(2\pi)-\frac{n}{2}\ln|\hat{\Sigma}| - \frac{np}{2} \qquad (3.1.4)$$

Under the alternative hypothesis, the log-likelihood is

$$\begin{aligned} l(\mu,\Sigma) &= -\frac{np}{2}\ln(2\pi)-\frac{n}{2}\ln|\Sigma|-\frac{1}{2}\sum_{i\neq j}(x_i-\mu)'\Sigma^{-1}(x_i-\mu) \\ &\qquad -(x_j-\mu-a)'\Sigma^{-1}(x_j-\mu-a) \\ &= -\frac{np}{2}\ln(2\pi)-\frac{n}{2}\ln|\Sigma|-\frac{(n-1)}{2}\mathrm{tr}(\Sigma^{-1}S_j) \\ &\qquad -\frac{(n-1)}{2}(\bar{x}_j-\mu)'\Sigma^{-1}(\bar{x}_j-\mu)-(x_j-\mu-a)'\Sigma^{-1}(x_j-\mu-a) \end{aligned}$$

where the subscript j in $\bar{x}_j$ and S_j denotes values in the reduced sample computed after x_j has been omitted from the n points. By taking

$$\hat{\hat{\mu}}=\bar{x}_j$$

and

$$\hat{a}=x_j-\hat{\hat{\mu}}$$

the fourth and fifth terms vanish. The remaining terms can be written as

$$l(\hat{\hat{\mu}},\Sigma)= -\frac{np}{2}\ln(2\pi)-\frac{n}{2}\ln|\Sigma| -\frac{n}{2}tr\{\Sigma^{-1}(n-1)S_j/n\}$$

so that equation (3.1.3) applies again with S replaced by $(n-1)S_j/n$. Consequently

$$\hat{\hat{\Sigma}} =(n-1)S_j/n$$

and the maximized log-likelihood is

$$l_1(\hat{\hat{\mu}},\hat{\hat{\Sigma}})= -\frac{np}{2}\ln(2\pi)-\frac{n}{2}\ln|\hat{\hat{\Sigma}}| -\frac{np}{2}$$

Comparing with (3.1.4), the change in maximized log-likelihood is

$$l_0-l_1= -\frac{n}{2}\ln|\hat{\Sigma}| +\frac{n}{2}\ln|\hat{\hat{\Sigma}}|$$

so that the likelihood ratio λ is given by

$$\lambda^{2/n}= \frac{|\hat{\hat{\Sigma}}|}{|\hat{\Sigma}|} = \frac{n-1}{n}\,\frac{|S_j|}{|S|}= \frac{|A_j|}{|A|} = \Lambda_j$$

where A_j and A are the sums of squares and products matrices for the reduced (by omission of x_j) and full samples. Thus the statistic Λ_j provides a test for H_0 against H_1 for specified j and, by taking its minimum over all choices of j, gives the outlier test statistic for unknown j.

For given j, the ratio Λ_j can be shown to have a beta distribution. This follows immediately from the well-known fact that Hotelling's T^2 statistic follows an F distribution, and from the relation (2.2.6) between Λ_j and T^2:

$$\Lambda_j^{-1} = 1+T_j^2/(n-2)$$

where $$\frac{(n-p-1)}{p(n-2)} T_j^2 \sim F_{p,n-p-1}$$

Hence

$$\frac{p(n-2)}{(n-p-1)T_j^2} \sim F_{n-p-1,p}$$

Now if $x \sim F_{a,b}$, then

$$\frac{ax}{(ax+b)} \sim B(a/2,b/2)$$

that is

$$(1+b/ax)^{-1} \sim B(a/2,b/2)$$

Therefore

$$\left(1+ \frac{p(n-p-1)T_j^2}{(n-p-1)p(n-2)}\right)^{-1} \sim B((n-p-1)/2,p/2)$$

and the left hand side is just

$$\left(1+ \frac{T_j^2}{(n-2)}\right)^{-1} = \Lambda_j$$

The outlier test statistic is the minimum over j of Λ_j. The distribution of this quantity has never been found. Wilks' answer was to use Bonferroni bounds, so that the percentage points for a conservative test at the α% level of significance are given by the lower α/n % points of the above Beta distribution. These values are tabulated by Wilks, for α=.01, .025, .05, .10 for 1 to 5 dimensions and for a selection of sample sizes up to 500. Parts of the

table are reproduced in Hawkins (1980a) and in Barnett and Lewis (1984).

It is generally accepted that, as in many other outlier problems, the true significance levels of these Bonferroni bounds are very close to the nominal ones. A partial check is possible through the fact that Wilks' statistic reduces to one of Grubbs' (1950) in the one-dimensional case. For certain values of n and α, the exact distribution is available and Wilks gives a table showing very close correspondence between these and the Bonferroni bounds. No exact distributions exist for two or more dimensions, so any further check on the accuracy of these bounds has to be by simulation. This check is included in Section 3.4.

3.2 Distributions of Λ for two or more outliers

Wilks went on to consider testing for two or more outliers in the sample. As indicated in Chapter 2, it is easy to see that the test statistic suitable for testing the hypothesis that a set T of points in the sample are drawn from populations whose means have slipped from the mean of the parent population (by amounts that are not assumed equal for each point) is

$$\Lambda_T = |A_T|/|A|$$

where A_T denotes the sum of squares and products matrix of the reduced sample consisting of the remaining points after the members of T have been omitted. This could be derived by the two-stage maximum likelihood procedure.

Wilks' analysis proceeded along the same lines as for the one-outlier case. As will be seen below, a Beta distribution again applies for the case of two outliers and Wilks gives Bonferroni bounds for conservative tests of significance. The distribution of a Λ ratio is more

complicated for higher numbers of outliers. A simple solution applicable for all p for each given number of outliers cannot be found. Wilks gives solutions for a small number of particular cases for 3 and 4 outliers. The distribution of the criterion for a specified set of t points is known to be the product of t independent Beta-distributed random variables (Anderson, 1958)

$$\Lambda \sim \prod_{i=1}^{t} B\left\{ \frac{n-p-i}{2} , \frac{p}{2} \right\} \qquad (3.2.1)$$

and this may be called the Λ distribution with parameters p, n-t-1 and t in the notation of Mardia, Kent and Bibby (1979, p.82):

$$\Lambda(p,n-t-1,t).$$

It is possible to simplify this distribution, as follows.

In the first place, it is easy to write down the moments of Λ. Because the terms in the product (3.2.1) are independent, the rth moment of Λ is just the product of the rth moments of each separate Beta in the product:

$$E(\Lambda^r)=E[(X_1 \ldots X_t)^r]=E(X_1^r)\ldots E(X_t^r)$$

Now any moment of a Beta-distributed random variable is just a product of gamma functions:

$$E(X^r)= \int_0^1 x^r \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}dx$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^{a+r-1}(1-x)^{b-1}dx$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+r)\Gamma(b)}{\Gamma(a+b+r)}$$

$$= \frac{\Gamma(a+b)\Gamma(a+r)}{\Gamma(a)\Gamma(a+r+b)}$$

$$= \frac{(a+r-1)\ldots(a+1)a}{(a+b+r-1)\ldots(a+b+1)(a+b)}$$

Hence $E(\Lambda^r)$ can be explicitly expressed in terms of the parameters a and b.

Secondly, it may be possible to recognize that the moments of Λ are equal to the moments of another product of random variables - particularly, the product of another set of Beta random variables, different from (3.2.1). If so, the standard theorem that equality of all moments implies that two distributions are identical can be applied. The point is that this second distribution may be easier to work with. To see how the method applies to the Λ distribution, consider two successive terms in the product (3.2.1)

$$B\left\{\frac{n-p-(2j-1)}{2}\ ,\ \frac{p}{2}\right\}\cdot B\left\{\frac{n-p-2j}{2}\ ,\ \frac{p}{2}\right\}$$

This product has rth moment:

$$\frac{\Gamma\left(\frac{n-2j+1}{2}\right)\ \Gamma\left(\frac{n-p-2j+1}{2}+r\right)\ \Gamma\left(\frac{n-2j}{2}\right)\ \Gamma\left(\frac{n-p-2j}{2}+r\right)}{\Gamma\left(\frac{n-2j+1}{2}+r\right)\ \Gamma\left(\frac{n-p-2j+1}{2}\right)\ \Gamma\left(\frac{n-2j}{2}+r\right)\ \Gamma\left(\frac{n-p-2j}{2}\right)}$$

$$= \frac{\left(\frac{n-p-2j+1}{2}+r-1\right)\cdots\left(\frac{n-p-2j+1}{2}\right)\cdot\left(\frac{n-p-2j}{2}+r-1\right)\cdots\left(\frac{n-p-2j}{2}\right)}{\left(\frac{n-2j+1}{2}+r-1\right)\cdots\left(\frac{n-2j+1}{2}\right)\ \cdot\ \left(\frac{n-2j}{2}+r-1\right)\cdots\left(\frac{n-2j}{2}\right)}$$

$$= \frac{(n-p-2j+1+2r-2)(n-p-2j+1+2r-4)\cdots(n-p-2j+1)}{(n-2j+1+2r-2)(n-2j+1+2r-4)\cdots(n-2j+1)}$$

$$\cdot\ \frac{(n-p-2j+2r-2)\cdots(n-p-2j)}{(n-2j+2r-2)\cdots(n-2j)}$$

$$= \frac{(n-p-2j+2r-1)\cdots(n-p-2j)}{(n-2j+2r-1)\cdots(n-2j)}$$

$$= \frac{\Gamma(n-p-2j+2r)\ \Gamma(n-2j)}{\Gamma(n-p-2j)\ \Gamma(n-2j+2r)}$$

This is the (2r)th moment of B(n-p-2j,p), and this has the same meaning as the rth moment of the square of B(n-p-2j,p). Hence each pair of Beta's in (3.2.1) can be expressed as the square of one Beta. If t is an odd number, one Beta from the original product is left over without another to be paired with. Hence the Λ distribution can be rewritten as

$$\Lambda \sim \begin{cases} \prod_{j=1}^{s} \{B(n-p-2j,p)\}^2, & t=2s \\ \prod_{j=1}^{s} \{B(n-p-2j,p)\}^2 . B(\frac{n-p-t}{2},\frac{p}{2}), & t=2s+1 \end{cases}$$

The terms are still independent, as in the original product.

This expression is simpler than the original because it contains fewer terms. In particular for the case t=2, $\Lambda=U^2$ where

$$U \sim B(n-p-2,p) \qquad (3.2.2)$$

This is why in the two-outlier case only a single distribution needs to be considered, as in the one-outlier case.

Three outliers

For the three-outlier case, the result is

$$\Lambda(p,n-4,3) = U^2V$$

where

$$U \sim B(n-p-2,p)$$

$$V \sim B(\frac{n-p-3}{2},\frac{p}{2})$$

The p.d.f. is

$$f(u,v) = \frac{\Gamma(n-2)}{\Gamma(n-p-2)\,\Gamma(p)} \cdot u^{(n-p-3)}(1-u)^{p-1} \cdot \frac{\Gamma((n-3)/2)}{\Gamma((n-p-3)/2)\,\Gamma(p/2)} v^{((n-p-3)/2)-1}(1-v)^{(p/2)-1}$$

$$0 \le u \le 1,\ 0 \le v \le 1$$

Percentage points are found by solving

$$P(r) = P(U^2V \le r) = \alpha$$

This does not lead to a simple general result along the lines of (3.2.2), but solutions can be derived for each particular case. Proceeding as in Wilks' paper

$$P(r) = \iint_D f(u,v)\,du\,dv$$

where D is the region shown in Figure 3.2.1

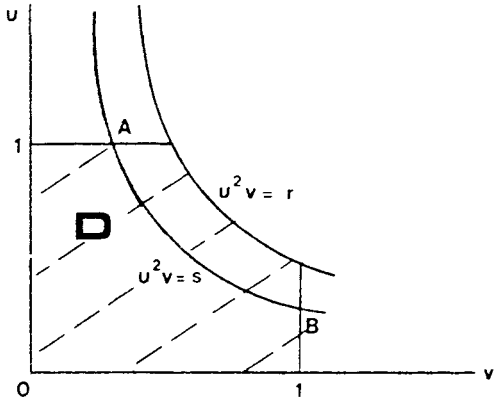


Figure 3.2.1 Transforming to (s,u), where $s=u^2v$, the range of integration in D is s=0 to r and $u=\sqrt{s}$ (point B) to 1 (point A).

Substituting $s=u^2v$,

$$P(r) = \int_{s=0}^{r} \int_{u=\sqrt{s}}^{1} f\left(u, \frac{s}{u^2}\right) \frac{1}{u^2} \, du\,ds$$

$$\propto \int_{s=0}^{r} \int_{u=\sqrt{s}}^{1} u^{n-p-3}(1-u)^{p-1} s^{(n-p-5)/2} u^{-(n-p-5)} \left(1-\frac{s}{u^2}\right)^{(p/2)-1} u^{-2} \, du\,ds$$

$$\propto \int_{s=0}^{r} s^{(n-p-5)/2} \int_{u=\sqrt{s}}^{1} (1-u)^{p-1} \left(1-\frac{s}{u^2}\right)^{(p/2)-1} du\,ds$$

where the constant of proportionality is the product of gamma functions. Now if p is even, the term $(1-s/u^2)^{(p/2)-1}$ can be expanded in powers of (s/u^2) and the integration is easy. Wilks gives the solution for the case p=2, namely

$$P(r) = \frac{(n-3)(n-4)(n-5)}{2} (\sqrt{r})^{n-5} \left\{ \frac{1}{n-5} - \frac{2\sqrt{r}}{n-4} + \frac{r}{n-3} \right\} \qquad (3.2.3)$$

For the case p=4, the distribution is a constant times

$$\int_{s=0}^{r} s^{(n-9)/2} \int_{u=\sqrt{s}}^{1} (1-u)^3 (1-s/u^2) \, du\,ds$$

The integral over u is

$$\int_{u=\sqrt{s}}^{1} (1-3u+3u^2-u^3)(1-s/u^2) \, du$$

$$= \frac{1}{4} - 2\sqrt{s} + 2s^{3/2} - \frac{s^2}{4} - \frac{3s}{2} \ln(s)$$

leading to

$$\int_{s=0}^{r} \left\{ \frac{s^{(n-9)/2}}{4} - 2s^{(n-8)/2} + 2s^{(n-6)/2} - \frac{s^{(n-5)/2}}{4} - \frac{3s^{(n-7)/2}}{2} \ln(s) \right\} ds$$

$$= \left[\frac{s^{(n-7)/2}}{2(n-7)} - \frac{4s^{(n-6)/2}}{n-6} + \frac{4s^{(n-4)/2}}{n-4} - \frac{s^{(n-3)/2}}{2(n-3)} \right]_0^r - \frac{3}{2} \int_0^r s^{(n-7)/2} \ln(s) \, ds$$

The second integral is

$$\left[\frac{2}{n-5}^{(n-5)/2}\ln(s)\right]_0^r - \int_0^r \frac{2}{n-5} s^{(n-5)/2} \frac{1}{s}\,ds$$

$$= \frac{2r^{(n-5)/2}\ln(r)}{n-5} - \frac{2}{n-5}\left[\frac{2}{n-5} s^{(n-5)/2}\right]_0^r$$

Finally,

$$P(r)=kr^{(n-7)/2}\left\{\frac{1}{2(n-7)} - \frac{4r^{1/2}}{n-6} + \frac{4r^{3/2}}{n-4} - \frac{r^2}{2(n-3)} + \frac{6r}{(n-5)^2} - \frac{3r\ln(r)}{n-5}\right\} \qquad (3.2.4)$$

where $k=\dfrac{\Gamma(n-2)\Gamma((n-3)/2)}{\Gamma(n-p-2)\Gamma((n-p-3)/2)\Gamma(p)\Gamma(p/2)}$

$$=(n-3)(n-4)(n-5)^2(n-6)(n-7)/24$$

since p=4.

If p is odd, Wilks' method does not work out. An alternative given by Anderson can be used (1958, § 8.5.3). In this method, the integral $P(r)=P(U^2V\leq r)$ is derived as the sum of the areas A and B in Figure 3.2.2

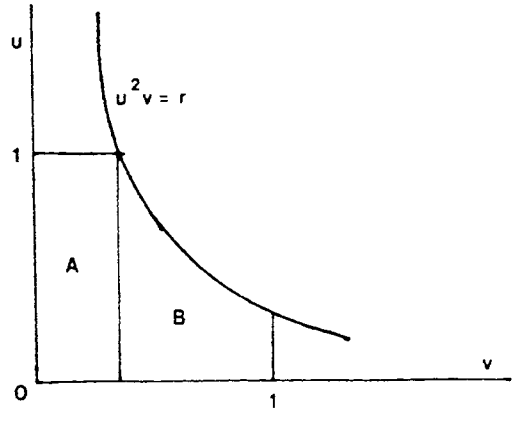


Figure 3.2.2 Integration of $P(U^2V\leq r)$

$$\text{Area } A = \int_{v=0}^{r} \int_{u=0}^{1} f(u,v)\,du\,dv = \int_{v=0}^{r} f(v)\,dv$$

$$= I_r\{(n-p-3)/2, p/2\},$$

the incomplete Beta integral.

$$\text{Area } B = \int_{v=r}^{1} \int_{u=0}^{\sqrt{r/v}} f(u,v)\ du\,dv$$

$$= k \int_{v=r}^{1} \left\{ \int_{u=0}^{\sqrt{r/v}} u^{n-p-3}(1-u)^{p-1}du \right\} v^{(n-p-5)/2}(1-v)^{(p-2)/2}\ dv$$

Now expand $(1-u)^{p-1}$, so the integral over u becomes

$$\int_{u=0}^{\sqrt{r/v}} u^{n-p-3} \sum_{i=0}^{p-1} \binom{p-1}{i} (-u)^i\ du$$

$$= \sum_{i=0}^{p-1} \binom{p-1}{i} (-1)^i \int_{u=0}^{\sqrt{r/v}} u^{n-p-3+i}\ du$$

$$= \sum_{i=0}^{p-1} \binom{p-1}{i} (-1)^i \left[\frac{u^{n-p-2+i}}{n-p-2+i} \right]_{u=0}^{\sqrt{r/v}}$$

$$= \sum_{i=0}^{p-1} \binom{p-1}{i} (-1)^i \frac{r^{(n-p-2+i)/2}}{n-p-2+i}\ v^{-(n-p-2+i)/2}$$

so Area B is

$$k \sum_{i=0}^{p-1} \binom{p-1}{i} (-1)^i \frac{r^{(n-p-2+i)/2}}{n-p-2+i} \cdot$$

$$\cdot \int_{v=r}^{1} v^{-(n-p-2+i)/2} v^{(n-p-5)/2} (1-v)^{(p-2)/2} dv$$

and the integral is

$$\int_{v=r}^{1} v^{-(i+3)/2}(1-v)^{(p-2)/2}dv$$

For even p, $(1-v)^{(p-2)/2}$ could be expanded as before, just as in Wilks' method. For odd p, expand $(1-v)^{(p-3)/2}$ in powers of v, leaving over a factor $\sqrt{1-v}$. The integrand is therefore the sum of powers of v multiplied by this factor, and can always be solved by standard substitutions and lengthy, routine algebra. Anderson works out the case p=3:

$$\int_{v=r}^{1} v^{-(i+3)/2}(1-v)^{(p-2)/2}\,dv \qquad \text{for} \quad i=0,1,2$$

$$= \int_{v=r}^{1} v^{-3/2}(1-v)^{1/2}\,dv \;;$$

$$\int_{v=r}^{1} v^{-2}(1-v)^{1/2}\,dv; \qquad \text{and}$$

$$\int_{v=r}^{1} v^{-5/2}(1-v)^{1/2}\,dv.$$

Finally

$$P(r)=I_r\left\{\frac{n-3}{2},\frac{3}{2}\right\}+kr^{(n-6)/2}\left[\frac{2\sqrt{1-r}}{(n-4)(n-5)}+\frac{2\sqrt{r}\{\sin^{-1}(2r-1)-(\pi/2)\}}{(n-5)}\right.$$

$$\left.+\frac{2r}{n-4}\ln\left(\frac{1+\sqrt{1-r}}{\sqrt{r}}\right)+\frac{2(1-r)^{3/2}}{3(n-3)}\right] \qquad (3.2.5)$$

where

$$k=\frac{(n-3)(n-4)(n-5)}{\sqrt{\pi}}\frac{\Gamma((n-3)/2)}{\Gamma((n-6)/2)}$$

Four outliers

Turning now to the case of 4 outliers, the Λ distribution can be rewritten as $\Lambda=U^2V^2$, where

$$U \sim B(n-p-2,p)$$

$$V \sim B(n-p-4,p)$$

so that

$$P(r)=P(U^2V^2 \leq r)$$

$$= \int_D \int f(u,v)\ dudv$$

where the area D is shown in Figure 3.2.3.

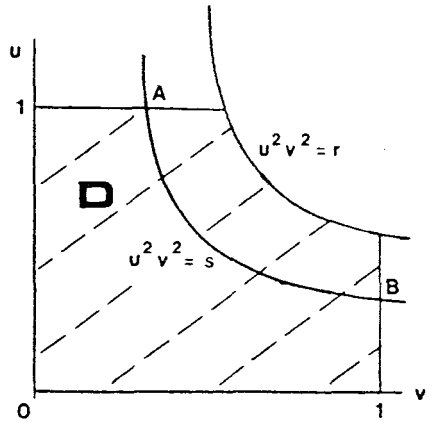


Figure 3.2.3 Transforming to (s,u), where $s=u^2v^2$, the range of integration in D is s=0 to r and $u=\sqrt{s}$ (point B) to 1 (point A).

Hence,

$$P(r)= \int_{s=0}^{r} \int_{u=\sqrt{s}}^{1} f(u,\sqrt{s}/u)\,(2u\sqrt{s})^{-1}\ duds$$

$$=k \int_{s=0}^{r} \int_{u=\sqrt{s}}^{1} u^{n-p-3}(1-u)^{p-1}(\sqrt{s}/u)^{n-p-5}(1-\sqrt{s}/u)^{p-1}u^{-1}s^{-1/2}\ duds$$

$$=k \int_{s=0}^{r} s^{(n-p-6)/2} \int_{u=\sqrt{s}}^{1} u\{(1-u)(1-\sqrt{s}/u)\}^{p-1}\ duds$$

as given by Wilks (eq. 4.18), where

$$k=\frac{1}{2}\,\frac{\Gamma(n-2)}{\Gamma(n-p-2)\Gamma(p)}\,\frac{\Gamma(n-4)}{\Gamma(n-p-4)\Gamma(p)}$$

As in the three-outlier problem, each particular case can be solved explicitly. Wilks gives the solution for p=2, namely

$$P(r)=\frac{(n-3)!}{6(n-7)!}(\sqrt{r})^{n-6}\left\{\frac{1}{n-6}-\frac{3\sqrt{r}}{n-5}+\frac{3r}{n-4}-\frac{(\sqrt{r})^3}{n-3}\right\} \qquad (3.2.6)$$

For p=3,

$$P(r)\propto\int_{s=0}^{r} s^{(n-9)/2}\int_{u=\sqrt{s}}^{1} u(1-2u+u^2)(1-2\sqrt{s}/u+s/u^2)\,du\,ds$$

Expanding and integrating over u gives

$$P(r)\propto\frac{1}{12}\int_{s=0}^{r}\{s^{(n-9)/2}-8s^{(n-8)/2}+8s^{(n-6)/2}-s^{(n-5)/2}-6s^{(n-7)/2}\ln(s)\}\,ds$$

$$=\frac{1}{12}\,r^{(n-7)/2}\left\{\frac{2}{n-7}-\frac{16\sqrt{r}}{n-6}+\frac{16r\sqrt{r}}{n-4}-\frac{2r^2}{n-3}\right\}-\frac{1}{2}\int_{s=0}^{r} s^{(n-7)/2}\ln(s)\,ds$$

The second term integrates by parts to give

$$\frac{2r^{(n-5)/2}}{n-5}\ln(r)-\frac{4}{(n-5)^2}\,r^{(n-5)/2}$$

so finally

$$P(r)=kr^{(n-7)/2}\left\{\frac{2}{n-7}-\frac{16\sqrt{r}}{n-6}+\frac{16r\sqrt{r}}{n-4}-\frac{2r^2}{n-3}-\frac{12r\ln(r)}{n-5}+\frac{24r}{(n-5)^2}\right\}$$

where $$k=\frac{1}{96}(n-3)(n-4)(n-5)^2(n-6)(n-7)$$

This is exactly the same as the result (3.2.4) for the case p=4 and t=3. This equivalence is a particular illustration of the result (3.3.3) to be given in the next section.

For the case p=4,

$$P(r) \propto \int_{s=0}^{r} s^{(n-10)/2} \int_{u=\sqrt{s}}^{1} u\{(1-u)(1-\sqrt{s}/u)\}^3 \, du\,ds$$

$$= \int_{s=0}^{r} s^{(n-10)/2} \left[\frac{1}{20} - \frac{3\sqrt{s}}{4} - 4s + 4s\sqrt{s} + \frac{3s^2}{4} - \frac{s^2\sqrt{s}}{20} - \frac{3s\ln(s)}{2} - \frac{3s\sqrt{s}\ln(s)}{2}\right] ds$$

and the integration gives finally

$$P(r) = kr^{(n-8)/2} \left[\frac{1}{n-8} - \frac{15\sqrt{r}}{n-7} - \frac{80r}{n-6} + \frac{60r}{(n-6)^2} + \frac{80r\sqrt{r}}{n-5} + \frac{15r^2}{n-4} - \frac{r^2\sqrt{r}}{n-3} + \frac{60r\sqrt{r}}{(n-5)^2} - \frac{30r\ln(r)}{n-6} - \frac{30r\sqrt{r}\ln(r)}{n-5}\right] \qquad (3.2.8)$$

where $k=(n-3)(n-4)(n-5)^2(n-6)^2(n-7)(n-8)/720$

The performance of Bonferroni percentage points derived from the distributions worked out for these particular cases will be examined below in § 3.4.

3.3 Exact and approximate F distributions for Λ

Because of the close relationship between the Beta and F distributions, it is not surprising that the F distribution can be used in relation to Λ. One special case, wherein the F distribution applies exactly to a simple function of Λ, is provided by the case p=2 and any number of outliers. In general, the distribution Λ(2,r,s) can be transformed exactly to F as follows:

$$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}} \sim \frac{s}{r-1} F_{2s,2(r-1)} \qquad (3.3.1)$$

(e.g. Mardia, Kent and Bibby, 1979, equation 3.7.10). In the present application, we have Λ(2,n-t-1,t) for the t-outlier problem, so that

$$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}} \sim \frac{t}{n-t-2} F_{2t,2(n-t-2)}$$

or

$$\Lambda \sim \left\{1+ \frac{t}{n-t-2} F_{2t,2(n-t-2)}\right\}^{-2} \qquad (3.3.2)$$

Percentage points for Λ obtained in this way for t=3 are exactly the same as can be derived from (3.2.3).

The above result can be derived by applying the result

$$\Lambda(p,r,s)=\Lambda(s,r+s-p,p) \qquad (3.3.3)$$

which can be shown by rewriting the ratio of determinants in Λ as the ratio of determinants of two other matrices after a suitable orthonormal rotation, as shown in Theorem 3.7.4 of Mardia, Kent and Bibby (1979). In the t-outlier application,

$$\Lambda(2,n-t-1,t)=\Lambda(t,n-3,2)$$

But this is the Λ criterion for a two-outlier problem, and we have already seen that its distribution, given by the product of two independent Beta's, can be re-expressed in terms of a single Beta. Specifically, $\Lambda(t,n-3,2)=U^2$ where

$$U \sim B(n-t-2,t)$$

from (3.2.2). Now the standard definition of the Beta distribution as a transformation of the F distribution gives

$$\frac{2tU}{2(n-t-2)(1-U)} \sim F_{2(n-t-2),2t}$$

or, taking the reciprocal,

$$\frac{(n-t-2)(1-U)}{tU} \sim F_{2t,2(n-t-2)}$$

so that

$$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}} \sim \frac{t}{n-t-2} F_{2t,2(n-t-2)}$$

as above.

For the Λ criterion with other parameters (excepting

the cases p=1 and t=1), exact distributional results in terms of the F or related distributions do not exist. However, approximations may be found. An obvious one exploits the derivation of Λ as a likelihood-ratio criterion, which implies that the standard asymptotic result applies, expressing the log-likelihood ratio as proportional to a X^2. Specifically, for the t-outlier problem,

$$-\{n-t-1-(p-t+1)/2\}\ \ln \Lambda(p,n-t-1,t) \sim X^2_{pt}$$

as $n \to \infty$. The multiplying factor includes an adjustment given by Box (1949). Further adjustments are given in the tables of Pearson and Hartley (1972). This result was seen earlier as (2.6.4) and was used by Bacon-Shone and Fung (1987) in their graphical method of searching for outliers.

A better approximation, due to Rao (1951, 1973), gives an asymptotic F distribution for a function of Λ. Applied to the t-outlier criterion, $\Lambda(p,n-t-1,t)$, the result is

$$\frac{ms-2\lambda}{pt}.\frac{(1-\Lambda^{1/s})}{\Lambda^{1/s}} \sim F_{pt,ms-2\lambda} \qquad (3.3.4)$$

where $\lambda=(pt-2)/4$

$m=n-1-(p+t+1)/2$

$s^2=(p^2t^2-4)/(p^2+t^2-5)$

The degrees of freedom are not necessarily integers. This approximation is employed to test the Λ criterion when it is used in other multivariate problems (for example, MANOVA and discriminant analysis) in well-known statistical packages such as BMDP and SPSS.

The following illustrative results give an idea of the accuracy of Rao's approximation by comparing approximate percentage points from (3.3.4) with exact ones obtained from particular cases for t=3 and 4 outliers worked out in

§ 3.2. Notice that the approximation (3.3.4) reduces to the exact expression (3.3.1) for p=2, so this case will not be considered further. Results from the F approximation in Table 3.3.1 were obtained by obtaining percentage points of F using the IMSL routine MDFI, then transforming to percentage points of Λ as in (3.3.4). Exact percentage points were obtained by solving equations (3.2.4), (3.2.5), (3.2.8) using Newton-Raphson iteration (Appendix I).

Table 3.3.1 Percentage points for Λ derived from Rao's F approximation (upper line), in comparison to points derived from exact distributions (lower line).

	n=10				n=20			
Case	1%	2.5%	5%	10%	1%	2.5%	5%	10%
p=3, t=3 (1)	.0155	.0262	.0397	.0615	.2448	.2909	.3336	.3858
	.0155	.0262	.0396	.0614	.2448	.2909	.3336	.3858
p=3, t=4 (2)	.0029	.0058	.0100	.0176	.1709	.2078	.2429	.2872
p=4, t=3 (2)	.0029	.0057	.0098	.0174	.1709	.2077	.2429	.2872
p=4, t=4 (3)	.00019	.00049	.0010	.0023	.1064	.1330	.1591	.1932
	.00016	.00043	.0009	.0021	.1064	.1329	.1591	.1932

(1) Exact distribution from equation (3.2.5)
(2) Equation (3.2.4)
(3) Equation (3.2.8)

3.4 Simulation studies of Wilks' statistic

The purpose of the studies described in this section is twofold - to provide a check on the accuracy of Bonferroni bounds for Wilks' statistic and to provide tables of simulated percentage points as an alternative to the Bonferroni bounds for one or two outliers. The design of the simulations for examining the one-outlier and two-outlier cases was as follows. Samples of size n= 10, 15, 20, 25, 30, 40, 50, 75 and 100 were examined, with p= 2, 3, 4 and 5. For each combination of n and p, 40,000

samples were generated, in five batches of 8,000. Each batch started from a different seed for the IMSL subroutine GGNSM which was used for the generation of multivariate normal data. For each sample of n independent and identically distributed vectors, the following were recorded:

the value of Wilks' one-outlier statistic;
the value of Wilks' two-outlier statistic;
whether or not the Bonferroni bound was exceeded by the one-outlier statistic;
whether or not the Bonferroni bound was exceeded by the two-outlier statistic.

From the distributions of the values of the one- and two-outlier statistics, 1, 2.5, 5 and 10% points were obtained. These are presented in Tables 3.4.1 (a-d) and 3.4.2 (a-d). The correspondence to the Bonferroni bounds is very close for t=1 (Table 3.4.1) but less so for t=2 (Table 3.4.2). How important the discrepancies are is illustrated by Tables 3.4.3 (a-d) and 3.4.4 (a-d), which shows how often the Bonferroni percentage points were in fact exceeded. For t=1, these percentages are extremely close to the nominal values. For t=2, on the other hand, it can be seen that there are sizeable departures, becoming more marked as n increases. Even at n=20, the nominal 10% level test actually is at only a little over half of that. As claimed by Hawkins, these are indeed rather poor approximations.

Because of the very heavy computing which is involved, the cases of three and four outliers have been investigated in less detail. Results (simulated percentage points and simulated exceedence probabilities) were obtained for the particular cases whose distributions were worked out in § 3.2. That is, for 3 outliers, simulations were carried out for p= 2, 3 and 4 dimensions, using sample sizes of n=10 and 20. For 4 outliers,

samples of size 10 were not considered because it is not very realistic to test for 4 outliers in 10 points. Also, the case of p=3 for 4 outliers is identical to p=4 for 3 outliers: hence simulations for 4 outliers were carried out only for p=2 and p=4, for n=20. Simulations were carried out as for one and two outliers, but using single batches of 2000 samples for each combination of n and p.

Simulated percentage points are shown in Table 3.4.5 and exceedence probabilities in Table 3.4.6. It can be seen that the exceedence probabilities for given n and p continue to decrease as the number of outliers being examined increases, although the differences in results between 3 and 4 outliers or between 2 and 3 outliers (comparing with Table 3.4.2) generally seem to be less dramatic than between 1 and 2 outliers. The slight improvement as the number of dimensions p increases for fixed n and number of outliers can also be seen, as in Table 3.4.2.

Discrepancies of this kind do not render Wilks' test with Bonferroni bounds unusable for practical purposes. The Bonferroni bound has the great virtue of providing a conservative test: that is, the true significance level does not exceed the nominal level. If a test result clearly gives evidence against the null hypothesis, say with p=0.01, then the true situation is that the evidence is even stronger than this. The difficulty comes when the evidence appears less clear. A poor Bonferroni bound means that these cases are not being held to be as good evidence against the null hypothesis as they in fact are, so that sensitivity is lost here for t>1. Consequently it is desirable to look for other tests, or for other, improved ways of implementing this test.

Table 3.4.1a Simulated percentage points for Wilks' one-outlier test statistic based on 40,000 simulations, α=0.01. Bonferroni bounds in parentheses.

Sample size n	Dimensions,p 2	3	4	5
10	.13712	.07753	.03809	.01604
	(.13895)	(.07781)	(.03866)	(.01523)
15	.29303	.22458	.16698	.12089
	(.29556)	(.22330)	(.16678)	(.12128)
20	.40614	.34473	.28456	.23258
	(.40893)	(.34019)	(.28354)	(.23506)
25	.49229	.42859	.37315	.32618
	(.49102)	(.42815)	(.37513)	(.32861)
30	.55570	.49735	.44573	.39850
	(.55263)	(.49547)	(.44663)	(.40320)
40	.64052	.58942	.54808	.50976
	(.63870)	(.59091)	(.54949)	(.51217)
50	.69596	.65491	.61910	.58554
	(.69598)	(.65514)	(.61947)	(.58711)
75	.78057	.75197	.72259	.69860
	(.78048)	(.75062)	(.72432)	(.70026)
100	.82810	.80191	.78214	.76150
	(.82704)	(.80352)	(.78271)	(.76361)

Table 3.4.1b. Simulated percentage points for Wilks' one-outlier test statistic, α=0.025. Bonferroni bounds in parentheses.

Sample size, n	Dimensions, p 2	3	4	5
10	.17942	.10541	.05492	.02442
	(.18053)	(.10601)	(.05606)	(.02420)
15	.34454	.26479	.20181	.14952
	(.34433)	(.26485)	(.20171)	(.14998)
20	.45361	.38503	.32477	.26718
	(.45547)	(.38281)	(.32239)	(.27020)
25	.53195	.46936	.41214	.36208
	(.53367)	(.46860)	(.41336)	(.36460)
30	.59218	.53237	.48233	.43486
	(.59144)	(.53303)	(.48287)	(.43806)
40	.67226	.62473	.57988	.54279
	(.67113)	(.62301)	(.58117)	(.54333)
50	.72405	.68258	.64652	.61403
	(.72365)	(.68286)	(.64715)	(.61466)
75	.80030	.77250	.74465	.72003
	(.80060)	(.77108)	(.74503)	(.72116)
100	.84337	.81949	.79867	.77848
	(.84281)	(.81967)	(.79916)	(.78030)

Table 3.4.1c. Simulated percentage points for Wilks' one-outlier test statistic, α=0.05. Bonferroni bounds in parentheses.

Sample size, n	Dimensions, p 2	3	4	5
10	.21947	.13396	.07424	.03461
	(.22007)	(.13408)	(.07438)	(.03440)
15	.38716	.30272	.23472	.17666
	(.38650)	(.30154)	(.23319)	(.17642)
20	.49340	.42034	.35661	.29885
	(.49417)	(.41876)	(.35558)	(.30060)
25	.56646	.50299	.44477	.39437
	(.56838)	(.50188)	(.44513)	(.39477)
30	.62288	.56302	.51183	.46402
	(.62260)	(.56347)	(.51248)	(.46674)
40	.69678	.64892	.60737	.56801
	(.69675)	(.64857)	(.60654)	(.56843)
50	.74497	.70579	.66907	.63609
	(.74532)	(.70472)	(.66909)	(.63659)
75	.81572	.78804	.76068	.73700
	(.81616)	(.78700)	(.76122)	(.73754)
100	.85500	.83218	.81203	.79267
	(.85494)	(.83214)	(.81192)	(.79329)

Table 3.4.1d. Simulated percentage points for Wilks' one-outlier test statistic, α=0.10. Bonferroni bounds in parentheses.

Sample size, n	Dimensions, p 2	3	4	5
10	.26941	.17018	.10003	.05014
	(.26827)	(.16978)	(.09888)	(.04901)
15	.43524	.34415	.27175	.20846
	(.43383)	(.34358)	(.26995)	(.20789)
20	.53606	.46031	.39481	.33535
	(.53615)	(.45832)	(.39255)	(.33484)
25	.60564	.53940	.48098	.42835
	(.60535)	(.53774)	(.47967)	(.42784)
30	.65673	.59642	.54477	.49673
	(.65540)	(.59583)	(.54420)	(.49768)
40	.72387	.67636	.63487	.59500
	(.72335)	(.67531)	(.63326)	(.59499)
50	.76768	.72880	.69214	.65978
	(.76763)	(.72738)	(.69195)	(.65955)
75	.83265	.80488	.77835	.75482
	(.83203)	(.80331)	(.77787)	(.75446)
100	.86759	.84561	.82538	.80707
	(.86725)	(.84486)	(.82497)	(.80663)

Table 3.4.2a. Simulated percentage points for Wilks' two-outlier test statistic, based on 40,000 simulations, α=0.01. Bonferroni bounds in parentheses.

Sample size, n	Dimensions, p 2	3	4	5
10	.19038	.10923	.05397	.01949
	(.18308)	(.10490)	(.05181)	(.01887)
15	.36909	.28633	.21784	.16237
	(.35641)	(.27605)	(.21050)	(.15589)
20	.48081	.41003	.35061	.28856
	(.46935)	(.39764)	(.33655)	(.28296)
25	.56085	.49622	.43996	.38769
	(.54722)	(.48390)	(.42895)	(.37975)
30	.61999	.56010	.50813	.46319
	(.60409)	(.54778)	(.49839)	(.45370)
40	.69701	.64818	.60598	.56906
	(.68184)	(.63596)	(.59527)	(.55804)
50	.74569	.70421	.66935	.63730
	(.73276)	(.69411)	(.65962)	(.62789)
75	.81708	.78898	.76267	.73906
	(.80704)	(.77928)	(.75434)	(.73124)
100	.85581	.83370	.81357	.79457
	(.84771)	(.82601)	(.80645)	(.78828)

Table 3.4.2b. Simulated percentage points for Wilks' two-outlier statistic, α=0.025. Bonferroni bounds in parentheses.

Sample size, n	Dimensions, p 2	3	4	5
10	.22653	.13374	.06845	.02706
	(.21444)	(.12703)	(.06572)	(.02579)
15	.40748	.31875	.24637	.18559
	(.38897)	(.30468)	(.23522)	(.17670)
20	.51505	.44148	.37700	.31769
	(.49859)	(.42504)	(.36203)	(.30642)
25	.59166	.52457	.46714	.41476
	(.57307)	(.50887)	(.45292)	(.40262)
30	.64598	.58485	.53450	.48690
	(.62707)	(.57036)	(.52046)	(.47517)
40	.71790	.66907	.62717	.59004
	(.70050)	(.65465)	(.61389)	(.57651)
50	.76347	.72287	.68860	.65623
	(.74841)	(.70994)	(.67555)	(.64386)
75	.82964	.80233	.77544	.75254
	(.81815)	(.79066)	(.76593)	(.74299)
100	.86552	.84398	.82396	.80567
	(.85632)	(.83487)	(.81552)	(.79753)

Table 3.4.1c. Simulated percentage points for Wilks' two-outlier test statistic, α=0.05. Bonferroni bounds in parentheses.

Sample size, n	Dimensions, p 2	3	4	5
10	.25802	.15787	.08448	.03508
	(.24188)	(.14702)	(.07881)	(.03273)
15	.43786	.34582	.27082	.20578
	(.41573)	(.32853)	(.25609)	(.19451)
20	.54389	.46764	.40133	.34062
	(.52205)	(.44722)	(.38282)	(.32571)
25	.61560	.54863	.49019	.43706
	(.59354)	(.52878)	(.47215)	(.42108)
30	.66633	.60716	.55529	.50720
	(.64513)	(.58821)	(.53799)	(.49230)
40	.73393	.68707	.64451	.60701
	(.71501)	(.66926)	(.62850)	(.59106)
50	.77782	.73819	.70293	.67097
	(.76052)	(.72224)	(.68798)	(.65635)
75	.83975	.81290	.78662	.76394
	(.82670)	(.79944)	(.77489)	(.75210)
100	.87367	.85237	.83271	.81434
	(.86292)	(.84169)	(.82251)	(.80467)

Table 3.4.1d Simulated percentage points for Wilks' one-outlier test statistic, α=0.05. Bonferroni bounds in parentheses.

Sample size, n	Dimensions, p 2	3	4	5
10	.29495	.18555	.10396	.04607
	(.27305)	(.17039)	(.09468)	(.04161)
15	.47341	.37748	.29925	.23001
	(.44453)	(.35452)	(.27909)	(.21439)
20	.57396	.49673	.42848	.36685
	(.54676)	(.47079)	(.40506)	(.34649)
25	.64128	.57500	.51571	.46230
	(.61487)	(.54966)	(.49243)	(.44064)
30	.69012	.63077	.57852	.53009
	(.66380)	(.60677)	(.55631)	(.51026)
40	.75260	.70556	.66393	.62586
	(.72990)	(.68431)	(.64361)	(.60614)
50	.79251	.75448	.71883	.68725
	(.77288)	(.73485)	(.70074)	(.66921)
75	.85073	.82382	.79849	.77547
	(.83537)	(.80837)	(.78403)	(.76141)
100	.88220	.86095	.84152	.82377
	(.86960)	(.84859)	(.82961)	(.81193)

Table 3.4.3a. Simulated null probability of obtaining a value of Wilks'one-outlier test statistic less than the Bonferroni approximation at α=0.01, based on 40,000 simulations.

Sample size, n	Dimensions, p 2	3	4	5
10	.0109	.0101	.0103	.0090
15	.0105	.0098	.0101	.0101
20	.0107	.0091	.0097	.0109
25	.0098	.0099	.0104	.0108
30	.0093	.0098	.0104	.0113
40	.0094	.0103	.0103	.0109
50	.0100	.0101	.0103	.0105
75	.0098	.0095	.0107	.0109
100	.0092	.0108	.0105	.0114

Table 3.4.3b. Simulated null probability of obtaining a value of Wilks' one-outlier test statistic less than the Bonferroni approximation at α=0.025.

Sample size, n	Dimensions, p 2	3	4	5
10	.0256	.0254	.0263	.0245
15	.0250	.0250	.0250	.0253
20	.0258	.0236	.0238	.0266
25	.0260	.0245	.0259	.0264
30	.0249	.0256	.0255	.0272
40	.0240	.0239	.0260	.0255
50	.0250	.0253	.0256	.0256
75	.0251	.0236	.0256	.0264
100	.0245	.0253	.0256	.0274

Table 3.4.3c. Simulated null probability of obtaining a value of Wilks' one-outlier test statistic less than the Bonferroni approximation at α=0.05.

Sample size, n	Dimensions, p 2	3	4	5
10	.0507	.0503	.0505	.0499
15	.0497	.0493	.0483	.0500
20	.0506	.0488	.0492	.0519
25	.0515	.0489	.0502	.0506
30	.0497	.0506	.0507	.0530
40	.0563	.0498	.0489	.0503
50	.0503	.0489	.0499	.0506
75	.0506	.0482	.0512	.0513
100	.0500	.0493	.0496	.0519

Table 3.4.3d. Simulated null probability of obtaining a value of Wilks' one-outlier test statistic less than the Bonferroni approximation at α=0.10.

Sample size, n	Dimensions, p 2	3	4	5
10	.0988	.0994	.0973	.0962
15	.0979	.0993	.0971	.0989
20	.1001	.0970	.0962	.0991
25	.0996	.0970	.0978	.0991
30	.0978	.0985	.0990	.1021
40	.0985	.0975	.0958	.1001
50	.0996	.0962	.0995	.0993
75	.0976	.0937	.0981	.0988
100	.0982	.0962	.0974	.0978

Table 3.4.4a. Simulated null probability of obtaining a value of Wilks' two-outlier test statistic less than the Bonferroni approximation at α=0.01, based on 40,000 simulations.

Sample size, n	Dimensions, p 2	3	4	5
10	.0082	.0085	.0090	.0094
15	.0071	.0072	.0080	.0079
20	.0067	.0070	.0074	.0081
25	.0061	.0066	.0068	.0077
30	.0051	.0063	.0068	.0070
40	.0054	.0059	.0063	.0063
50	.0051	.0058	.0062	.0064
75	.0051	.0048	.0057	.0056
100	.0046	.0052	.0055	.0058

Table 3.4.4b. Simulated null probability of obtaining a value of Wilks' two-outlier test statistic less than the Bonferroni approximation at α=0.025.

Sample size, n	Dimensions, p 2	3	4	5
10	.0187	.0202	.0215	.0223
15	.0162	.0167	.0174	.0183
20	.0160	.0157	.0165	.0174
25	.0145	.0153	.0156	.0169
30	.0127	.0153	.0158	.0159
40	.0116	.0137	.0142	.0142
50	.0119	.0131	.0134	.0136
75	.0108	.0117	.0132	.0136
100	.0102	.0110	.0117	.0129

Table 3.4.4c. Simulated null probability of obtaining a value of Wilks' two-outlier test statistic less than the Bonferroni approximation at α=0.05.

Sample size, n	Dimensions, p 2	3	4	5
10	.0359	.0374	.0401	.0417
15	.0306	.0329	.0335	.0346
20	.0295	.0297	.0299	.0330
25	.0264	.0283	.0290	.0311
30	.0244	.0280	.0284	.0303
40	.0221	.0251	.0265	.0263
50	.0215	.0244	.0242	.0252
75	.0210	.0209	.0240	.0244
100	.0198	.0207	.0223	.0230

Table 3.4.4d. Simulated null probability of obtaining a value of Wilks' two-outlier test statistic less than the Bonferroni approximation at α=0.10.

Sample size, n	Dimensions, p 2	3	4	5
10	.0668	.0698	.0743	.0777
15	.0573	.0607	.0623	.0644
20	.0523	.0539	.0550	.0587
25	.0491	.0514	.0530	.0555
30	.0460	.0496	.0518	.0552
40	.0426	.0448	.0484	.0484
50	.0392	.0432	.0453	.0465
75	.0368	.0377	.0425	.0437
100	.0348	.0367	.0398	.0415

Table 3.4.5 Simulated percentage points for particular cases for 3 and 4 outliers from 2000 simulations; Bonferroni bounds in parentheses.

Case	Significance level 0.01	0.025	0.05	0.10
3 outliers				
p=2, n=10	.00880 (.00733)	.01278 (.01071)	.01762 (.01430)	.02459 (.01914)
p=2, n=20	.13920 (.12224)	.16341 (.13937)	.18502 (.15400)	.20788 (.17028)
p=3, n=10	.00135 (.00120)	.00235 (.00193)	.00346 (.00277)	.00568 (.00400)
p=3, n=20	.08564 (.07537)	.10828 (.08716)	.12104 (.09739)	.13838 (.10893)
p=4, n=10	.00012 (.00010)	.00023 (.00019)	.00037 (.00030)	.00062 (.00049)
p=4, n=20	.05397 (.04614)	.06437 (.02218)	.07292 (.02533)	.08359 (.02897)
4 outliers				
p=2, n=20	.08169 (.06842)	.09991 (.07879)	.11303 (.08773)	.13329 (.09775)
p=4, n=20	.02465 (.01865)	.02875 (.02218)	.03405 (.02533)	.03994 (.02897)

Table 3.4.6 Simulated exceedence probabilities of Bonferroni percentage points derived from certain exact distributions for 3 and 4 outliers (2000 simulations).

Case	Nominal significance level 0.01	0.025	0.05	0.10
3 outliers				
p=2, n=10	.0075	.0165	.0345	.0610
p=2, n=20	.0045	.0105	.0175	.0305
p=3, n=10	.0090	.0165	.0335	.0600
p=3, n=20	.0060	.0105	.0150	.0265
p=4, n=10	.0080	.0215	.0370	.0710
p=4, n=20	.0035	.0105	.0185	.0370
4 outliers				
p=2, n=20	.0035	.0070	.0150	.0235
p=4, n=20	.0030	.0080	.0115	.0260

CHAPTER 4

SEQUENTIALLY APPLIED TESTS

4.1 Introduction

The bulk of the large number of outlier test statistics to be found in the literature is designed for use when the number of possible outliers is specified. For example, a test for two outliers is usually a test of the null hypothesis of no outliers against the alternative hypothesis of two outliers. In itself, it has nothing directly to say about the possibilities of one outlier or three or more outliers being in the sample. If, in fact, the information that there are either two or no outliers is wrong and the number of outliers is not two, then the two-outlier test could be a very poor means of detecting any outliers, as when "masking" occurs. However, it is surely the exception for such firm knowledge to be available. A test for two outliers is usually made because an inspection of the data has suggested either that this is the number present or that this falls in a range of possible numbers of outliers. For example, there could be, graphically, one very clear outlier and two less distinct from the body of the sample; tests for one, two and three outliers might then all be carried out.

Therefore, except for the rare occasions when some a priori specification of the possible number of outliers exists, testing for outliers is part of a multistage process, as Collett and Lewis (1976) point out. Even if a display of the data has very clearly suggested k outliers, then the k-outlier test has been preceded at least by the stage of deciding to use this test and, before that, by choosing that particular display and perhaps by deciding to look for outliers at all. This means that the structure of the test is not as simple as it appears to

be, so that the significance levels may not have their claimed meanings. When there is no clear indication of the possible number of outliers, and instead tests are made over a range of possible numbers, it is even more obvious that each particular k-outlier test is part of a multistage procedure. Since results at each stage must be conditional to some extent on results at earlier stages, this dependence ought to be built into a proper testing procedure. There have been various suggestions for outlier-testing methodologies which do specify the successive application of tests for different numbers of outliers. (Barnett and Lewis, 1984, chapter 5, p.136-143; Hawkins, 1980b, chapter 5). The purpose of this chapter is to develop procedures of this kind for multivariate data. The procedures to be examined will be multivariate applications of tests introduced by Rosner (1975, 1977, 1983) for univariate data.

4.2 Testing Strategies

As this section must include some comments on the terms used for describing different approaches to testing over a range of possible numbers of outliers, it is appropriate to first remark on another point of terminology. It is quite common to speak of such tests as "sequential tests": this is the label used by Hawkins (1980a) and is found in many of the journal articles. Barnett and Lewis (1984, p.136) object to this use of "sequential", because they assert that "sequential test" in statistics means a test in which the sample size is not fixed - each stage involves the accumulation of more data. They prefer to speak of "consecutive" testing for outliers. In fact, since sequential testing in its original sense does not have any application in outlier testing, there doesn't really seem to be any problem caused by talking of sequential tests. In this thesis, the terminology "sequentially applied tests" is used, to retain the

connection with the most widely used terminology.

In sequential tests in their original meaning, the testing process can only be carried out in one "direction", namely, on successively bigger sample sizes. A procedure working on reducing sample sizes would be pointless. In general, however, where a statistical procedure involves multiple testing, there are different paths to follow. In multiple regression, for example, identification of a subset of significant predictors usually proceeds either by choosing the best single predictor, then the next best conditional on the first choice and so on (the forward selection procedure), or by starting with all the potential predictors and eliminating the one making least contribution, then the conditionally next worst and so on (the backward elimination procedure). Besides these two common procedures, there are others, such as procedures aiming to define the best subset of predictors of a given size and procedures which can move both backwards and forwards. There is never any guarantee that these various procedures will reach the same conclusion. The multiple regression example is particularly apt, because Hawkins (1980a) borrows its terminology to describe different strategies in the outlier problem. The basic alternatives in testing sequentially for, say, one up to four outliers, are either to start at four and work downwards - which Hawkins calls backward elimination - or to start at one and work upwards - forward selection. He chooses these terms because he actually has in mind a multiple regression formulation of the outlier problem (his section 7.3). In this, each observation is represented by a dummy variable and an associated regression coefficient, so that "forward selection" in the usual regression sense means successively identifying which coefficients are significantly different from zero so that the corresponding observation is to be regarded as an outlier:

in other words, forward selection identifies increasing numbers of outliers. On the other hand, backward elimination removes points from the set of those being considered as outliers.

Barnett and Lewis (1984) also avoid this terminology, speaking instead of inward and outward consecutive procedures, corresponding to forward selection and backward elimination. These terms refer to direction of movement: inward means starting at the extreme of the sample, the most outlying point, and successively examining points lying closer to the centre of the distribution; outward means moving away from the centre. This thesis adopts the Barnett and Lewis terminology in preference to that of Hawkins in this case. The reason is that the term backward elimination seems too confusing, because 'elimination' seems to suggest the removal of points from the sample, since this is in a sense the intention of the outlier testing procedure, just as well as it suggests removal from the set of possible outliers.

The procedures for sequential application of tests for different numbers of outliers which will be used here are outward testing procedures, starting with a test for a chosen number, k, of outliers and then testing for k-1 and so on if necessary. However, the test statistics employed are derived by inward construction. These procedures will be described in the following sections.

Are there any general grounds for preferring either the inward or the outward procedures over the other? One point is that the inward procedures generally do not avoid the danger of masking, although they can do if a test statistic such as that of Rosner (1975) is used, wherein trimming of the mean and standard deviation ensures that any outliers, up to a specified number, cannot contaminate these estimates. On the other hand, there is no question

of masking in the outward procedures, unless in fact there are actually more than the k outliers at which the testing procedure is started. Hawkins (1980a), as discussed in the next section, finds a problem in the statistical size of the current outward procedures, but eventually comes down in favour of such procedures employed in conjunction with his definitions of critical regions. The calculation of critical regions unfortunately may require simulations; inward procedures, on the other hand, only require standard percentage points.

4.3 Rosner's first procedure for sequentially applied tests

The procedure to be applied here to the case of multivariate data is that introduced by Rosner (1975, 1977) and applied by Prescott (1978, 1979) to use of Grubb's statistic in univariate samples and by Kimber (1982) to testing for outliers in univariate exponential samples. The steps of the method are as follows:

(1) A maximum k is specified as the greatest number of outliers one is prepared to consider.

(2) A single outlier statistic is computed to identify the most extreme member of the sample (without testing), which is then removed from the sample.

(3) The step (2) is repeated on the reduced sample and so on until the k most extreme points have been identified.

(4) The significance test for the kth possible outlier (from step 3) is carried out. If this point is confirmed as an outlier, so too are 1,2,...,(k-1) without further testing. Percentage points are obtained as described below.

(5) Otherwise, test the (k-1)th possible outlier, and so on.

As mentioned in the preceding section, the test statistics used here are constructed 'inwards'. The test itself, however, is an outward test when it comes to actual declaration of outliers.

The percentage points for the tests of significance are determined from the joint distribution of the k outlier test statistics at (2) and (3). This almost inevitably calls for simulations. Rosner proposes the following definition of the size of the test, to which his calculation of percentage points corresponds. Let the test statistics be $D_1,\ldots,D_k$. Critical values $\lambda_1,\ldots,\lambda_k$ are required so that under the null hypothesis of no outliers,

$$\Pr[\bigcup_{j=1}^{k}\{D_j<\lambda_j\}]=\alpha \qquad (4.3.1)$$

for chosen significance level α. (The notation $D_j<\lambda_j$ is used here because Wilks' statistic, which will be used in the multivariate application, looks for values in the lower tail of the distribution; the statistic used by Rosner declares values in the upper tail to be significant, so his notation is $D_j>\lambda_j$.) This can be satisfied in many ways: the condition chosen for a unique solution is to impose equality at each step, so that

$$\Pr\{D_j<\lambda_j\}=\beta, \quad j=1,\ldots,k \qquad (4.3.2)$$

The idea will be that in the simulation study, the marginal and joint distributions of the D_j are recorded in sufficient detail that different values of β in (4.3.2) can be tried and those critical values $\lambda_j(\beta)$ finally selected are those which lead to (4.3.1) being satisfied.

This definition of critical regions for the test is

disliked by Hawkins (1980a, 1980b). Its interpretation is that α is the probability of declaring that there are any outliers (the actual number declared being between one and k) when in fact there are none. Hawkins' objection is that nothing is said about the result when there are outliers. He proposes the alternative definition that, if there are actually $m<k$ outliers, the test should declare more than m outliers with probability α, which is to be the same for all m. This subsumes Rosner's definition (case $m=0$), so the question is whether or not its extra conditions are as desirable as Hawkins claims. Perhaps it is a matter of basic opinions about what is the purpose of outlier testing. In seeking to control the probability of declaring too many outliers, while not apparently being concerned to control the probability of declaring too few, Hawkins seems to be feeling that the danger to avoid is overenthusiastic rejection of points from the sample. This ties in with some introductory remarks in his book (1980a, p8):

"It is this author's experience that statisticians tend to detect outliers that are not present, and to regard the non-significance of outlier test statistics as a reflection on the poor power of the tests rather than an indication that the suspicious-looking observation is statistically quite plausible."

On the other hand, Prescott (1980) responds that further investigation of all the indicated points ought to be the rule anyway - it is not just a question of using this test, throwing away the 'outliers' and then doing on the reduced sample the analysis one had in mind in the first place.

4.4 Rosner's second procedure

The second procedure published by Rosner (1983) recomputes critical values to adopt Hawkins' definition of size. As with his first procedure, the application is to univariate data; the test statistic is max $(x_i-\bar{x})/s$. The size definition (4.3.1) needs to be rewritten as

$$\Pr\{\bigcup_{j=m+1}^{k}(D_j<\lambda_j|H_m)\}=\alpha \qquad (4.4.1)$$

for m=0,1,...,k-1

where H_m is the hypothesis that there are m outliers. (As in the previous section, the event $D_j<\lambda_j$ is sought because Wilks' statistic looks for values in the lower tail.) Equivalently to (4.4.1),

$$\Pr\{\bigcap_{j=m+1}^{k}(D_j\geq\lambda_j|H_m)\}=1-\alpha \qquad (4.4.2)$$

and Rosner conjectures that this probability depends essentially on D_{m+1}; that is, if

$$\Pr(D_{m+1}\geq\lambda_{m+1}|H_m)=1-\alpha' \qquad (4.4.3)$$

then α' is close to α.

The number of outliers declared by the test is the highest value m for which $D_m\leq\lambda_m$ is true, or zero if this is never true for any m up to the maximum considered. For the purpose of discussion, it may be convenient to talk as if the test is executed by starting from m=1 and carrying out every test up to the maximum value of m allowed. In practice, a computer program might start at the maximum value and decrease m until a result which is statistically significant at the chosen level arises, after which point no more tests are made.

In fact, it is obvious from (4.4.2) and (4.4.3) that

$1-\alpha'$ exceeds $1-\alpha$, so that α' is less than α. The error in using (4.4.3) to approximate is (4.4.2) is the probability of the event

$$(D_{m+1} \geq \lambda_{m+1}) \cap \{ \bigcup_{j=m+2}^{k} (D_j < \lambda_j) \} \qquad (4.4.4)$$

under H_m. To take a specific situation, if there are 2 outliers (H_2 holds), then the error of approximation is the probability of the event that the test statistic for examining three outliers (which means examining the most extreme member of the sample of size n-2 after the two most extreme have been eliminated) is not significant at the chosen level, but that one of the statistics for 4, 5 or more outliers, is significant at the same chosen level. In other words, after deleting two apparent outliers, the remaining sample of n-2 points must contain at least two more apparent outliers but none of these must be sufficiently extreme to avoid being masked by the rest in the test for one outlier in n-2 points. It seems quite reasonable to suppose that this probability is small; Rosner (1983) in his Table 1 gives simulation results which show that the approximation is very good for $n \geq 25$ under H_0 in his univariate application. This being so, a very valuable procedure has been obtained, which is very simple because the approximate critical values found from (4.4.3) by putting $\alpha'=\alpha$ are nothing more than the usual critical values for the chosen test statistic in samples of size n-m. In the end, no adaptation of the levels for the sequential application of the test statistic has been made. An equivalent procedure for multivariate data is a very attractive proposition, to avoid the extremely heavy computing demands of multivariate simulation. As will be seen below, however, rather more should be said about the performance of the test than Rosner's simple conclusion suggests.

Some simulation results will now be presented, showing

further details of the performance of Rosner's test in the univariate application given by Rosner himself. These runs were carried out for confirmation after corresponding features had been observed in the simulations for the multivariate case, described below in § 4.6. The investigation looks at the performance of the test in the presence of one or two outliers. This is the situation examined by Rosner in his Table 2, and the details of the simulation are the same, namely, 2000 runs are made for each case and the sample size is 25.

Table 4.4.1 shows the proportion of samples in which outliers were declared, at nominal 1% and 5% levels (Rosner presents only the latter), under two versions of the test. In one, the maximum number of outliers allowed was two. Rosner looked at this for a direct comparison with his earlier procedure. In the other, up to 10 outliers were allowed. If there are two contaminating points, only this second version provides any information about the probability of declaring "too many" outliers, but Rosner did not examine this at all.

Comparison between the results here for the 5% level test with up to 2 outliers allowed, and Rosner's results in his Table 2, shows good agreement. One particularly interesting result is the probability of 0.0190 of declaring two outliers when there is actually just one, with a slippage of 2. Rosner gives 0.01 at this point. The results for up to 10 outliers, and at the 1% level, agree that the probability here is substantially below the nominal level. Since Rosner presents results for the two-outlier case only for testing for up to two outliers, he gives no corresponding information about the probability of declaring more than two outliers in this case. This information is now supplied by the extreme right-hand column of Table 4.4.1. It can be seen that the probability is again well below the nominal level, except

when both slippages are very large (-4 and 6). It therefore seems that Rosner's test can sometimes be very conservative.

This result, when first seen in the multivariate case, was rather a surprise because the evidence of Rosner's Table 1 (under H_0) suggested a liberal test, not a conservative one. The use of (4.4.3) to approximate (4.4.2) also makes it seem that the test at each stage separately should be liberal: under H_m, there is an $\alpha\%$ probability of declaring one further outlier at this stage plus the probability of not declaring one at this stage but declaring more than one at a later stage. The true error probability therefore exceeds $\alpha\%$ (and the use of the conservative Bonferroni bounds has relatively little impact, because they are known to be quite accurate in the single-outlier tests which are carried out at each stage.) However, this reasoning applies strictly only to the very first step, under H_0. Under H_1, for example, the distribution theory for the test for an outlier in the reduced sample of size n-1 depends on the contaminating observation having been identified correctly in the first test and removed from the sample. In fact, except when the slippage is so large that the contaminant is virtually always found, removing the most extreme point slightly truncates the distribution. Thus the distributional assumptions do not apply for H_1, H_2,... as they do for H_0.

Table 4.4.1. Proportion of times that given numbers of outliers were declared by Rosner's procedure in 2000 simulations: n=25.

(a) Nominal 1% level

Outliers	Slippage(s)	Up to 2 allowed			Up to 10 allowed				
		0	1	2	0	1	2	≥3	>true
0	-	.9830	.0160	.0010	.9905	.0085	.0010	0	.0095
1	2	.9555	.0420	.0025	.9700	.0280	.0010	.0010	.0020
1	6	.0515	.9390	.0095	.0590	.9260	.0130	.0020	.0150
2	2, 2	.9620	.0320	.0060	.9625	.0295	.0065	.0015	.0015
2	-2, 6	.0830	.8830	.0340	.0815	.8765	.0390	.0030	.0030
2	-4, 6	.1340	.4325	.4335	.1240	.4250	.4445	.0065	.0065

(b) Nominal 5% level

0	-	.9390	.0510	.0100	.9415	.0425	.0070	.0090	.0585
1	2	.8665	.1145	.0190	.8770	.1050	.0100	.0080	.0180
1	6	.0110	.9335	.0555	.0130	.9205	.0495	.0170	.0665
2	2, 2	.8500	.1190	.0310	.8485	.1055	.0300	.0160	.0160
2	-2, 6	.0130	.8815	.1055	.0165	.8520	.1010	.0305	.0305
2	-4, 6	.0245	.3160	.6595	.0195	.3045	.6175	.0585	.0585

A further possible factor making the performance of the test rather unpredictable is that the sequential removal of points does not necessarily lead to the "best" set of points being removed (that is, the ones that would have been removed if all were tested simultaneously in a set of appropriate size). This again should not arise if the slippages are very large. Both factors may be more important when the sample size is small.

Another interesting feature of the results in Table 4.4.1 lies in the comparison of the results for two outliers with slippages of -2 and 6 against those for -4 and 6. It might be expected that an outlier test should be more likely to declare outliers in the second case than

the first, and in fact it can be seen to be far more likely to declare two outliers in the second case. But it turns out that it is also more likely not to declare any at all. The reason for this behaviour is that the tests in the two different cases are not comparable, since they have different sizes as seen in the final column of the table.

4.5 Sequential application of Wilks' test statistic

In using Wilks' test statistic with the methods proposed by Rosner, the first step is the construction of the statistics $D_1,\ldots,D_k$, from the most extreme to the kth most extreme points in the sample. The most extreme point is the point j such that the ratio

$$\frac{|A_j|}{|A|} \qquad (4.5.1)$$

is the minimum over such ratios for all sample points, where A and A_j are the sample sums of squares and products (SSP) matrices respectively before and after deletion of point j for the sample. The value of this ratio is the statistic D_1. The corresponding point j is now removed from the sample and the most extreme of the remaining n-1 points identified. This is point h, such that

$$\frac{|A_{jh}|}{|A_j|} \qquad (4.5.2)$$

is a minimum over all n-1 choices of points, where A_{jh} denotes the SSP matrix of the n-2 points remaining from the sample after deletion of both j and h. This ratio is D_2. Similar minimizations and deletions lead to $D_3,\ldots,D_k$.

In fact, calculation is a little simpler if an

alternative form is used for the Wilks statistic. As shown in § 2.2, (4.5.1) can also be written as

$$1-(n/(n-1))(x_j-\bar{x})'A^{-1}(x_j-\bar{x})$$

and similarly (4.5.2) as

$$1-((n-1)/(n-2))(x_h-\bar{x}_j)'A_j^{-1}(x_h-\bar{x}_j)$$

where $\bar{x}_j$ denotes the mean of the sample of n-1 points remaining after deletion of point j. Similar expressions follow for the rest of the D-statistics. The advantage of this form is that the usual updating formula gives A_j^{-1} in terms of A^{-1} without the need for an actual matrix inversion. Specifically,

$$A_j^{-1} = A^{-1} + \frac{nA^{-1}(x_j-\bar{x})(x_j-\bar{x})'A^{-1}}{(n-1)\{1-(n/(n-1))(x_j-\bar{x})'A^{-1}(x_j-\bar{x})\}}$$

(e.g. Morrison, 1976, p.69), with similar expressions for A_{jh}^{-1} in terms of A_j^{-1} and so on.

Computation of critical values for tests at chosen levels of significance using $D_k,\ldots,D_1$ follows the methodology described in the previous sections. In the more complicated case of the first method (§ 4.3), results were derived by simulation for tests at the 10, 5, 2.5 and 1% levels for maximum number of outliers k=2 and 3. Within each chosen combination of sample size n and dimensionality p, results for both values of k and all values of significance level were obtained from the same simulated data. At each of these combinations, 40000 samples of the required size n were generated for calculation of the distribution of the $\{D_j\}$. These samples were obtained as 5 lots of 8000, each lot starting with a different seed for the IMSL pseudo-random generator GGNSM for multivariate normal vectors.

The simulated critical values are displayed in Tables 4.5.1 for k=2 and 4.5.2 for k=3.

TABLE 4.5.1 Critical values for testing for up to 2 outliers using the sequentially applied version of Wilk's test.

Dimensions:		2		3		4		5	
α	n	λ_1	λ_2	λ_1	λ_2	λ_1	λ_2	λ_1	λ_2
0.01	15	.2606	.3131	.2007	.2361	.1467	.1707	.1020	.1172
	20	.3775	.4509	.3103	.3765	.2594	.3174	.2126	.2569
	25	.4588	.5457	.4026	.4802	.3497	.4208	.3044	.3726
	30	.5246	.6093	.4724	.5542	.4236	.5001	.3781	.4505
	50	.6726	.7482	.6337	.7081	.5983	.6737	.5694	.6408
	100	.8153	.8622	.7902	.8388	.7707	.8201	.7510	.8023
.025	15	.3065	.3579	.2383	.2733	.1754	.2032	.1290	.1470
	20	.4230	.4915	.3514	.4163	.2938	.3508	.2433	.2870
	25	.4983	.5792	.4409	.5132	.3878	.4549	.3371	.4007
	30	.5623	.6370	.5052	.5806	.4567	.5295	.4115	.4789
	50	.7031	.7669	.6611	.7268	.6275	.6928	.5958	.6601
	100	.8323	.8713	.8085	.8500	.7875	.8304	.7679	.8119
.05	15	.3453	.3971	.2703	.3090	.2023	.2320	.1525	.1715
	20	.4588	.5236	.3858	.4480	.3248	.3780	.2721	.3170
	25	.5342	.6071	.4719	.5391	.4168	.4796	.3667	.4264
	30	.5936	.6618	.5349	.6036	.4844	.5508	.4393	.5034
	50	.7258	.7816	.6842	.7421	.6501	.7081	.6175	.6743
	100	.8442	.8786	.8217	.8572	.8014	.8388	.7819	.8203
.10	15	.3889	.4385	.3082	.3480	.2373	.2667	.1816	.2002
	20	.4996	.5592	.4248	.4819	.3605	.4110	.3055	.3482
	25	.5730	.6354	.5084	.5672	.4502	.5076	.4001	.4530
	30	.6267	.6869	.5680	.6282	.5172	.5764	.4707	.5279
	50	.7487	.7966	.7073	.7578	.6741	.7245	.6408	.6908
	100	.8574	.8861	.8352	.8653	.8148	.8464	.7964	.8290

TABLE 4.5.2 Critical values for testing for up to 3 outliers using sequentially applied version of Wilk's test.

Dimensions:		2			3			4			5		
α	n	λ_1	λ_2	λ_3	λ_1	λ_2	λ_3	λ_1	λ_2	λ_3	λ_1	λ_2	λ_3
0.01	15	.2458	.2956	.2895	.1854	.2212	.2083	.1359	.1560	.1364	.0899	.1063	.0903
	20	.3611	.4344	.4566	.2960	.3620	.3724	.2451	.3003	.2948	.2026	.2451	.2434
	25	.4419	.5322	.5589	.3877	.4637	.4873	.3366	.4092	.4220	.2900	.3556	.3643
	30	.5080	.5958	.6284	.4578	.5432	.5643	.4112	.4893	.5137	.3641	.4394	.4610
	50	.6632	.7417	.7707	.6236	.6997	.7314	.5896	.6657	.6918	.5559	.6322	.6608
	100	.8095	.8587	.8769	.7839	.8358	.8544	.7648	.8159	.8372	.7448	.7986	.8175
0.025	15	.2856	.3348	.3315	.2207	.2563	.2438	.1623	.1875	.1664	.1169	.1338	.1145
	20	.4040	.4748	.4935	.3333	.3987	.4057	.2800	.3360	.3344	.2305	.2734	.2741
	25	.4836	.5663	.5922	.4239	.4985	.5194	.3725	.4412	.4561	.3207	.3875	.3984
	30	.5470	.6261	.6547	.4897	.5687	.5921	.4431	.5178	.5387	.3968	.4675	.4910
	50	.6923	.7598	.7850	.6529	.7191	.7448	.6162	.6849	.7096	.5853	.6520	.6779
	100	.8257	.8680	.8839	.8020	.8460	.8623	.7816	.8265	.8446	.7618	.8083	.8265
0.05	15	.3231	.3751	.3707	.2519	.2892	.2779	.1881	.2137	.1956	.1386	.1568	.1380
	20	.4404	.5059	.5246	.3668	.4309	.4361	.3057	.3635	.3640	.2558	.2995	.3013
	25	.5147	.5877	.6170	.4553	.5253	.5458	.4011	.4660	.4802	.3489	.4120	.4238
	30	.5780	.6496	.6749	.5192	.5917	.6155	.4694	.5396	.5590	.4243	.4895	.5121
	50	.7143	.7739	.7966	.6725	.7512	.7708	.6387	.7008	.7230	.6064	.6671	.6909
	100	.8389	.8751	.8896	.8157	.8538	.8687	.7950	.8351	.8351	.7754	.8165	.8327
0.10	15	.3661	.4155	.4155	.2881	.3259	.3163	.2171	.2480	.2326	.1652	.1845	.1652
	20	.4778	.5405	.5575	.4040	.4639	.4701	.3407	.3936	.3983	.2873	.3312	.3322
	25	.5538	.6211	.6425	.4899	.5524	.5717	.4327	.4924	.5071	.3835	.4390	.4515
	30	.6098	.6745	.6972	.5519	.6162	.6371	.5000	.5641	.5831	.4551	.5162	.5345
	50	.7382	.7896	.8092	.6984	.7512	.7708	.6632	.7170	.7373	.6301	.6836	.7045
	100	.8520	.8830	.8956	.8297	.8619	.8751	.8090	.8429	.8566	.7902	.8254	.8395

The performance of the test procedure using these critical values was investigated by simulation (2000 samples at each combination of n and p) in the presence of different numbers of outliers with different amounts of slippage in the mean. The results are discussed in the following section. Some comments on the choice of slippage in this and other simulations in this thesis will be found in Appendix II.

For the second of Rosner's methods, simulation was used only to examine the performance of the method. Critical values were obtained from (4.4.3). Specifically, since D_{h+1} is the value of Wilks' statistic in the sample of n-h points remaining after h extreme points have been deleted from the original sample of n points, the rest at this step consists of comparing D_{h+1} to the $\alpha/(n-h)\%$ point of the Beta distribution with parameters (n-h-p-1)/2 and p/2. These are the standard Bonferroni approximations, just as Rosner (1983) employs in his univariate application.

The simulations presented below involved generating 2000 samples for each combination of n, p and number and type of outliers. Tests were made for all numbers of outliers up to the minimum of n/2, 10 and n-p-1; the first two of these conditions were used by Rosner and the third is a detail which ensures that the matrices being examined remain non-singular.

4.6 Performance of the two procedures

The results of simulation studies of the two methods are presented here in tables showing the proportion of simulated samples in which outliers were declared, under various conditions. Such results would normally be called size and power; however, this terminology is liable to become confused because a whole sequence of hypotheses $(H_k: k=0, 1, 2, \ldots)$ is under consideration. The first

table (Table 4.6.1) studies the first procedure, in the versions testing for up to 2 and up to 3 outliers, in the presence of either one (H_1), two (H_2) or three (H_3) contaminating observations. No study under H_0 is needed, because the error level is fixed by construction. The second procedure is examined under H_0 in Table 4.6.2 and under H_1, H_2 and H_3 in Table 4.6.3. Some results for those combinations of n, p and number and type of contaminants which were examined under both methods are gathered together in Table 4.6.5: the intervening table, 4.6.4, augments the results of 4.6.3.

Table 4.6.1 shows the proportion of simulated samples in which outliers are declared, in the presence of either one, two or three contaminants.

Table 4.6.1a Performance of first sequentially applied procedure in presence of one or two outliers, at 1% level.

p	n	Out-liers	Squared slippage distance	Type[1]	Outliers tested ≤2: Outliers declared 0	1	2	Outliers tested ≤3: Outliers declared 0	1	2	3
2	15	1	15	+	.8710	.1110	.0180	.8930	.0905	.0120	.0045
			30	+	.5580	.4165	.0255	.6100	.3700	.0160	.0040
		2	20	++	.7690	.0155	.2155	.7755	.0100	.1815	.0330
			20	+-	.7555	.0210	.2235	.7750	.0160	.1850	.0240
			20	+±	.6150	.2155	.1695	.6655	.1985	.1240	.0120
2	25	1	15	+	.7925	.1780	.0295	.8170	.1515	.0225	.0090
			30	+	.3370	.6225	.0405	.3835	.5750	.0290	.0125
		2	20	++	.5235	.0465	.4300	.5380	.0355	.3760	.0505
			20	+-	.4960	.0660	.4380	.5285	.0530	.3705	.0480
			20	+±	.3690	.2315	.3995	.4155	.2065	.3335	.0445
4	15	1	15	+	.9525	.0365	.0110	.9615	.0275	.0060	.0050
			30	+	.8165	.1675	.0160	.8480	.1390	.0090	.0040
		2	20	++	.9405	.0045	.0550	.9390	.0030	.0435	.0145
			20	+-	.9280	.0150	.0570	.9380	.0095	.0385	.0140
			20	+±	.8255	.1205	.0540	.8585	.0955	.0400	.0060
4	25	1	15	+	.9045	.0750	.0205	.9190	.0575	.0160	.0075
			30	+	.5740	.3875	.0385	.6155	.3465	.0285	.0095
		2	20	++	.7790	.0185	.2025	.7660	.0150	.1700	.0490
			20	+-	.7660	.0370	.1970	.7760	.0290	.1605	.0345
			20	+±	.6200	.1875	.1925	.6590	.1595	.1550	.0265

[1] Key : + equal slippage added to each dimension
- equal slippage subtracted to each dimension
± equal slippage added to 1st dimension (to 1st and 3rd for p=4) and subtracted from the 2nd dimension (from 2nd and 4th for p=4).

(Details of slippage calculations in Appendix II)

Table 4.6.1b Performance of first sequentially applied procedure in presence of one or two outliers, at 5% level.

p	n	Out-liers	Squared slippage distance	Type	Outliers tested: ≤2 Outliers declared: 0	1	2	Outliers tested: ≤3 Outliers declared: 0	1	2	3
2	15	1	15	+	.6545	.2695	.0760	.7110	.2125	.0535	.0230
			30	+	.2800	.6270	.0930	.3350	.5715	.0675	.0260
		2	20	++	.5520	.0405	.4075	.5430	.0290	.3360	.0920
			20	+-	.4870	.0655	.4475	.5150	.0550	.3560	.0740
			20	+±	.2775	.3035	.4190	.3510	.2745	.3190	.0555
2	25	1	15	+	.5735	.3275	.0990	.6125	.2860	.0580	.0435
			30	+	.1380	.7335	.1285	.1760	.7005	.0770	.0465
		2	20	++	.2830	.0885	.6285	.3050	.0715	.4735	.1500
			20	+-	.2410	.1135	.6455	.2655	.1045	.4890	.1410
			20	+±	.1495	.2295	.6210	.1855	.2220	.4505	.1420
4	15	1	15	+	.8320	.1175	.0505	.8445	.0945	.0320	.0290
			30	+	.5695	.3490	.0815	.6275	.3100	.0440	.0185
		2	20	++	.8165	.0250	.1585	.8165	.0170	.1140	.0525
			20	+-	.7905	.0525	.1570	.8025	.0360	.1115	.0500
			20	+±	.5830	.2430	.1740	.6260	.2215	.1135	.0390
4	25	1	15	+	.7245	.2065	.0690	.7590	.1615	.0445	.0350
			30	+	.3290	.5495	.1215	.3635	.5080	.0795	.0490
		2	20	++	.5770	.0660	.3570	.5540	.0510	.2705	.1245
			20	+-	.5110	.0965	.3925	.5360	.0705	.2915	.1020
			20	+±	.3425	.2500	.4075	.3920	.2235	.2790	.1055

Table 4.6.1c Performance of first sequentially applied procedure in presence of three outliers, all with squared generalized distance 20; 1% significance level.

			Outliers tested							
			≤2			≤3			
			Outliers declared			Outliers declared			
p	n	Type	0	1	2	0	1	2	3
2	15	+++	.9835	.0030	.0135	.8670	.0025	.0075	.1230
		++-	.9635	.0060	.0305	.8365	.0040	.0180	.1415
		++±	.7235	.1760	.1005	.6365	.1105	.0755	.1775
		+-±	.6830	.1895	.1275	.6275	.1295	.0930	.1500
2	25	+++	.8360	.0105	.1535	.5630	.0055	.0555	.3760
		++-	.7820	.0160	.2020	.5320	.0065	.0815	.3800
		++±	.4435	.2370	.3195	.3400	.1180	.1515	.3905
		+-±	.3900	.1545	.4655	.3270	.0890	.2050	.3790
4	15	+++	.9885	.0060	.0055	.9770	.0060	.0030	.0140
		++-	.9815	.0085	.0100	.9670	.0070	.0060	.0200
		++±	.8915	.0760	.0325	.8845	.0610	.0225	.0320
		+-±	.8855	.0740	.0405	.8920	.0550	.0245	.0285
4	25	+++	.9435	.0110	.0455	.8520	.0075	.0240	.1165
		++-	.9130	.0155	.0715	.8100	.0100	.0455	.1345
		++±	.6915	.1565	.1520	.6355	.1090	.1025	.1530
		+-±	.6570	.1395	.2035	.6135	.0975	.1345	.1545

Table 4.6.1d Performance of first sequentially applied procedure in presence of three outliers, all with squared generalized distance 20; 5% significance level.

			Outliers tested							
			≤2			≤3			
			Outliers declared			Outliers declared			
p	n	Type	0	1	2	0	1	2	3
2	15	+++	.9155	.0155	.0690	.6990	.0065	.0320	.2625
		++-	.8420	.0315	.1265	.6025	.0175	.0625	.3175
		++±	.4280	.3400	.2320	.3360	.1815	.1365	.3460
		+-±	.3530	.2945	.3525	.3165	.1695	.1750	.3390
2	25	+++	.5790	.0370	.3840	.3385	.0175	.0925	.5515
		++-	.4570	.0545	.4885	.2845	.0205	.1125	.5825
		++±	.1580	.2355	.6065	.1225	.1090	.1755	.5930
		+-±	.1280	.1250	.7470	.1140	.0710	.2110	.6040
4	15	+++	.9485	.0230	.0285	.9110	.0175	.0150	.0565
		++-	.9245	.0330	.0425	.8870	.0230	.0275	.0625
		++±	.7070	.1785	.1145	.7020	.1315	.0680	.0985
		+-±	.6615	.1920	.1465	.6805	.1405	.0885	.0905
4	25	+++	.8165	.0410	.1425	.6670	.0215	.0620	.2495
		++-	.7575	.0565	.1860	.6065	.0300	.0815	.2820
		++±	.4255	.2395	.3350	.3625	.1510	.1800	.3065
		+-±	.3610	.2090	.4300	.3365	.1300	.2080	.3255

The two versions of the test, for a maximum of either two (T_2) or three (T_3) outliers, are considered, so that the possible number of outliers which can be declared are zero, one and two (in test T_2), or zero, one, two and three (in T_3). A variety of distances and directions are covered for the slippage vectors. The table first shows results at the 1% and 5% levels for one or two contaminants. The following observations can be made:

a) in most cases, the test T_2 for up to 2 outliers is more likely to declare any outliers than the test T_3 for up to 3, if in fact there are one or two contaminants;

b) the opposite is true if there are three contaminants;

c) if there is one contaminant, the result in a) is due to a greater probability of declaring exactly one outlier with test T_2 than with test T_3 - the probability of declaring more than one outlier is about the same under both tests;

d) the probability of declaring more than the true number of contaminants can go up to about 5% for the test at the 1% level (under H_0) and 15% for the test at the 5% level: these figures depend on the test (T_2 or T_3) and the nature of the outliers, but in almost all cases exceeds the size under H_0.

The result a) is to be expected because some of the, say, 5% error probability which is all "used up" in testing for one or two outliers in the test T_2 must be allotted to testing for three outliers in the test T_3. This means reducing the probability of declaring outliers in the tests for one and two outliers, and this is not made up for by the small probability of declaring three outliers (since there are actually only one or two contaminating points). On the other hand, if there are in fact three outliers, only the test T_3 can declare this,

and - with the test T_2 suffering from masking - the result b) follows. Result c) is also entirely as expected; changes in the critical value have a bigger absolute effect at the level of one outlier since declaring two or three outliers is a relatively rarer event.

Result d) illustrates the point of Hawkins' criticism of the construction of critical values at each level of the test, that error levels under H_1, H_2,... are not controlled. The fact that the error level increases over that applying to H_0 bears out the remark in Rosner (1983) and the example in Hawkins (1980a).

For the second sequentially applied procedure, Table 4.6.2 shows the proportion of samples in which outliers were declared, when in fact there were no contaminants.

Table 4.6.2a Performance of second sequentially applied procedure under H_0 (no outliers), at nominal 1% level. (Blanks denote values that are the same as the preceding ones in the same row; dashes denote tests which were not carried out.)

		Outliers declared (cumulative proportions)									
p	n	1	2	3	4	5	6	7	8	9	10
2	10	.0090	.0130	.0150	.0205	-	-	-	-	-	-
	15	.0075	.0095	.0095	.0105	.0115	.0115	.0130	-	-	-
	20	.0130	.0140	.0145	.0150	.0150	.0150	.0150	.0150	.0155	
	25	.0150	.0160	.0175	.0175	.0175	.0175	.0175	.0175	.0180	
	30	.0095	.0105								
	50	.0130									
	100	.0110									
3	10	.0100	.0145	.0185	.0230	-	-	-	-	-	-
	15	.0105	.0105	.0120	.0135	.0140	.0165	.0190	-	-	-
	20	.0120	.0125	.0125	.0130	.0135	.0140	.0145	.0150	.0150	.0165
	25	.0100	.0115								
	30	.0090									
	50	.0110									
	100	.0120									
4	10	.0100	.0135	.0200	.0245	-	-	-	-	-	-
	15	.0075	.0090	.0095	.0105	.0125	.0130	.0150	-	-	-
	20	.0090	.0105	.0105	.0105	.0105	.0110	.0110	.0115	.0125	.0150
	25	.0120	.0130	.0130	.0130	.0130	.0135				
	30	.0115	.0120								
	50	.0090	.0145								
	100	.0140									
5	10	.0075	.0120	.0160	.0250	-	-	-	-	-	-
	15	.0085	.0115	.0125	.0145	.0180	.0245	.0290	-	-	-
	20	.0100	.0105	.0105	.0105	.0115	.0115	.0125	.0140	.0145	.0170
	25	.0110	.0120	.0120	.0125	.0125	.0125	.0125	.0130	.0135	
	30	.0120									
	50	.0095									
	100	.0080									

Table 4.6.2b Performance of second sequentially applied procedure under H_0 (no outliers), at nominal 5% level. (Blanks denote values that are the same as the preceding ones in the same row; dashes denote tests which were not carried out.)

		Outliers declared (cumulative proportions)									
p	n	1	2	3	4	5	6	7	8	9	10
2	10	.0460	.0695	.0885	.1090	-	-	-	-	-	-
	15	.0470	.0605	.0625	.0705	.0760	.0865	.0950	-	-	-
	20	.0445	.0510	.0535	.0580	.0595	.0625	.0640	.0660	.0690	.0745
	25	.0475	.0555	.0585	.0600	.0600	.0600	.0610	.0615	.0625	.0630
	30	.0390	.0460	.0470	.0470	.0475	.0480	.0485			
	50	.0495	.0515	.0520							
	100	.0530	.0560								
3	10	.0480	.0655	.0865	.1110	-	-	-	-	-	-
	15	.0450	.0540	.0650	.0735	.0800	.0930	.1025	-	-	-
	20	.0415	.0495	.0525	.0545	.0585	.0620	.0640	.0670	.0690	.0765
	25	.0450	.0535	.0565	.0570	.0570	.0580	.0585	.0590	.0590	.0610
	30	.0500	.0525	.0535	.0535	.0540	.0545	.0555			
	50	.0540	.0575	.0580							
	100	.0390	.0400								
4	10	.0440	.0665	.0900	.1150	-	-	-	-	-	-
	15	.0375	.0490	.0535	.0645	.0720	.0850	.0980	-	-	-
	20	.0490	.0560	.0590	.0605	.0645	.0700	.0735	.0790	.0860	.0945
	25	.0470	.0545	.0560	.0600	.0605	.0610	.0615	.0625	.0625	.0635
	30	.0470	.0520	.0545	.0550	.0550	.0550	.0550	.0560	.0560	.0565
	50	.0485	.0535	.0540							
	100	.0530	.0535								
5	10	.0380	.0590	.0845	.1230	-	-	-	-	-	-
	15	.0415	.0545	.0620	.0720	.0820	.1065	.1240	-	-	-
	20	.0460	.0520	.0555	.0590	.0635	.0675	.0695	.0765	.0820	.0915
	25	.0540	.0605	.0640	.0670	.0675	.0695	.0710	.0730	.0740	.0745
	30	.0455	.0530	.0530	.0530	.0535	.0540	.0545	.0550	.0560	.0570
	50	.0450	.0475	.0475	.0480						
	100	.0455									

(Note, as a check, that the first column - one outlier declared - simply gives the size of Wilks' test using Bonferroni bounds at the nominal level of 1% and 5%, as investigated in more detail in Chapter 3.) This table

suggests that one needs a sample size of about 25 to 30 before the true significance levels are very close to the nominal 1% and 5% levels. This result appears not to depend on the dimensionality, p. The finding agrees very well with Rosner's (1983) recommendation that the approximation (4.4.3) is acceptable for $n \geq 25$ in his univariate application of the methodology.

It might also be expected that Table 4.6.3 would similarly confirm, at least for sufficiently large n, the adequacy of approximation (4.4.3) in the presence of outliers.

Table 4.6.3a Performance of second sequentially applied procedure in presence of one or two outliers, at nominal 1% level.

p	n	Out-liers	Squared slippage distance	Type[1]	Outliers declared 0	1	2	3	≥4	More than correct no
2	15	1	15	+	.8125	.1725	.0075	.0010	.0065	.0150
			30	+	.4600	.5240	.0085	.0010	.0065	.0160
		2	20	++	.8195	.0245	.1285	.0185	.0090	.0275
			20	+-	.7995	.0530	.1270	.0085	.0080	.0165
			20	+±	.5070	.3930	.0870	.0550	.0075	.0130
2	25	1	15	+	.2795	.7055	.0125	.0000	.0025	.0150
			30	+	.0115	.9770	.0100	.0005	.0010	.0115
		2	20	++	.1575	.4460	.3865	.0050	.0010	.0060
			20	+-	.1195	.4635	.4095	.0070	.0005	.0075
			20	+±	.1320	.4935	.3680	.0055	.0010	.0065
2	50	1	15	+	.1865	.8050	.0085	.0000	.0000	.0085
			30	+	.0015	.9885	.0100	.0000	.0000	.0100
		2	20	++	.0435	.4705	.4810	.0050	.0000	.0050
			20	+-	.0420	.4580	.4945	.0055	.0000	.0055
			20	+±	.0430	.4655	.4875	.0040	.0000	.0040
4	15	1	15	+	.9230	.0670	.0005	.0015	.0080	.0100
			30	+	.7490	.2285	.0085	.0030	.0110	.0225
		2	20	++	.9290	.0160	.0330	.0075	.0145	.0220
			20	+-	.9240	.0245	.0320	.0050	.0145	.0195
			20	+±	.8665	.1950	.0225	.0050	.0110	.0160
4	25	1	15	+	.8790	.1130	.0050	.0000	.0030	.0080
			30	+	.5155	.4775	.0050	.0015	.0005	.0070
		2	20	++	.8595	.0480	.0820	.0085	.0030	.0105
			20	+-	.8450	.0695	.0815	.0035	.0005	.0040
			20	+±	.6135	.3330	.0520	.0010	.0005	.0015
4	50	1	15	+	.8090	.1895	.0015	.0000	.0000	.0015
			30	+	.3490	.6460	.0040	.0010	.0000	.0050
		2	20	++	.6895	.1530	.1545	.0030	.0000	.0030
			20	+-	.6580	.2120	.1250	.0050	.0000	.0050
			20	+±	.4485	.4305	.1190	.0020	.0000	.0020

Table 4.6.3b Performance of second sequentially applied procedure in presence of one or two outliers, at nominal 5% level.

p	n	Out-liers	Squared slippage distance	Type	Outliers declared 0	1	2	3	≥4	More than correct no
2	15	1	15	+	.5590	.3580	.0320	.0145	.0365	.0830
			30	+	.1885	.7200	.0415	.0130	.0370	.0915
		2	20	++	.5345	.0690	.2915	.0480	.0570	.1050
			20	+-	.4730	.1240	.3110	.0370	.0550	.0920
			20	+±	.1825	.4715	.2650	.0300	.0510	.0810
2	25	1	15	+	.1105	.8315	.0440	.0065	.0075	.0580
			30	+	.0000	.9295	.0550	.0050	.0105	.0705
		2	20	++	.0395	.3135	.5985	.0355	.0130	.0485
			20	+-	.0200	.3110	.6145	.0385	.0160	.0545
			20	+±	.0205	.3220	.6030	.0380	.0165	.0545
2	50	1	15	+	.0710	.8795	.0465	.0030	.0000	.0495
			30	+	.0000	.9495	.0475	.0030	.0000	.0505
		2	20	++	.0090	.2855	.6650	.0360	.0045	.0405
			20	+-	.0090	.2885	.6610	.0380	.0035	.0415
			20	+±	.0095	.2895	.6580	.0395	.0035	.0430
4	15	1	15	+	.7385	.1870	.0195	.0115	.0435	.0745
			30	+	.4425	.4585	.0305	.0140	.0545	.0990
		2	20	++	.7515	.0580	.0950	.0315	.0640	.0955
			20	+-	.7230	.0835	.1020	.0315	.0600	.0915
			20	+±	.4625	.3615	.1015	.0180	.0565	.0745
4	25	1	15	+	.7030	.2520	.0270	.0045	.0135	.0450
			30	+	.2645	.6790	.0365	.0075	.0125	.0565
		2	20	++	.6015	.1180	.2195	.0380	.0235	.0615
			20	+-	.5805	.1895	.1865	.0270	.0165	.0435
			20	+±	.3190	.4690	.1765	.0185	.0170	.0355
4	50	1	15	+	.6335	.3440	.0205	.0015	.0005	.0225
			30	+	.1805	.7825	.0335	.0020	.0015	.0370
		2	20	++	.4385	.2455	.2940	.0195	.0025	.0220
			20	+-	.3860	.3185	.2760	.0175	.0020	.0195
			20	+±	.2180	.4885	.2785	.0125	.0025	.0150

Table 4.6.3c Performance of second sequentially applied procedure in presence of three outliers (all with squared generalized distance = 20), at nominal 1% level.

p	n	Outlier types	Outliers declared 0	1	2	3	4	≥5	More than correct number
2	15	+++	.8885	.0060	.0020	.0670	.0205	.0160	.0365
		++-	.8990	.0145	.0065	.0630	.0120	.0050	.0170
		++±	.6130	.2480	.0415	.0825	.0120	.0030	.0150
		+-±	.6105	.2545	.0500	.0765	.0050	.0035	.0085
2	25	+++	.2065	.0530	.3200	.4085	.0100	.0020	.0120
		++-	.1990	.0535	.3430	.3935	.0085	.0025	.0110
		++±	.1940	.0700	.3460	.3775	.0095	.0030	.0125
		+-±	.1815	.0775	.3505	.3805	.0080	.0020	.0100
2	50	+++	.0340	.0795	.4295	.4515	.0045	.0010	.0055
		++-	.0255	.0615	.4455	.4645	.0025	.0005	.0030
		++±	.0280	.0525	.4255	.4880	.0050	.0010	.0060
		+-±	.0245	.0680	.4100	.4905	.0065	.0005	.0070
4	15	+++	.9615	.0085	.0020	.0125	.0035	.0120	.0155
		++-	.9600	.0110	.0045	.0080	.0035	.0130	.0165
		++±	.8430	.1100	.0160	.0120	.0060	.0130	.0190
		+-±	.8430	.1040	.0250	.0140	.0095	.0045	.0140
4	25	+++	.9225	.0160	.0090	.0295	.0130	.0100	.0230
		++-	.9125	.0235	.0135	.0315	.0130	.0060	.0190
		++±	.6715	.2295	.0530	.0310	.0105	.0045	.0150
		+-±	.6670	.2345	.0590	.0330	.0055	.0010	.0065
4	50	+++	.8025	.0925	.0430	.0555	.0065	.0000	.0065
		++-	.7725	.1000	.0630	.0605	.0040	.0000	.0040
		++±	.4665	.3420	.1340	.0555	.0020	.0000	.0020
		+-±	.4405	.3620	.1450	.0510	.0015	.0000	.0015

Table 4.6.3d Performance of second sequentially applied procedure in presence of three outliers (all with squared generalized distance 20), at nominal 5% level.

p	n	Outlier types	Outliers declared 0	1	2	3	4	≥5	More than correct number
2	15	+++	.6580	.0295	.0235	.1630	.0600	.0660	.1260
		++-	.6085	.0475	.0410	.2070	.0470	.0490	.0960
		++±	.2415	.3270	.1050	.2390	.0455	.0420	.0875
		+-±	.2275	.3330	.1670	.2020	.0305	.0400	.0705
2	25	+++	.0355	.0290	.2730	.6085	.0395	.0145	.0540
		++-	.0285	.0265	.3005	.5915	.0365	.0165	.0530
		++±	.0240	.0330	.3050	.5780	.0440	.0160	.0600
		+-±	.0255	.0350	.2970	.5825	.0440	.0160	.0600
2	50	+++	.0030	.0200	.2935	.6435	.0370	.0030	.0400
		++-	.0020	.0180	.2850	.6545	.0360	.0045	.0405
		++±	.0035	.0110	.2795	.6655	.0355	.0050	.0405
		+-±	.0035	.0140	.2775	.6645	.0370	.0035	.0405
4	15	+++	.8130	.0375	.0190	.0445	.0250	.0610	.0860
		++-	.7890	.0585	.0220	.0470	.0295	.0540	.0835
		++±	.6195	.2445	.0605	.0605	.0270	.0525	.0795
		+-±	.5530	.2480	.0715	.0565	.0240	.0470	.0710
4	25	+++	.7325	.0590	.0340	.0940	.0400	.0405	.0805
		++-	.6855	.0870	.0510	.1035	.0415	.0315	.0730
		++±	.3660	.3220	.1350	.1170	.0360	.0240	.0600
		+-±	.3310	.3525	.1570	.1190	.0210	.0195	.0405
4	50	+++	.5095	.1605	.1375	.1680	.0185	.0060	.0245
		++-	.4360	.2005	.1765	.1660	.0175	.0035	.0210
		++±	.2135	.3375	.2670	.1650	.0160	.0010	.0170
		+-±	.2005	.3490	.2895	.1475	.0120	.0015	.0135

The relevant results will be found in the last column of each section of the table, giving the proportion of simulated samples in which the number of outliers declared exceeds the true number of outliers. As expected, these proportions substantially exceed the nominal 1% and 5% in the sets of simulations for n=15, but are at about the right value for n=25. However, for n=50 they are substantially below these nominal levels. This finding leads to the re-examination of Rosner's own application as described in § 4.4.

It was shown in that section that results depended on the amount of the slippage. It will be noted that in Table 4.6.3 the slippages are not particularly large: simulation (2000 runs) shows that Wilks' ordinary test for one outlier has a power of about 89% (at the nominal 5% level using Bonferroni bounds) for the combination of n=25, p=2 and squared distance=30, while the ordinary test for two outliers has a power of about 71% to 90% (depending on the directions of slippages) for the combination of n=25, p=2 and squared distance=20 (in this cases using simulated 5% critical values, from Chapter 3). For the latter combination, but with n=50, the power becomes 85% to 93%. Most of these powers are well below the 99% for Rosner's univariate sequential procedure with a slippage of 6 (see Table 4.4.1). Some supplementary runs for the multivariate case were therefore undertaken with a larger slippage, namely 50 in each dimension. Results are shown in Table 4.6.4.

Table 4.6.4

		Outliers declared at 1%				
n	outliers(s)	0	1	2	3	4+
15	+	0	.9835	.0090	.0020	.0055
50	+	0	.9920	.0080	.0000	.0000
50	+-	0	0	.9840	.0135	.0025
50	+±	0	0	.9890	.0100	.0010
		Outliers declared at 5%				
n	outlier(s)	0	1	2	3	4+
15	+	0	.9170	.0395	.0110	.0325
50	+	0	.9560	.0415	.0025	.0000
50	+-	0	0	.9300	.0520	.0180
50	+±	0	0	.9340	.0510	.0150

(Note: 2000 simulations; p=2 throughout)

One would hope to be able to predict the results of

this investigation on the following argument. If there is one contaminant and the slippage is very large, the outlier should be identified correctly by the test for one outlier with high probability. The contaminant is removed and the test in the reduced sample is now an application of a single-outlier test in a sample which strictly conforms to the null hypothesis. The error level should therefore be close to the nominal level, with a small excess due to the possibility of declaring two or more outliers in cases where the single-outlier test is not significant. Similar remarks apply to the case of two outliers, assuming that two successive separate identifications of the most extreme single outlier have the same effect as directly detecting the most outlying pair. The results in Table 4.6.4 do seem to agree with these predictions, remembering that the critical values are based on simulations for the case of two outliers, so introduce some inaccuracy.

Finally, Table 4.6.5 gathers together some results already presented in earlier tables, in order to give side-by-side comparisons between the performance of the two test procedures. However, the meaningfulness of thesecomparisons is limited, because of the restricted scope of the first procedure (inability to declare more than 2 or 3 outliers, depending on the version) and the different sizes of the procedures.

Table 4.6.5a Comparative performance of first and second outlier detection procedures in presence of one or two contaminants: proportion of times in 5000 simulations that less than correct number, correct number or more than corect number of points declared as outliers, at nominal 1% level.

			Procedure:	First, up to 2 Outl. declared			First, up to 3 Outl. declared			Second Outl. declared		
p	n	Out-liers	Squared dist-ances	Type	<	Correct	>	<	Correct	>	<	Correct	>
2	15	1	15	+	.8710	.1110	.0180	.8930	.0905	.0165	.8125	.1725	.0150
			30	+	.5580	.4165	.0255	.6100	.3700	.0200	.4600	.5240	.0160
		2	20	++	.7845	.2155	-	.7855	.1815	.0330	.8440	.1285	.0275
			20	+-	.7765	.2235	-	.7910	.1850	.0240	.8525	.1270	.0165
			20	+±	.8305	.1695	-	.8640	.1240	.0120	.9000	.0870	.0130
2	25	1	15	+	.7925	.1780	.0295	.8170	.1515	.0315	.2795	.7055	.0150
			30	+	.3370	.6225	.0405	.3835	.5750	.0415	.0115	.9770	.0115
		2	20	++	.5700	.4300	-	.5735	.3760	.0505	.6035	.3865	.0060
			20	+-	.5620	.4380	-	.5815	.3705	.0480	.5830	.4095	.0075
			20	+±	.5605	.3995	-	.6220	.3335	.0445	.6255	.3680	.0085
4	15	1	15	+	.9525	.0365	.0110	.9615	.0275	.0110	.9230	.0670	.0100
			30	+	.8165	.1675	.0160	.8480	.1390	.0130	.7490	.2285	.0225
		2	20	++	.9450	.0550	-	.9420	.0435	.0145	.9450	.0330	.0220
			20	+-	.9430	.0570	-	.9475	.0385	.0140	.9485	.0320	.0195
			20	+±	.9460	.0540	-	.9540	.0400	.0060	.9615	.0225	.0160
4	25	1	15	+	.9045	.0750	.0205	.9190	.0575	.0235	.8790	.1130	.0080
			30	+	.5740	.3875	.0385	.6155	.3465	.0380	.5155	.4775	.0070
		2	20	++	.7975	.2025	-	.7810	.1700	.0490	.9075	.0820	.0105
			20	+-	.8030	.1970	-	.8050	.1605	.0345	.9145	.0815	.0040
			20	+±	.8075	.1925	-	.8185	.1550	.0265	.9465	.0520	.0015
2	15	3	20	+++	1	-	-	.8770	.1230	-	.8965	.0670	.0365
				++-	1	-	-	.8585	.1415	-	.9200	.0630	.0170
				++±	1	-	-	.8225	.1775	-	.9025	.0825	.0150
				+-±	1	-	-	.8500	.1500	-	.9150	.0765	.0085
2	25	3	20	+++	1	-	-	.6240	.3760	-	.5795	.4085	.0120
				++-	1	-	-	.6200	.3800	-	.5955	.3935	.0110
				++±	1	-	-	.6095	.3905	-	.6100	.3775	.0125
				+-±	1	-	-	.6210	.3790	-	.6095	.3805	.0100
4	15	3	20	+++	1	-	-	.9860	.0140	-	.9720	.0125	.0155
				++-	1	-	-	.9800	.0200	-	.9755	.0080	.0165
				++±	1	-	-	.9680	.0320	-	.9690	.0120	.0190
				+-±	1	-	-	.9715	.0285	-	.9720	.0140	.0140
4	25	3	20	+++	1	-	-	.8835	.1165	-	.9475	.0295	.0230
				++-	1	-	-	.8655	.1345	-	.9495	.0315	.0190
				++±	1	-	-	.8470	.1530	-	.9540	.0310	.0150
				+-±	1	-	-	.8455	.1545	-	.9605	.0330	.0065

Table 4.6.5b Comparative performance of first and second outlier detection procedures in presence of one or two contaminants: proportion of times in 5000 simulations that less than correct number, correct number or more than corect number of points declared as outliers, at nominal 5% level.

				Procedure:	First, up to 2 Outl. declared			First, up to 3 Outl. declared			Second Outl. declared		
p	n	Out-liers	Squared dist-ances	Type	<	Correct	>	<	Correct	>	<	Correct	>
2	15	1	15	+	.6545	.2695	.0760	.7110	.2125	.0765	.5590	.3580	.0830
			30	+	.2800	.6270	.0930	.3350	.5715	.0935	.1885	.7200	.0915
		2	20	++	.5925	.4075	-	.5720	.3360	.0920	.6035	.2915	.1050
			20	+-	.5525	.4475	-	.5700	.3560	.0740	.5970	.3110	.0920
			20	+±	.5910	.4190	-	.6255	.3190	.0555	.6540	.2650	.0810
2	25	1	15	+	.5735	.3275	.0990	.6125	.2860	.1015	.1105	.8315	.0580
			30	+	.1380	.7335	.1285	.1760	.7005	.1235	.0000	.9295	.0705
		2	20	++	.3715	.6285	-	.3765	.4735	.1500	.3530	.5985	.0485
			20	+-	.3545	.6455	-	.3700	.4890	.1400	.3310	.6145	.0545
			20	+±	.3780	.6210	-	.4075	.4505	.1420	.3425	.6030	.0545
4	15	1	15	+	.8320	.1175	.0505	.8445	.0945	.0610	.7385	.1870	.0745
			30	+	.5695	.3490	.0815	.6275	.3100	.0625	.4425	.4585	.0990
		2	20	++	.8415	.1585	-	.8335	.1140	.0525	.8095	.0950	.0955
			20	+-	.8430	.1570	-	.8385	.1115	.0500	.8065	.1020	.0915
			20	+±	.8260	.1740	-	.8475	.1135	.0390	.8240	.1015	.0745
4	25	1	15	+	.7245	.2065	.0690	.7590	.1615	.0795	.7030	.2520	.0450
			30	+	.3290	.5495	.1215	.3635	.5080	.1285	.2645	.6790	.0565
		2	20	++	.6430	.3570	-	.6050	.2705	.1245	.7190	.2195	.0615
			20	+-	.6075	.3925	-	.6065	.2915	.1020	.7700	.1865	.0435
			20	+±	.5925	.4075	-	.6155	.2790	.1055	.7880	.1765	.0355
2	15	3	20	+++	1	-	-	.7375	.2625	-	.7110	.1630	.1260
			20	++-	1	-	-	.6825	.3175	-	.6970	.2070	.0960
			20	++±	1	-	-	.6540	.3460	-	.6735	.2390	.0875
			20	+-±	1	-	-	.6610	.3390	-	.7275	.2020	.0705
2	25	3	20	+++	1	-	-	.4485	.5515	-	.3375	.6085	.0540
			20	++-	1	-	-	.4175	.5825	-	.3555	.5915	.0530
			20	++±	1	-	-	.4070	.5930	-	.3620	.5780	.0600
			20	+-±	1	-	-	.3960	.6040	-	.3575	.5825	.0600
4	15	3	20	+++	1	-	-	.9435	.0565	-	.8695	.0445	.0860
			20	++-	1	-	-	.9375	.0625	-	.8695	.0470	.0835
			20	++±	1	-	-	.9015	.0985	-	.9245	.0605	.0795
			20	+-±	1	-	-	.9095	.0905	-	.8725	.0565	.0710
4	25	3	20	+++	1	-	-	.7505	.2495	-	.8255	.0940	.0245
			20	++-	1	-	-	.7180	.2820	-	.8235	.1035	.0210
			20	++±	1	-	-	.6935	.3065	-	.8230	.1170	.0170
			20	+-±	1	-	-	.6745	.3255	-	.8405	.1190	.0135

4.7 Use of sequentially applied test statistics

One role for outlier detection methods is in the automatic screening of the data, advocated by Gentleman and Wilk (1975). This is particularly valuable when the data are unlikely to be inspected closely by a trained eye; this situation might arise for various reasons including the automation of data collection and data reporting. Routine use of these methods also removes the difficulties over significance level introduced by the subjective decision to employ outlier testing in the light of some impression gathered from inspection of the data (Collett and Lewis, 1976).

Under these circumstances, the only difficulty remaining with the sequentially applied test is the choice of k, the maximum possible number of outliers. This too is effectively avoided if the test is used repeatedly in the same situation, for example with batches of similar data from the same laboratory. In this situation, it is possible to estimate the frequency of outlying values in the long run. The upper limit k could then be chosen so that the probability of a sample containing more than k outliers is sufficiently small. Table 4.7.1 shows the probability of having more than 2 or 3 contaminants in a sample of given size, given the probability that a randomly selected point is a contaminant.

Table 4.7.1 Probability of more than k discordant points in a sample of size n if a randomly selected point is discordant with probability p.

		p		
	n	0.01	0.02	0.05
k=2	10	.000114	.000864	.011504
	30	.003317	.021717	.187822
	50	.013817	.078427	.459465
	100	.079372	.323313	.881737
k=3	10	.000030	.000239	.003648
	30	.000996	.007599	.092535
	50	.004651	.032925	.294562
	100	.033623	.186606	.777056

These are simply binomial probabilities. Since a contaminant is not necessarily an outlier, and vice versa, this is not precisely the same as predicting the number of outliers, but should serve as close guide. Probably the test with k=3 would be thought very adequate with sample sizes of 10 even if the probability of a discordant observation were as big as 0.05. With p=0.01, perhaps a much more realistic value than 0.05 in most circumstances, k=3 seems adequate even for n=50.

As an illustration, however, the sequentially applied test is used here on a unique set of data rather than one from a series, in order to compare to another published method. Bacon-Shone and Fung (1987) illustrated their graphical method with a set of three-dimensional data (n=36) on milk transportation costs which they attribute to Johnson and Wichern (1982). These data are given here as Table 4.7.2.

Table 4.7.2 Transportation cost data (Johnson and Wichern, (1982), copied from Bacon-Shone and Fung, 1987): costs in dollars per mile of transporting milk from farm to dairy plant.

Fuel	Repair	Capital
16.44	12.43	11.23
7.19	2.70	3.92
9.92	1.35	9.75
4.24	5.78	7.78
11.20	5.05	10.67
14.25	5.78	9.88
13.50	10.98	10.60
13.32	14.27	9.45
29.11	15.09	3.28
12.68	7.61	10.23
7.51	5.80	8.13
9.90	3.63	9.13
10.25	5.07	10.17
11.11	6.15	7.61
12.17	14.26	14.39
10.24	2.59	6.09
10.18	6.05	12.14
8.88	2.70	12.23
12.34	7.73	11.68
8.51	14.02	12.01
26.16	17.44	16.89
12.95	8.24	7.18
16.93	13.37	17.59
14.70	10.78	14.58
10.32	5.16	17.00
8.98	4.49	4.26
9.70	11.59	6.83
12.72	8.63	5.59
9.49	2.16	6.23
8.22	7.95	6.72
13.70	11.22	4.91
8.21	9.85	8.17
15.86	11.42	13.06
9.18	9.18	9.49
12.49	4.67	11.49
17.32	6.86	4.44

Bacon-Shone and Fung obtain (their Table 3.2) the following results for Wilks' test, with α indicating the unconditional significance levels:

Table 4.7.3

No of outliers tested	Wilks' statistic	α	Observations selected
1	0.481	<0.005	9
2	0.278	<<0.005	9,21
3	0.196	<<0.005	9,21,36
4	0.148	<<0.005*	9,21,36,20

* This value, not given by Bacon-Shone and Fung, was computed here from 1000 simulations.

They then employ their graphical method and assert that there is no evidence for more than two outliers, the result of the three outlier test being due to the effect known as swamping. They interpret their graphs as indicating that definitely point 9 and probably point 21 should be regarded as outliers. Evidence for point 21 is drawn too from what they call a sequential procedure. That is, point 9 is eliminated and the standard Wilks tests for one, two, and more outliers are carried out on the remaining 35 points in the reduced sample, giving (their Table 3.3, with the addition of the result for three outliers, computed from 4000 simulations):

Table 4.7.4

No. of outliers	Wilks' statistic	α	Observations selected
1	0.577	0.02	21
2	0.407	0.04	21,36
3	0.287	0.028	21,36,20

The significant result of the two-outlier test presumably indicates that swamping is still in effect, although Bacon-Shone and Fung make no remark on this result.

To obtain comparative results using Rosner's first sequentially applied test, critical values have been computed for n=36 and p=3. These are as follows:

Table 4.7.5

Max. no of outliers	outliers	α 0.01	0.025	0.05	0.10
3	1	0.5201	0.5527	0.5794	0.6096
	2	0.6063	0.6287	0.6488	0.6702
	3	0.6375	0.6575	0.6741	0.6926
2	1	0.5335	0.5656	0.5954	0.6237
	2	0.6134	0.6385	0.6594	0.6808

Alternatively, these could be obtained approximately by interpolation from the previous tables.

The three most extreme points are numbers 9, 21 and 36 in the data file, with test statistics D_1=0.4815, D_2=0.5770 and D_3=0.7058. All these details match the results of Bacon-Shone and Fung. It can be seen that D_3 is so large that we would not accept the existence of three outliers even at the 10% level. Testing therefore passes to D_2, which is below the 1% critical value of 0.6063. Consequently, the evidence seems to be very clear that this set of data contains two outliers, points 9 and 21.

Rosner's second procedure simply uses the unadjusted critical values of Wilks' statistic test at each step. The results, for testing for up to 3 outliers, are set out in Table 4.7.6. Details of tests beyond 3 outliers (up to 10) are not presented because all calculated values of Wilks' statistic were above 0.7, whereas even the 10% critical value was only 0.64 or less.

Table 4.7.6

Critical values of Wilks' statistic

Sample size	Omit point	Wilks' statistic	1%	2.5%	5%	10%
36	9	.481	.558	.592	.619	.648
35	21	.577	.548	.583	.611	.640
34	36	.706	.539	.574	.602	.632

Accepting the validity of the approximation involved in the use of these unadjusted significance levels, the conclusion would be that the existence of two outliers (9,21) can be accepted at the 2.5% level, but three outliers would not be accepted even at 10%.

The results from either version of Rosner's test indicate the same conclusion, that 9, 21 can be regarded as outliers; this is without needing graphical supplement to overcome the swamping problem.

CHAPTER 5

ROHLF'S GENERALIZED GAP TEST FOR MULTIVARIATE OUTLIERS

5.1 Introduction

In univariate problems, any outliers must lie at the extremes of the ordered sample values. This simple observation makes possible certain characterizations of outliers. In particular, if there are exactly k upper outliers, then the gap between the successive order statistics $X_{(n-k+1)}-X_{(n-k)}$ should be unusually large: this test was proposed by Irwin (1925). The idea was revived by Tietjen and Moore (1972) and Tiku (1975), who propose using the gaps to see how many outliers to test for and then using an optimal test for that number. The gap tests themselves are not optimal but, as Hawkins (1980a) says, they are very attractive if there is reason to believe that the contaminants all follow the same distribution, for then the data ought to fall into two well-separated clusters. Various other tests using gaps were proposed by Dixon (1950). His test criteria are ratios of differences between order statistics: one example was shown in Example 1 of § 1.4. Hawkins (1980a) comments that these criteria have lost favour, partly because they do not extend to other situations, such as linear models.

Gap tests thus form, or have formed, a significant part of the theory on testing for outliers in univariate samples. For multivariate data, on the other hand, they cannot fill the same role. It is difficult to define gaps in terms of order statistics, because of the lack of a convenient concept of ordering (Barnett, 1976). Nonetheless, a form of gap test has been proposed for multivariate data, by Rohlf (1975). Far from having lost favour, this seems from the literature never to have had any. But potentially it has advantages over the main

general-purpose test, the Wilks test, if there is more than one outlier. Rohlf claims that his test should be less susceptible to masking than the Wilks test and points out that it requires the same computational effort whatever number of outliers is examined, whereas the Wilks procedure requires a rapidly increasing number of comparisons. (The reason for this is that Rohlf's test is not in fact a test for any particular number of outliers: if the test results in the declaration of the presence of outliers, then their number is inferred from the structure of the sample, as described below.) A third possible advantage which could be added is that the Bonferroni bounds usually used for the Wilks test are poor approximations for more than one outlier (Hawkins, 1980a; see also Chapter 3 of this thesis). For these reasons, it is worth investigating Rohlf's test in some detail.

Rohlf's procedure starts by defining a distance measure between the points in a sample of independent p-dimensional data vectors, and constructing the minimum spanning tree (MST) for these distances. A spanning tree of a set of n points is a set of n-1 out of the $\binom{n}{2}$ possible edges (connections between pairs of points) such that:

(i) each point is connected to at least one another;

(ii) every point is accessible from every other point by following some path along edges of the tree;

(iii) there are no closed loops - for any (i,j), there is a unique path of edges by which point j can be reached from point i.

The MST is that spanning tree which has minimum sum of length of edges. It has a wide variety of applications in operational research problems. Its importance in multivariate data analysis is due to its equivalence to single linkage cluster analysis and its usefulness as a supplement to graphical displays in two dimensions (Gower

and Ross, 1969).

Rohlf then suggests that the distances (or edges) in the MST be taken as analogous to gaps in a univariate sample. In particular, the presence of an outlier would be indicated by the fact that the largest distance in the MST was unusually large in comparison to the rest. Rohlf's idea is basically to examine this largest distance in the MST. This might be done informally in a probability plot, but Rohlf also gives a more formal, significance testing approach (although in subsequent correspondence, he seems rather defensive about this and claims that the plot was intended to be the main method; Rohlf, 1977). One reason for the lack of popularity of this procedure may be that its theoretical basis is not very sound. Rohlf argues for a gamma distribution for the set of distances from the MST and proposes estimating the parameters of this gamma, followed by a test using these parameter estimates as if they were true values. He presents a table of Bonferroni upper bounds for the ratio of maximum to average distance in the MST. Strangely, Barnett and Lewis (1984) refer to this table without pointing out that it is the same, except for multiplying each entry by the sample size, as the first table in their book, for testing discordancy in a gamma sample.

The contents of the analysis of Rohlf's gap test are as follows. Firstly, the performance of the test in the form suggested by Rohlf is investigated. Secondly, a modified testing procedure, using simulated rather than approximated percentage points, is considered. Finally, a further variation, replacing Rohlf's Euclidean distances by generalized distances, is examined.

5.2 Examination of Rohlf's procedure

5.2.1 Choice of distance measure

Rohlf's first step is the selection of a distance

measure. He opts for a Euclidean distance on standardized variables, so that given the nxp data matrix $X=(x_{ik})$ and sample standard deviation s_k for variable k, the distance between points i and j is

$$d_{ij}=\{\sum_{k=1}^{p}(x'_{ik}-x'_{jk})^2/p\}^{1/2} \qquad (5.2.1)$$

where $x'_{ik}=x_{ik}/s_k$. The standardization is to equalize the impacts of variables with differing variances. Rohlf actually suggests using some unspecified form of robust estimator of s_k, his description of the first step of his procedure being:

"Perform a univariate test for outliers (such as Dixon's [1950] gap test) so that one can obtain fairly good estimates of the standard deviations for each variable."

Robust estimation seems logically necessary after choosing to use standardization in (5.2.1). Otherwise, no matter how large an outlier might be in a particular dimension k on the original scale of measurement, its inclusion in the computation of s_k means that distances to this outlier in this dimension are constrained to be of only the same order of magnitude as distances between points on other dimensions where no outliers appear. This would mean a very severe lack of sensitivity.

The robust estimation was carried out in the present study by simply trimming the sample (separately in each dimension) by omitting either the most extreme observation at each end of the ordered sample values or the two most extreme at each end. The standard deviation of the remaining n-2 or n-4 observations was then computed and used in place of s_k in (5.2.1). This quantity does not of course estimate the population standard deviation unless it is adjusted; however, the adjustment would be the same for each dimension, so has no effect in the analysis.

Rohlf also remarked that the familiar generalized

squared distances

$$d_{ij}^2=(x_i-x_j)'S^{-1}(x_i-x_j) \qquad (5.2.2)$$

could be used in place of those given by (5.2.1), without elaborating on this. The analysis presented here will first be in terms of (5.2.1), before considering use of (5.2.2).

5.2.2 Distribution of distances

Rohlf argues that if the data were independent vectors from $N_p(\mu,I)$, then the squared Euclidean distance between two randomly selected points would be distributed as $2X_p^2$.

In fact this applies more generally than Rohlf states, to standardized squared distances from $N_p(\mu,\Sigma)$ where $\Sigma=\text{diag}(\sigma_1^2,\ldots,\sigma_p^2)$, since

$$x_{ik}-x_{jk} \sim N(0,2\sigma_k^2)$$

independently in each dimension $k=1,\ldots,p$, so

$$\frac{(x_{ik}-x_{jk})}{\sqrt{2}\,\sigma_k} \sim N(0,1)$$

and

$$\sum_{k=1}^{p} \frac{(x_{ik}-x_{jk})^2}{2\sigma_k^2} \sim X_p^2$$

(However, the same distribution cannot hold when σ_k is replaced by a sample estimate s_k, when the N(0,1) quantities are replaced by t distributions.) The distribution $2X_p^2$ is the same as the gamma distribution $G(1/4,p/2)$ with scale parameter $\lambda=1/4$ and shape parameter $\eta=p/2$. Rohlf then says that, if the variables are correlated, a randomly selected d_{ij}^2 should still approximately follow the gamma distribution, but with different parameter values.(Note that Rohlf divides the squared distance by p; this is not important, because a

constant multiple of a gamma distributed random variable is also gamma distributed.)

Some of the details behind Rohlf's statement can be found in Gnanadesikan (1977; p.233). If A is a non-random matrix and vectors Y_i are a random sample from $N(0,\Phi)$, then squared distances $Y_i'AY_i$ are distributed as the linear combination $c_1\chi_1^2+\ldots+c_r\chi_r^2$, where the c_i are positive eigenvalues of $A\Phi$, the χ^2's are independent chi-squared values each with one d.f. and r is the rank of A. It is then a well-known approximation that a gamma distribution comes close to this combination of chi-squared variates, for suitable choice of parameters. The same result is used, as a further approximation, when A is an estimate from the sample. To apply this to distances between points x_i distributed as $N(\mu,\Sigma)$, one looks at $d_{ij}^2=Y_{ij}'AY_{ij}$ where $Y_{ij}=x_i-x_j \sim N(0,2\Sigma)$.

Further approximation is introduced through the Y_{ij}'s not being independent. In fact the sample of n points provides n(n-1)/2 values of d_{ij}^2 and clearly their heavy interdependence could mean that it is unlikely that this entire set of values follows the gamma distribution at all well. However, interest here lies only in the selected subset of n-1 distances which make up the MST. These are certainly not a random sample of all distances, and Rohlf (1977) gives some simulation results showing that their statistical properties are quite different from those of randomly selected distances. Rohlf also claims that empirically a gamma distribution does fit the squares of these distances quite well. Because the MST distances tend to be among the smaller values of d_{ij}, the theoretical parameters of the gamma distribution would not apply even if the variables were independent, so the parameters λ and η need to be estimated from the data, that is, from the MST.

In the special case of generalized distances, given by

$A=S^{-1}$ in $Y_i'AY_i$ an exact distributional result is known. For a randomly selected pair of points, $d^2_{ij}/2(n-1)$ follows the Beta distribution with parameters $p/2$ and $(n-p-1)/2$ (Gnanadesikan and Kettenring, 1972). There are also some results on correlations between distances within the same sample. Specifically, any squared distances d^2_{ij} and $d^2_{i',j'}$ have asymptotic correlation 0 if $i \neq i'$ and $j \neq j'$ but 0.5 if one index is in common (Gnanadesikan and Kettenring, 1972). Again, these results seem to have little bearing on the special sample of distances making up the MST. A gamma approximation may be tried, as with the standardized Euclidean distances. Rohlf remarks that "preliminary simulation runs do not seem to indicate any distinct advantage" in using generalized rather than Euclidean distances. This appears to be correct: Figures 5.2.1 to 5.2.4 show examples of gamma probability plots of MST distances for both Euclidean and generalized distances. These plots were not selected; they are simply the first of a number of runs. The general indication seems to be that a gamma distribution is a reasonably good approximation and that this holds just as well for either form of distance measure.

5.2.3 Testing procedure

Following the above discussion, Rohlf's procedure is based on the fit of a gamma distribution to the distances of the edges in the MST. Since theory does not supply values for the parameters of the distribution, they must be estimated. These parameter estimates can then be used either to construct a gamma probability plot for visual assessment of the MST distances or to permit formal testing of some aspect of the MST. Attention here will be focused on the formal testing, which is carried out for the length of the longest edge in the MST in relation to the total length, because visual assessment is impractical with the large-scale simulation study which will be called for. In any case, Rohlf claims advantages for his

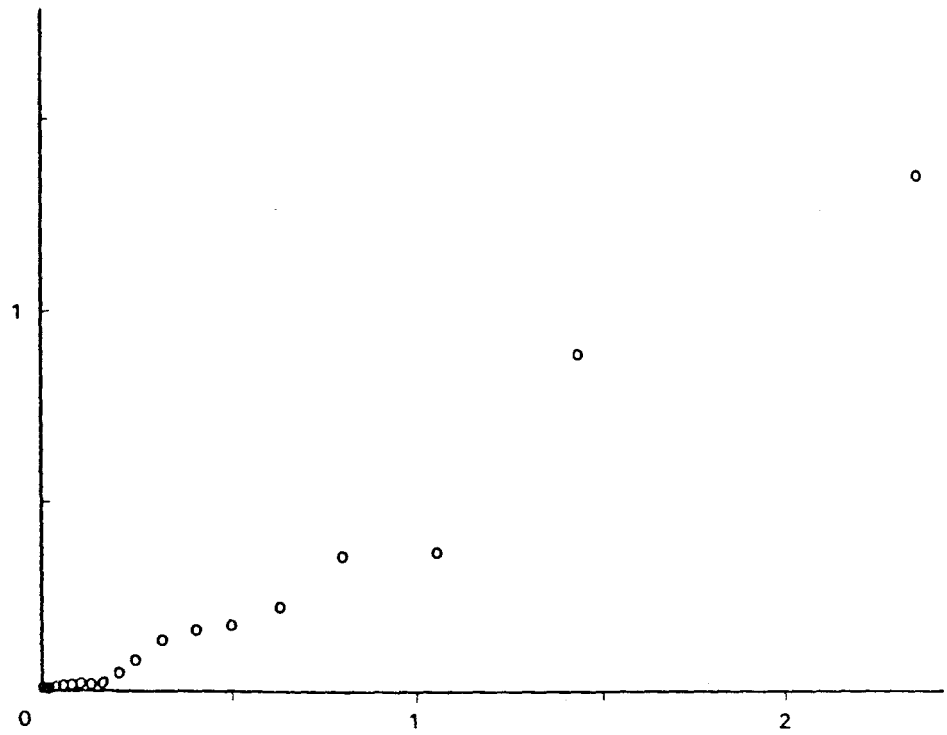


Figure 5.2.1 Q-Q plot of Euclidean MST distances (ordinate) against expected gamma order statistics (abscissa): n=20, p=2, ρ=0.6.

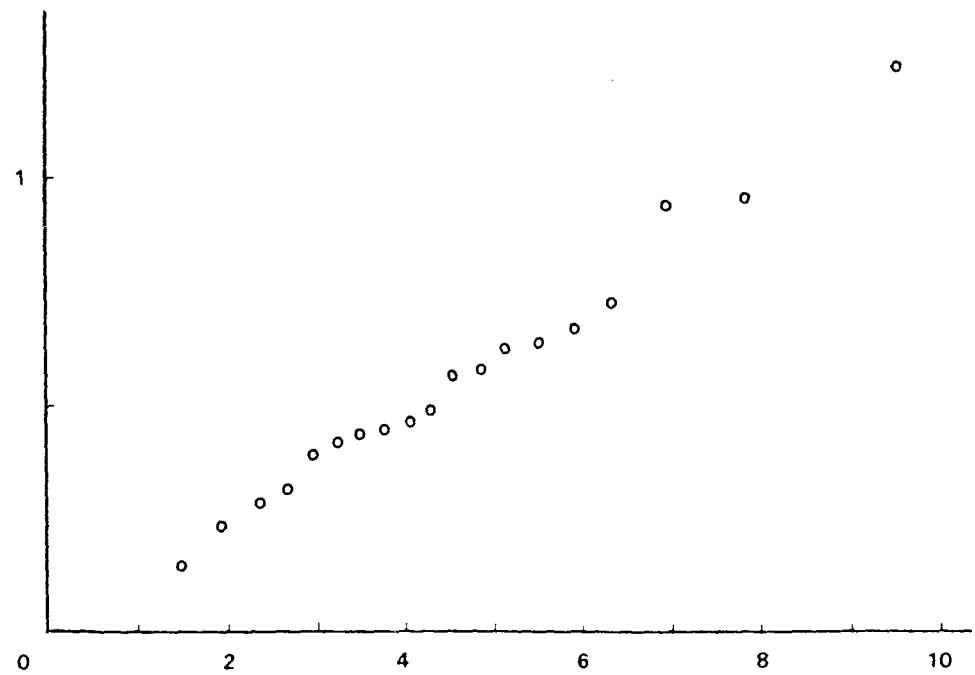


Figure 5.2.2 Gamma Q-Q plot of Euclidean MST distances: n=20, p=4, ρ=0.6.

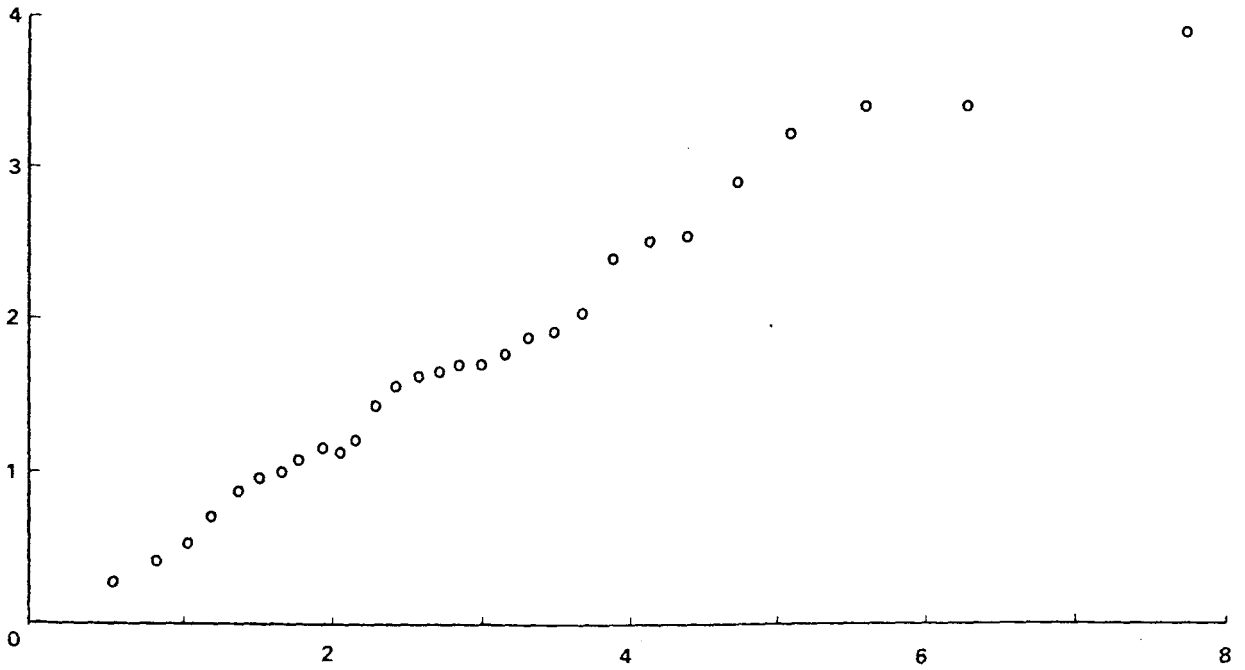


Figure 5.2.2 Gamma Q-Q plot of generalized MST distance: n=30, p=4, ρ=0.

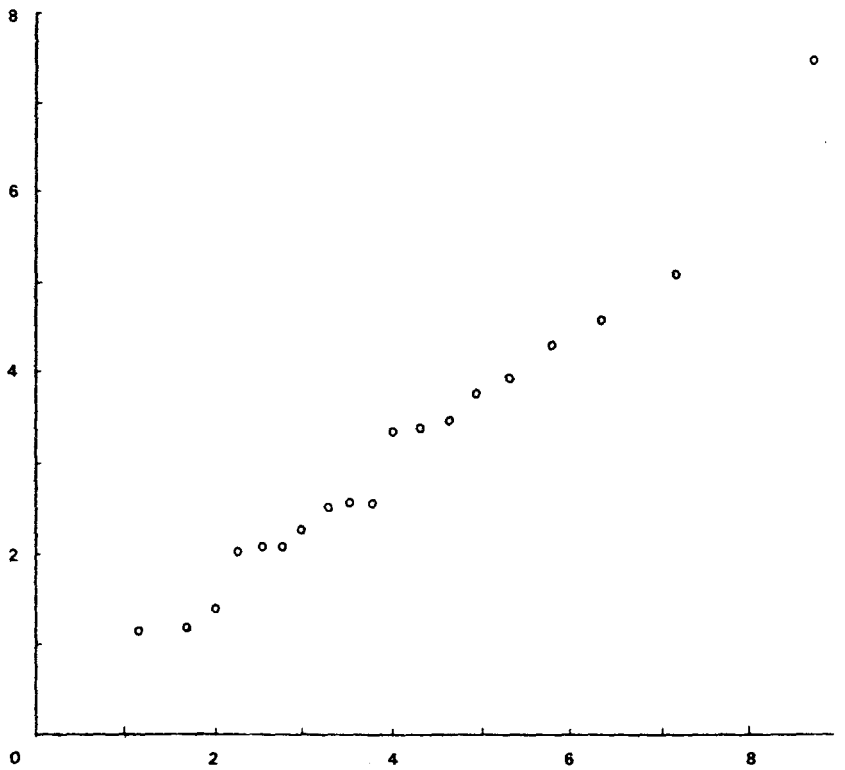


Figure 5.2.4 Gamma Q-Q plot of generalized MST distance: n=20, p=5, ρ=0.

procedure over that of Wilks and so, since Wilks' is a formal test, it is this aspect of Rohlf's procedure which must be examined.

Rohlf's procedure depends on the following argument. Suppose that the MST distances $\{d_i\}$ follow the gamma distribution $G(\lambda,\eta)$, with known parameters. Then if $Y_i=d_i^2$, any ratio

$$\frac{Y_i}{\sum Y_i}$$

follows a beta distribution with parameters η and $(m-1)\eta$. Here, m is being used to denote the number of edges in the MST, so equals the sample size minus one. The m such ratios are not independent, so, as usual, obtaining the distribution of the maximum ratio is not possible, but the first Bonferroni approximation can be used. Rohlf presents a table of these approximations to the 1% and 5% points, for the test statistic $\max Y_i/\bar{Y}$, for a range of values of m and of η. He proposes using this table (by interpolation) with the sample estimate $\hat{\eta}$ substituted for the true unknown η. There are consequently three approximations involved in the test procedure:

(i) the gamma distribution is an approximation to the distribution of distances in the MST;

(ii) the theory cannot take account of the use of an estimate of η;

(iii) Bonferroni approximations are needed for the percentage points.

Any one of these might be a good approximation on its own. Unfortunately, their cumulative effect seems to result in unacceptably imprecise values for the tests. This can be seen from the following results of simulation comparisons to Wilks' test for one outlier.

The details of the computation of the test statistic

are as follows. The algorithm of Ross (1969) was used to obtain the MST, using standardized Euclidean distances with trimmed estimates of standard deviations. Maximum likelihood estimates of the parameters λ and η of the gamma distribution fitted to the MST distances were found by Newton-Raphson iteration, incorporating Bernardo's (1976) algorithm to compute the psi (digamma) function. Probabilities from the beta distribution were obtained by using the IMSL subroutine MDBETA to evaluate the incomplete beta integral. Direct computation of the probabilities seemed more convenient for a simulation study than interpolation in the published table. In fact, it seems the best option for use of the test in practice, since anyone with the facilities to evaluate the MST and fit the gamma distibution can also evaluate the incomplete beta integral.

5.2.4 The number of outliers declared

As mentioned earlier, Rohlf's test is not a test for any specific number of outliers: it is applied in the same form in all circumstances and does not have different variants depending on the hypothesized possible number of outliers. If the significance test leads to a declaration that the sample is not homogeneous, then the number of outliers must be determined by inspection of the MST. If the largest edge links a single point to the rest of the sample, then one outlier is indicated. Otherwise, the number of outliers is the minimum of the number of points making up the two isolated clusters which would result on removing the longest edge of the MST. Notice that this means that multiple outliers can only be detected if they form a cluster in this sense. Hence the presence of outliers in opposite directions away from the main body of data would never be detectable.

Determination of the number of points in a cluster is a trivial matter by eye, but for a simulation study,

inspection must be automated. An algorithm to determine the number of outliers is as follows. Ross' (1969) algorithm to construct the MST from a sample of size n returns a vector B of dimension n, in which element B(i) contains the index of one of the points to which sample member i is joined in the MST ($i \geq 2$). Distances between i and B(i) are held in array C. The first step in determining the number of outliers is to search through C to identify the points k and l=B(k) which are joined by the longest edge of the MST. The algorithm then operates on k, but first the special case l=1 has to be checked. In this case, search B to see if B(i)=1 for any other $i \geq 2$. If not, then 1 is not connected to any point other than k, so there is only one outlier. Otherwise, the procedure continues as in the general case.

The steps are as follows:

1. Let OUTL=1.

2. Search through B for $i \geq 2$ for the first time that B(i)=k, at i=m, say. If it does not happen at all, then go to step 4.

3. Set B(m)=0
   Set OUTL=OUTL+1
   Run through B for $i \geq 2$ and for every j with B(j)=m, set B(j)=k.
   Go to 2.

4. Number of outliers=min(OUTL,n-OUTL).

The logic behind step 3 is the identification of any other points m, besides l, which are linked to k in the MST. Points linked to m must also be counted as falling in the same cluster of points as k and setting B(j)=k for such points ensures that this will be done, while setting B(m)=0 prevents the recounting of this point. The logic of step 4 is that it is not known whether, in choosing to

work with k rather than 1, one has selected a member of the main body of points or a member of the smaller cluster of outliers.

A Fortran coding of this algorithm is as follows. The input arrays C and B of dimension N are as in Ross' algorithm. The parameter NOUT holds the number of outliers on output.

```
      SUBROUTINE NOUTL(N,C,B,NOUT)
C
      REAL C(N)
      INTEGER B(N),HALFN
C
      K=2
      L=B(2)
      XMAX=C(2)
C
      DO 200 I=3,N
        IF(C(I).GT.XMAX) GO TO 21
        GO TO 200
   21   XMAX=C(I)
        K=I
        L=B(K)
  200 CONTINUE
      IF(L.EQ.1) GO TO 33
      GO TO 2
   33 DO 210 I=2,N
        IF(I.EQ.K) GO TO 210
        IF(B(I).EQ.1) GO TO 2
  210 CONTINUE
      NOUT=1
      GO TO 500
    2 NOUT=1
    1 DO 260 I=2,N
        IF(B(I).EQ.K) GO TO 23
        GO TO 260
   23   B(I)=0
        NOUT=NOUT+1
        DO 270 J=2,N
  270     IF(B(J).EQ.I) B(J)=K
        GO TO 1
  260 CONTINUE
      HALFN=N/2
      IF(NOUT.GT.HALFN) NOUT=N-NOUT
  500 CONTINUE
C
      RETURN
      END
```

5.3 Simulation studies of Rohlf's procedure and modifications

5.3.1 Comparison between Rohlf's and Wilks' tests

In the first study, Rohlf's original test (standardized Euclidean distances, robust estimation of dispersion by trimming, Bonferroni bounds for the largest distance of the MST) was compared to Wilks' test. The comparison was in terms of the power in detecting a single outlier, at the 5% significance level, and also using Bonferroni bounds for Wilks' test. Table 5.3.1 shows results for various combinations of sample size n, dimensionality p and correlation ρ between dimensions (taken to be equal for all pairs of dimensions). Each figure is based on 8,000 simulations, with the two statistics computed from the same data. The data were generated from the multivariate normal distribution with mean zero and unit variances, with a single contaminant created by adding u units to each dimension for the first point in the sample, where u was determined so that the squared generalized distance of the slippage from the origin in the metric of the population covariance matrix was 30.

Table 5.3.1 Powers of Rohlf's and Wilks' statistics at nominal 5% level in the presence of a single outlier.

Data description			% of times outlier declared	
n	p	ρ	Rohlf	Wilks
10	2	-.4	64.1	64.3
10	2	.4	84.6	63.0
10	3	.5	89.5	41.2
10	3	0	75.0	40.5
50	2	.4	98.7	92.4
50	4	.4	97.7	82.2

At first sight, the selection of results in Table 5.3.1 might be taken to indicate that Rohlf's test is a very good one, with power vastly in excess of that of Wilks'

test, in most circumstances. However, it should be remembered that the Wilks' test is obtained by maximum likelihood under the model used in the data simulation. Furthermore, the conservative Bonferroni percentage points for the Wilks' test are, as discussed in Chapter 3, quite good. For example, a nominal 5% test appears from table 3.2.3c to be actually at about 4.9%. These two facts, that Wilks' test is based on maximum likelihood and that its percentage points are quite accurate, imply that Rohlf's test can only appear to be vastly better if its percentage points are quite inaccurate and not conservative, but in the opposite direction. For example, a nominal 5% test might be a true 10% test. Differences of this kind, and in this direction, render the test in this form unusable as a formal test statistic.

It should be noted that this conclusion is not dependent on the use of Euclidean rather than any other distance measure. The simulations selected in Table 5.3.1 include a case of ρ=0, where sample differences between different measures ought to be very small.

To which of the three approximations listed in § 5.2.3 is the inaccuracy of these percentage points owing? It cannot be to the fact that they are Bonferroni bounds, since that would cause conservatism. Nor, from the empirical plots, does it appear to be due to the assumption of a gamma distribution. It must therefore arise from the use of the estimated parameters of the gamma as if they were known values.

5.3.2 Use of an average value of η

Since the approximations involved in Rohlf's testing procedure appear to be too inaccurate and since no superior approximation presents itself, the investigation now turns to the use of simulation results to provide an improved procedure.

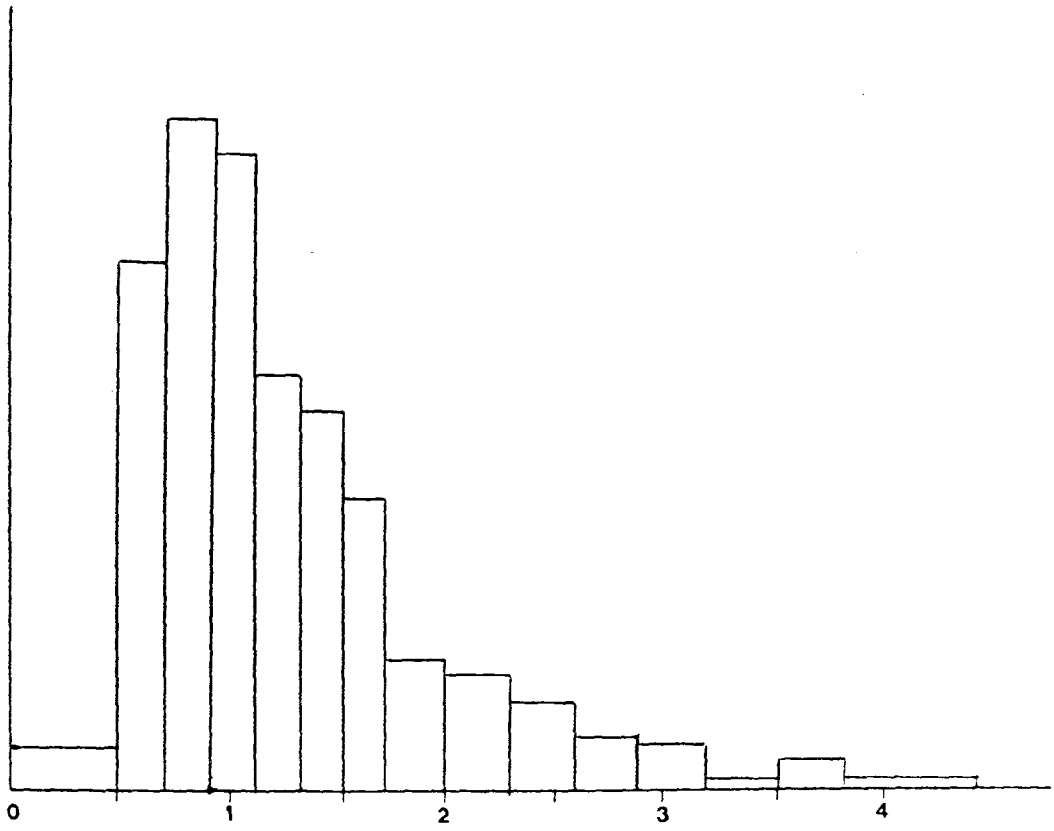


Figure 5.3.1 Sampling distribution of $\hat{\eta}$: 500 simulations, n=10, p=2, ρ=.2.

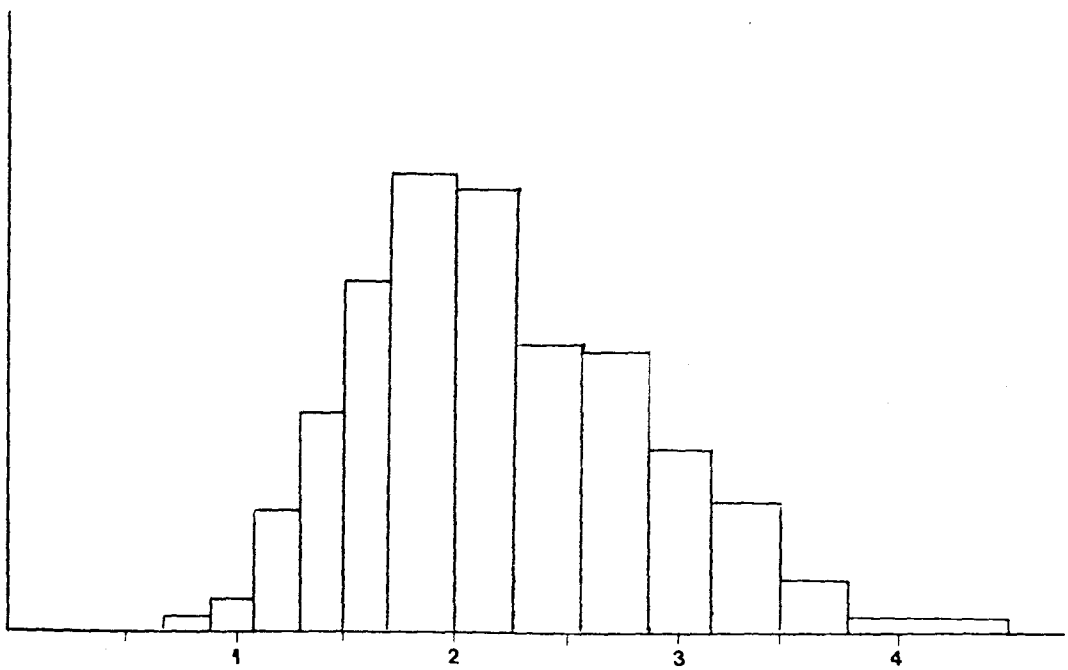


Figure 5.3.2 Sampling distribution of $\hat{\eta}$: 500 simulations, n=30, p=4, ρ=.4.

The stage of Rohlf's method which seems to be the particularly important source of inaccuracy is the use of the sample $\hat{\eta}$ as if it were the true η. Figures 5.3.1 and 5.3.2 show two examples of sampling distributions of $\hat{\eta}$, illustrating that this quantity is extremely variable between samples. Matters might therefore be improved if the average $\hat{\eta}$ for given n and p can be found and this used in place of the specific sample $\hat{\eta}$. This presumably only works if other aspects of the structure of the problem do not affect $\hat{\eta}$, including the correlations between the different dimensions. However, this seems unlikely to be the case with Euclidean distances and the results in Table 5.3.2 show that this is indeed so. If ρ is the correlation between any pair of dimensions, then the average value of $\hat{\eta}$ falls as ρ moves away from zero in either direction. The fall is rather large when ρ becomes large, such as 0.8. Usually, the covariance matrix is unknown and, consequently, the appropriate value of η - which is being selected to avoid sampling effects - could only be found if sample estimates of the covariances were used. In other words, one sampling effect is replaced by another, which is unlikely to lead to any advantage.

Table 5.3.2. Average values of gamma shape parameter estimated in minimum spanning tree of equicorrelated multivariate normal data, using Euclidean distances of the form 5.2.1 with trimmed standard deviations : 500 simulations.

		ρ				
		-0.4	0	0.2	0.4	0.8
p=2	n=10	1.3226	1.4236	1.3923	1.3608	1.0845
	20	1.0003	.9881	1.0133	.9768	.9138
	30	.9028	.8938	.9064	.9042	.8466
p=3	n=10	2.1107	2.5170	2.4424	2.3195	1.7806
	20	1.6420	1.8775	1.8141	1.7616	1.5118
	30	1.5610	1.6590	1.6695	1.5739	1.4269
		ρ=-.3				
p=4	n=10	3.2780	3.9612	3.7871	3.5280	2.4985
	20	2.2996	2.9705	2.7548	2.6905	2.2409
	30	2.1739	2.5192	2.5008	2.3924	1.9928

5.3.3 Simulated percentage points

Since procedures based on $\hat{\eta}$ or an average of $\hat{\eta}$ seem to be ineffective, attention will now be turned away away from attempting to construct percentage points using the gamma distribution; instead, the possibility of simulating percentage points will be investigated.

The investigation was carried out in the context of equicorrelated normal data, with

$$\Sigma=\sigma^2\{(1-\rho)I+\rho J\}$$

where J is the pxp matrix whose entries are all ones. Data were generated for different combinations of n, p and ρ, as follows:

p=2; n=10, 20 and 30; ρ=0,±.1,±.2,±.4,±.6,±.8,±.9
p=3; n=20; ρ=-.45,0,±.2,±.4,.6,.8,.9
p=4; n=20; ρ=-.3,0,±.2,.4,.6,.8,.9

At each combination, 2000 samples were simulated. The 1, 2.5, 5 and 10% critical values of $\max Y_i/\bar{Y}$ were recorded; these are shown in Table 5.3.3 (a-c).

Table 5.3.3a Percentage points of max $Y/\bar{Y}$ statistic for minimum spanning tree in equicorrelated normal data, obtained from 2000 simulations.

	p=2 n=10				p=2 n=20			
	1%	2.5%	5%	10%	1%	2.5%	5%	10%
-.9	.8246	.7677	.7133	.6341	.6983	.6434	.5706	.4949
-.8	.7799	.7233	.6591	.5799	.6425	.5757	.5161	.4385
-.6	.7226	.6568	.5935	.5212	.5797	.5176	.4525	.3889
-.4	.6898	.6182	.5607	.4987	.5480	.4805	.4281	.3663
-.2	.6969	.6224	.5523	.4816	.5046	.4442	.3966	.3455
-.1	.6720	.5945	.5428	.4800	.5181	.4516	.4016	.3533
0	.6856	.6042	.5490	.4696	.5135	.4651	.4115	.3493
.1	.6806	.6164	.5512	.4953	.5002	.4498	.4079	.3536
.2	.6766	.6169	.5498	.4945	.5181	.4657	.4147	.3634
.3	.6721	.6119	.5639	.4921	.5320	.4693	.4223	.3672
.4	.6764	.6226	.5782	.5049	.5463	.4705	.4202	.3690
.5	.6991	.6451	.5760	.4978	.5841	.4996	.4322	.3688
.6	.7191	.6470	.5865	.5174	.5949	.5190	.4719	.3869
.8	.8142	.7547	.6858	.5984	.6681	.6022	.5297	.4422
.9	.8024	.7531	.6968	.6251	.6879	.6346	.5625	.4855

Table 5.3.3b Percentage points of max $Y_i/\bar{Y}$ statistic for minimum spanning tree in equicorrelated normal data, obtained from 2000 simulations.

	p=2 n=30			
	1%	2.5%	5%	10%
-.9	.6588	.5646	.5049	.4297
-.8	.5638	.4964	.4574	.3936
-.6	.4983	.4431	.3942	.3366
-.4	.4987	.4221	.3669	.3151
-.2	.4514	.4008	.3571	.3091
-.1	.4514	.3836	.3463	.2999
0	.4363	.3912	.3485	.2988
.1	.4435	.3921	.3445	.3006
.2	.4585	.3998	.3467	.3022
.3	.4699	.3904	.3467	.2971
.4	.4822	.3982	.3524	.3025
.5	.5372	.4407	.3879	.3232
.6	.5150	.4348	.3845	.3210
.8	.6053	.5064	.4438	.3666
.9	.6672	.6007	.5373	.4485

Table 5.3.3c Percentage points of max $Y_i/\bar{Y}$ statistic for minimum spanning tree in equicorrelated normal data, obtained from 2000 simulations.

	p=3 n=20				p=4 n=20			
	1%	2.5%	5%	10%	1%	2.5%	5%	10%
-.45	.4449	.3864	.3514	.3045				
-.4	.4169	.3687	.3198	.2758				
-.3					.3180	.2810	.2523	.2254
-.2	.3644	.3224	.2895	.2499	.2924	.2523	.2315	.2054
0	.3330	.3028	.2770	.2393	.2769	.2461	.2238	.1983
.2	.3750	.3313	.2842	.2498	.2754	.2469	.2239	.2018
.4	.4010	.3438	.3059	.2595	.3219	.2790	.2489	.2160
.6	.4497	.3968	.3368	.2922	.3871	.3250	.2783	.2385
.8	.5447	.4845	.4206	.3570	.5156	.4454	.3777	.3134
.9	.6720	.5928	.4985	.4180	.6005	.5385	.4714	.3864

The power of Rohlf's test was investigated first under the assumption that ρ was known, so that data were generated from a population with given ρ and the value of Rohlf's test statistic was then compared to the simulated percentage points for the same ρ. Powers in comparison to Wilks' test calculated from the same data are shown in Table 5.3.4. In this comparison, the simulated percentage points for Wilks' test, obtained in Chapter 3, are used rather than the conservative Bonferroni points. It appears that Rohlf's test is much more powerful, especially in those cases (n small compared to p) where Wilks' test is not very effective.

Table 5.3.4 Power comparisons for detection of one outlier between Rohlf's and Wilks' tests in 5000 simulated samples for each combination of n, p and ρ, using simulated percentage points for both tests and treating ρ as known: outlier slippage equal in each dimension, squared generalized distance 30.

p	n	ρ	% of times outlier declared: Rohlf	Wilks	Rohlf-not Wilks	Wilks-not Rohlf
At 1% level						
2	10	0	54.80	31.68	24.32	1.20
		0.6	68.68	31.52	37.34	1.80
2	20	0	77.54	65.56	13.68	1.70
		0.6	83.88	66.34	18.58	1.04
4	10	0	44.18	7.42	37.70	0.94
		0.6	81.04	7.28	73.84	0.08
4	20	0	71.60	39.42	32.70	0.52
		0.6	89.38	39.86	49.60	0.08
At 5% level						
2	10	0	82.62	64.06	19.42	0.86
		0.6	87.84	63.92	24.22	0.30
2	10	0	88.88	86.80	4.70	2.70
		0.6	92.76	86.14	7.70	1.08
4	10	0	75.36	25.66	51.06	1.36
		0.6	91.74	25.44	66.76	0.46
4	20	0	84.70	66.04	19.68	1.02
		0.6	94.84	65.70	29.38	0.24

One reason for the better performance of Rohlf's test in this comparison is that it exploits the information concerning ρ. Wilks' test, on the other hand, is a general test for any correlation structure. A more appropriate comparison, therefore, could be between Rohlf's test and a version of Wilks' test which did benefit from knowledge of the correlation structure. An approximation to this is to take the case ρ=0, and use

$$\sum_{j=1}^{p} (x_{ij}-\bar{x}_j)^2/\sigma_j^2 \sim X_p^2, \quad i=1,\ldots,n$$

where $x_1,\ldots,x_n$ are independent vectors from $N_p(\mu,\Sigma)$ with $=\text{diag}(\sigma_1^2,\ldots,\sigma_p^2)$. Since σ_j^2 is unknown, it is replaced by its estimate s_j^2, giving the statistic

$$\sum_{j=1}^{p} (x_{ij}-\bar{x}_j)^2/s_j^2$$

This is the statistic suggested by Healy (1968) for probability plotting. Although it does not itself follow the X^2 distribution, this distribution can be kept as an approximation. For this outlier detecting application, the maximum value of the statistic over choices of x_i from $(x_1,\ldots,x_n)$ is taken, and Bonferroni % points from X_p^2 are used for significance testing. A power comparison between Rohlf's test statistic and this new test is shown in Table 5.3.5. The difference between the powers is, as expected, much less than was the case in the comparison with the standard Wilks' test statistic. This comparison could be refined further, by constructing simulated percentage points for the new test statistic instead of taking Bonferroni approximations to the critical values of an approximate distribution, but the effort does not seem to be worthwhile since the case of known ρ is seldom realistic. Instead, the use of unknown ρ will be looked at further.

Table 5.3.5 Power comparison for detection of one outlier between Rohlf's test and a Wilks'-type test statistic: details as Table 5.3.4., n=20, p=4, ρ=0.

Level of significance	% of times outlier declared by Rohlf	Wilks	Rohlf-not Wilks	Wilks-not Rohlf
1%	71.60	59.82	14.18	2.40
2.5%	79.00	72.86	9.72	3.58
5%	84.70	81.30	6.86	3.46
10%	88.70	88.20	4.36	3.86

The assumption of equicorrelation structure was maintained for this analysis. Under this assumption, ρ was estimated in each sample and critical values for this value of ρ interpolated from those obtained earlier and shown in Table 5.3.3. This estimate of ρ was thus being used as if it were a known value, since no better alternative procedure presented itself. Curves for interpolating critical values were obtained by fitting polynomials in ρ to the critical values for each combination of n, p and significance level. Coefficients for quartic fits are shown in Table 5.3.6 and Figure 5.3.3 illustrates observed and fitted values for the cases

n=20, p=2, ρ=0,±.9,±.8,±.6,±.4,±.2,±.1,.3,.5
n=20, p=3, ρ=-.45,±.4,±.2,0,.6,.8,.9

Table 5.3.6 Coefficients for quartic in ρ fitted to simulated 100% percentage points (ρ^3 term omitted for n=20, p=3 and p=4 because tolerance limit in the regression exceeded).

			Coefficients x 10^{-5}					
p	n	α	constant	ρ	ρ^2	ρ^3	ρ^4	R^2
2	10	.01	67506	-1668	10752	2980	8331	0.9238
		.025	60510	624	11497	-737	10815	0.9480
		.05	54718	1167	9530	-1444	12782	0.9691
		.10	48297	604	5729	-574	15331	0.9699
2	20	.01	50594	1183	26587	153	-6228	0.9867
		.025	45154	204	16133	1174	7213	0.9852
		.05	40388	1031	11789	-857	9821	0.9838
		.10	35346	691	4267	-926	14735	0.9875
2	30	.01	44735	105	21013	4896	1107	0.9616
		.025	39269	-2411	7748	6341	16737	0.9733
		.05	34856	-1700	4147	3131	20832	0.9766
		.10	30298	-1551	-607	1366	21992	0.9741
3	20	.01	35405	-4083	27115	-	18324	0.9769
		.025	31524	-3536	20524	-	20348	0.9870
		.05	28061	-3933	17272	-	16596	0.9898
		.10	24266	-3378	15415	-	11826	0.9881
4	20	.01	27385	-3734	33882	-	13116	0.9992
		.025	24597	-3079	19611	-	24389	0.9977
		.05	22694	-2886	9848	-	28474	0.9956
		.10	19872	-3227	10462	-	19470	0.9904

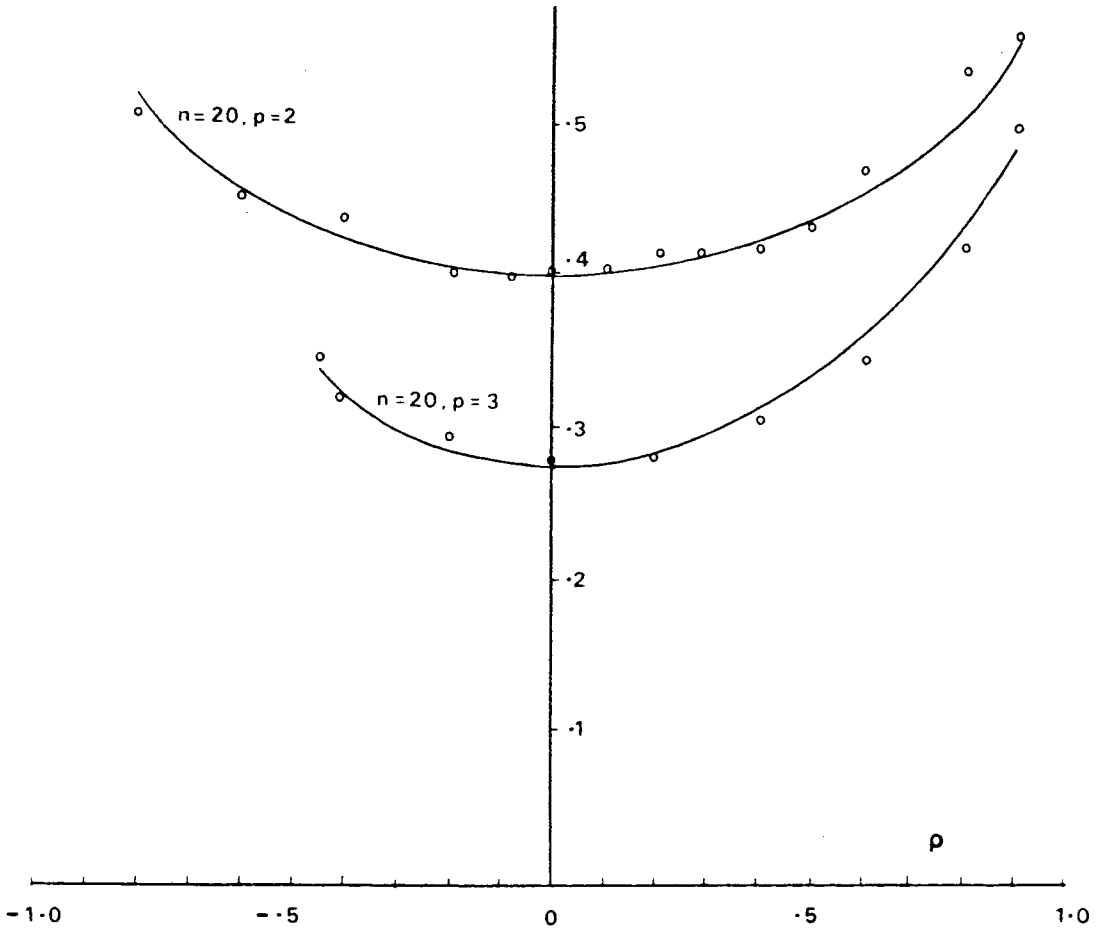


Figure 5.3.3 Simulated 5% critical values with quartic fits.

Firstly, it will be confirmed that it is necessary in practice to use different critical values for each value of ρ. It would not matter that ρ ought theoretically to be taken into account if the effect of, say, using the ρ=0 percentage points in a population with ρ=0.9 was only of the order of altering a 5% significance level into a 5.5% one. This is clearly not the case, however. The differences seen above for different values of ρ have large effects, as illustrated in the following power comparison between Rohlf's and Wilks' tests - (Table 5.3.6) - in which the percentage points for ρ=0 are applied to samples generated with ρ=0.6 as well as to the case ρ=0. The differences are so large that it would be unacceptable not to take ρ into account.

Table 5.3.7 Power comparison between Rohlf's and Wilks' tests in 1000 simulated samples with n=10, p=4 and ρ=0 or 0.6; outlier slippages= 4.5826 in each component (squared generalized distance 30); critical values of Rohlf's test are those simulated for ρ=0.

		% of times outliers are declared at			
		1% level		5% level	
ρ	no. of outliers	Rohlf	Wilks	Rohlf	Wilks
0	1	44.2	7.4	75.4	25.7
	2	18.8	2.3	53.9	10.0
0.6	1	88.9	7.3	96.6	25.4
	2	71.5	2.3	91.6	10.3

The performance of a test using these simulated percentage points was investigated by simulating further sets of data, estimating $\hat{\rho}$ assuming the equicorrelation model and using the polynomial in ρ to obtain critical values to be used as if this value of ρ were the true value. Samples were first generated under the null hypothesis (no outliers) to check the exceedance probabilities of this procedure. Results are shown in Table 5.3.8. It can be seen that this test is generally a little conservative. Some dependence on ρ is evident, with the most extreme values of the exceedence probabilities being associated with the extreme values of ρ. This might be expected, because the slopes of the curves in Figure 5.3.3 show that this is where the critical values are most sensitive to the value of ρ.

Since these results show that the size of this procedure is acceptably well controlled, it is reasonable to go on to power studies. These are again based on 2000 simulated samples at each combination of n, p, ρ and slippage vectors for one and two outliers. Wilks' test was computed (using simulated percentage points) on the same data for comparison to the above version of Rohlf's gap test.

Table 5.3.8 Exceedance probabilities obtained on estimating ρ.

			Observed exceedance probabilities at nominal level			
p	n	ρ	1%	2.5%	5%	10%
2	10	-.8	.008	.021	.050	.099
		-.2	.004	.015	.045	.089
		0	.007	.018	.038	.086
		.4	.006	.019	.038	.085
		.9	.010	.021	.042	.092
2	20	-.8	.010	.027	.045	.094
		-.2	.014	.029	.057	.120
		0	.009	.024	.044	.097
		.4	.005	.021	.046	.089
		.9	.016	.029	.062	.113
2	30	-.8	.013	.030	.050	.097
		-.2	.009	.023	.047	.091
		0	.009	.021	.047	.099
		.4	.008	.029	.055	.110
		.9	.007	.022	.044	.095
3	20	-.45	.017	.036	.067	.123
		-.2	.010	.024	.047	.095
		0	.011	.022	.045	.103
		.4	.010	.028	.055	.103
		.9	.006	.018	.055	.117

In the first instance, slippage vectors consisted of an equal quantity added to each dimension, the quantity being a function of ρ chosen so that the generalized distance of the slippage vector from the origin was constant over ρ. This meant that the power of Wilks' test was also constant over ρ. However, it was found that the power of Rohlf's test depended very strongly on ρ, being very low for large negative values of ρ, increasing steeply as ρ approaches zero and then increasing slightly as ρ increases through positive values. This behaviour is illustrated in Table 5.3.9 and Figure 5.3.4 for one outlier only.

Table 5.3.9 Power comparison for one outlier between Rohlf's and Wilks' tests in 2000 simulated samples for different n's, p's and ρ's; outlier slippages = $\sqrt{(1+(p-1)\rho).D^2}/\sqrt{p}$; (squared generalized distance $D^2=30$); critical values of Rohlf's test are interpolated from the quartic fit.

			% of times outliers are declared at the 5% level			
p	n	ρ	Rohlf	Wilks	Rohlf-not Wilks	Wilks-not Rohlf
2	10	-.9	10.85	62.80	1.00	52.95
		-.8	29.30	64.00	1.85	36.55
		-.6	56.40	64.50	6.10	14.20
		-.4	66.55	64.10	8.60	6.15
		-.2	72.60	62.85	12.85	3.10
		-.1	74.40	62.25	14.20	2.05
		0	76.15	62.15	15.70	1.70
		.1	76.35	62.70	14.90	1.25
		.2	76.85	63.70	14.25	1.10
		.3	78.20	62.65	16.50	.95
		.4	79.30	62.20	18.10	1.00
		.5	77.25	62.60	16.25	1.60
		.6	79.25	63.50	17.60	1.85
		.8	79.15	63.90	16.85	1.60
		.9	80.55	64.30	18.70	2.45
2	20	-.9	12.20	86.70	.30	74.80
		-.8	36.50	85.50	.35	49.65
		-.6	65.75	85.95	.85	21.05
		-.4	78.60	85.95	2.50	9.85
		-.2	82.90	85.80	3.65	6.55
		-.1	85.30	86.85	3.00	4.55
		0	85.10	85.30	3.95	4.15
		.1	85.40	85.40	4.00	4.00
		.2	88.85	86.45	4.55	2.15
		.3	87.25	85.50	4.35	2.60
		.4	88.70	86.20	5.30	2.80
		.5	90.40	87.85	5.30	2.75
		.6	87.95	84.80	6.05	2.90
		.8	89.60	85.70	6.20	2.30
		.9	89.55	86.20	6.55	3.20
3	20	-.45	7.85	76.65	.55	69.35
		-.4	22.65	77.25	.75	55.35
		-.2	72.10	76.25	5.40	9.55
		0	85.00	77.40	9.70	2.10
		.2	88.00	76.50	12.60	1.10
		.4	88.35	75.40	14.10	1.15
		.6	89.35	75.80	15.15	1.60
		.8	91.05	77.90	14.30	1.15
		.9	93.00	77.50	16.40	.90

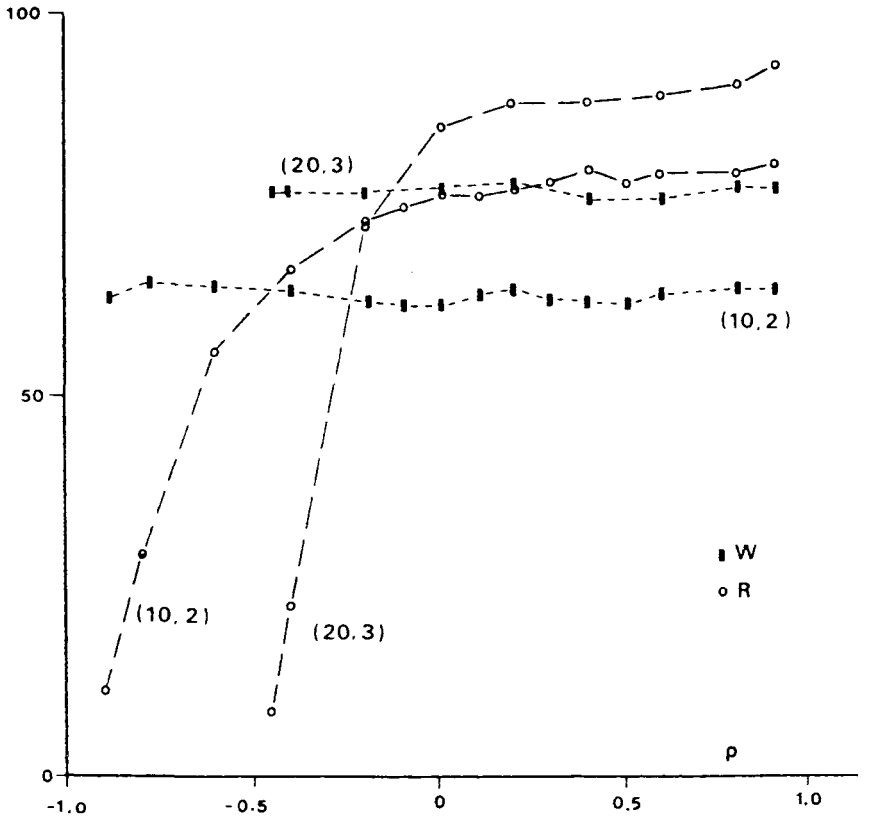


Figure 5.3.4 Powers of Wilks' (W) and Rohlf's (R) tests as a function of ρ, for (n,p)=(10,2) and (20,3).

The reason for this behaviour is easy to find. In the two-dimensional case, the generalized distance represented by the slippage vector (δ,δ) is easily shown to be

$$2\delta^2/(1+\rho)$$

if variances are unity. It follows that, for constant generalized distances, the value δ must be proportional to $\sqrt{(1+\rho)}$. This is a monotonically increasing function of ρ. Moreover its slope is a monotonically decreasing function of ρ. Consequently, the slippage is, as ρ increases, an increasing distance from the origin, with the rate of increase being greatest for the larger negative values of ρ. Since Rohlf's test uses Euclidean distance, this behaviour of the chosen slippage vector entirely agrees with the observed behaviour of the power function. As a check, it can be predicted that if slippages of $(\delta,-\delta)$ are used, then the desired α is proportional to $\sqrt{(1-\rho)}$: in this case, the opposite relation between the power of Rohlf's test and ρ should be observed, as indeed it is (Table 5.3.10).

Table 5.3.10 Power comparison for one outlier between Rohlf's and Wilks' tests in 2000 simulated samples for different n's, p's and ρ's; outlier slippages = $\sqrt{(1+(p-1)\rho).D^2}/\sqrt{p}$; (squared generalized distance $D^2=30$); critical values of Rohlf's test are interpolated from the quartic fit.

			% of times outliers are declared at the 5% level			
p	n	ρ	Rohlf	Wilks	Rohlf-not Wilks	Wilks-not Rohlf
2	10	-.9	80.30	63.45	18.90	2.05
		-.8	77.05	62.35	16.70	2.00
		-.6	79.95	63.40	18.05	1.50
		-.4	78.25	62.10	17.35	1.20
		-.2	77.10	62.05	16.15	1.10
		-.1	78.70	63.05	16.95	1.30
		0	76.70	63.70	14.85	1.85
		.1	75.45	62.30	15.55	2.40
		.2	74.60	63.95	13.35	2.70
		.3	70.75	62.85	11.75	3.85
		.4	67.40	62.75	9.95	5.30
		.5	61.10	63.20	7.20	9.30
		.6	53.45	63.60	4.90	15.05
		.8	29.20	64.80	1.75	37.35
		.9	12.10	62.05	.80	50.75
2	20	-.9	90.90	86.70	6.70	2.50
		-.8	90.70	87.65	5.65	2.60
		-.6	89.75	87.25	5.35	2.85
		-.4	88.55	87.00	4.50	2.95
		-.2	87.75	85.55	4.85	2.65
		-.1	87.95	86.35	4.30	2.70
		0	87.45	86.40	3.80	2.75
		.1	86.35	86.05	4.30	4.00
		.2	85.15	86.60	3.10	4.55
		.3	79.90	86.10	2.15	8.35
		.4	77.95	87.05	1.80	10.90
		.5	74.20	87.05	1.55	14.40
		.6	65.10	85.80	.50	21.20
		.8	35.35	85.55	.45	50.65
		.9	11.75	86.40	.25	74.90
2	30	-.9	91.55	90.30	3.90	2.65
		-.8	90.75	90.90	3.00	3.15
		-.6	91.45	91.35	2.75	2.65
		-.4	90.50	89.75	3.90	3.15
		-.2	89.35	89.65	3.20	3.50
		-.1	89.70	90.95	2.10	3.35
		0	89.10	91.50	1.50	3.90
		.1	88.15	91.20	2.25	5.30
		.2	85.50	90.25	1.95	6.70
		.3	82.45	91.00	.60	9.00
		.4	80.55	90.60	.80	9.40
		.5	76.80	90.70	.60	14.50
		.6	70.45	90.65	.40	20.60
		.8	39.15	90.20	.35	51.40
		.9	12.25	90.05	.10	77.90

5.4 Rohlf's test: Conclusion

The results of the previous sections have shown that it is not very difficult to modify Rohlf's original procedure to achieve controlled size, but that the power of the resulting test in relation to Wilks' test is very heavily dependent on the correlation structure of the data and is very much lower than the power of Wilks' test over a substantial region of parameter space. Strictly, this has been shown for the equicorrelation case but it is reasonable to infer that it applies more generally. It is, of course, a result which is entirely to be expected, with hindsight. The test was set up ignoring correlations between variables, but it is very optimistic to hope that the outcome would also be independent of correlations. The only question could be how strong would be the impact of correlations and the answer here is, very strong. It is concluded that Rohlf's test, in a form along the lines described, is not effective.

Can a useful test be obtained by retaining the structure of his procedure, but replacing the Euclidean distance by generalized distance so that correlations are taken into account? The difficulty lies in obtaining a robust estimate of the covariance matrix. The necessity for robust estimation will be demonstrated by first showing what happens if it is not employed.

Table 5.4.1 shows simulated 5% percentage points for Rohlf's statistic based on generalized distances, obtained from 8,000 simulations of samples of uncorrelated multivariate normal data at each combination of n and p. It also shows powers of the test using these percentage points, in comparison to Wilks' tests for one and two outliers. The striking result is that the power of Rohlf's test to detect any outliers when there are actually two outliers quickly becomes very low in comparison to the power of Wilks' test as the sample size

increases. Furthermore, only in the minority of cases where Rohlf's test declares any outliers does it declare that there are two outliers. Thus Rohlf's test is rather poor at detecting two outliers, even though in the situation simulated these are very distinct from the main body of the sample. Moreover, they have the same slippage, so are the only kind of multiple outliers in a cluster which can be detected by Rohlf's test; see § 5.2.4.

Table 5.4.1 Simulated powers of Rohlf's and Wilks' tests for one and two outliers at 5% level with p=5, in 8000 simulated samples. Slippage vector(s) 2.4495 in each component (squared generalized distance 30). R=Rohlf's test; W_1=Wilks' one-outlier test; W_2=Wilks' two-outlier test.

	sample size n			
	10	20	30	50
Critical value	.23695	.16890	.13040	.09238
One outlier: % of times declared by				
W_1	15.1	54.8	69.5	77.3
R	12.8	44.8	58.9	67.5
W_1 not R	8.0	14.5	13.8	12.2
R not W_1	5.7	4.5	3.2	2.5
Two outliers: % of times any outliers are declared by				
W_2	6.4	42.5	68.8	85.9
R	6.1	15.8	28.2	45.8
W_2 not R	4.8	29.2	42.0	40.6
R not W_2	4.5	2.7	1.3	0.5
R declares 2:	0	0.2	3.7	14.1

The problem can be demonstrated by considering the set

of data shown in Table 5.4.2 and Figure 5.4.1, consisting of ten points generated from the bivariate normal N(0,I), with slippages of (12,12) added to two of the points.

Table 5.4.2 Illustrative data (n=10,p=2) for problem of failure of Rohlf's test to declare two outliers.

Point	1	2	3	4	5	6	7	8	9	10
x_1	12.73	-0.45	0.24	-0.18	-0.51	-1.02	0.98	0.52	0.25	10.29
x_2	11.63	0.13	0.35	0.47	-1.20	-0.08	-0.17	-0.44	2.10	13.32

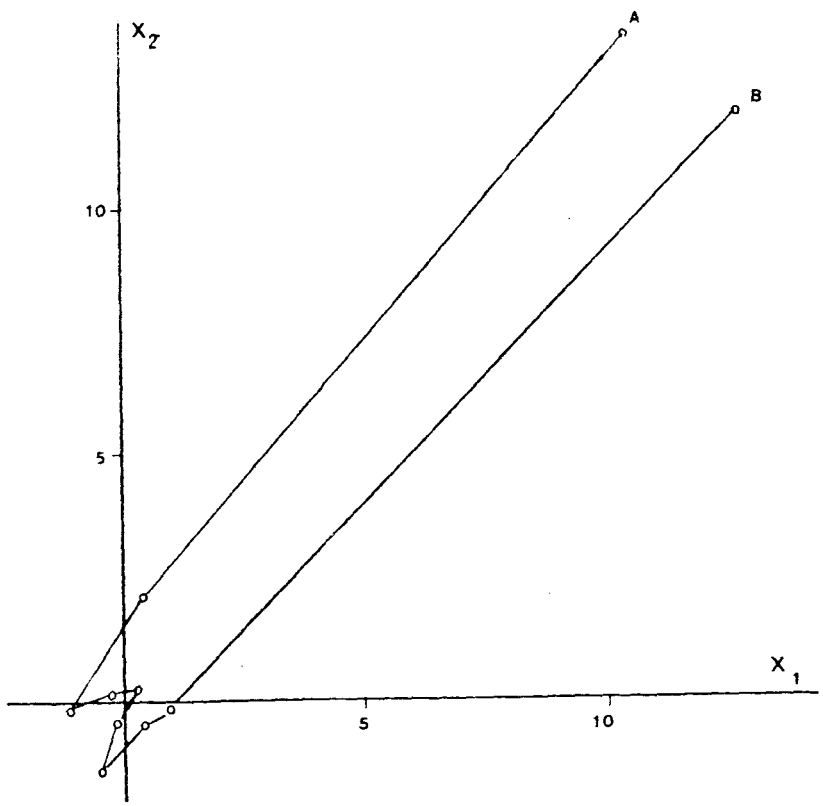


Figure 5.4.1 Scattergram of data of Table 5.4.2.

A reasonable test for outliers would surely be expected to declare that both the points to which the slippages have been added, A and B, are outliers from the main sample comprising the remaining 8 points. This is what happens with Wilks' test for two outliers. The value of the ratio

$$|A_{AB}|/\ |A| = 0.069$$

omitting these 2 points is significant well beyond the 1% level (simulated critical value = .19038, n=10, p=2). However, the MST using generalized distances (superimposed on Figure 5.4.1) does not link A and B, so it is not possible for both of them to be declared outliers. Rohlf's test statistic $Y_{max}/\bar{Y}$ takes the value of 0.503, which falls between the simulated critical values (from 40,000 samples) of 0.518 (5%) and 0.463 (10%), so there is not even very strong evidence to declare one outlier using Rohlf's test.

The reason for this behaviour is that the two large outliers induce a high correlation in the full sample of 10 points : r=0.967 in fact. The effect of this can be seen be comparing the standardized Euclidean distance

$$E_{ij}^2 = \frac{1}{2}\left\{\frac{(x_{i1}-x_{j1})^2}{s_1^2} + \frac{(x_{i2}-x_{j2})^2}{s_2^2}\right\}$$

to the generalized distance

$$D_{ij}^2 = \frac{1}{1-r^2}\left\{\frac{(x_{i1}-x_{j1})^2}{s_1^2} + \frac{(x_{i2}-x_{j2})^2}{s_2^2} - \frac{2r}{s_1 s_2}(x_{i1}-x_{j1})(x_{i2}-x_{j2})\right\}$$

$$= \frac{2}{1-r^2}\left\{E_{ij}^2 - \frac{r}{s_1 s_2}(x_{i1}-x_{j1})(x_{i2}-x_{j2})\right\}$$

Since r is large, the second term can have a big effect.

If the ith and jth points x_i and x_j both lie along the

direction from the origin to the vicinity of A and B, then $(x_{i1}-x_{j1})(x_{i2}-x_{j2})$ is positive, so that D^2_{ij} is much reduced in comparison to E^2_{ij}. If on the other hand, the line joining x_i to x_j is orthogonal to that direction, then $(x_{i1}-x_{j1})(x_{i2}-x_{j2})$ is negative, so D^2_{ij} is substantially increased over E^2_{ij}. This is what happens to the distance AB in this example.

It is obvious that this version of Rohlf's test fails to work in such cases because the geometry of the sample is distorted so severely by the outliers which the test is trying to detect. Wilks' two-outlier test can succeed because it also evaluates an undistorted statistic $|A_{AB}|$ for comparison to the distorted one $|A|$. To be effective in this situation, Rohlf's test also needs an undistorted statistic, which was the point of using robust estimators of dispersion in the case of the standardized Euclidean distance. The equivalent idea here would be to employ a robust estimator of the covariance matrix from which to construct generalized distances. However, a little reflection shows that this is not a direction worth pursuing. Calculating a robust covariance matrix is not such a simple matter as trimming, as employed earlier. It has been seen earlier that one way of doing it, as described in Chapter 2, section 2.7, actually reveals a great deal of information directly related to detecting outliers in the sample, in the form of the weights attached to each point. Consequently, there is no need to go on to Rohlf's test after making such a calculation; the MST would only be useful for its customary purpose of data display.

The conclusion of this chapter therefore remains negative towards Rohlf's test, which is not seen to be usable as an effective outlier test.

CHAPTER 6

UNION-INTERSECTION TESTING

6.1 Introduction

The union-intersection method of test construction is, along with likelihood ratio, one of the two principal approaches to hypothesis testing in the multivariate case. When likelihood ratio is applied to the slippage of the mean model used in outlier problems, it leads to the Wilks statistic via the two-stage procedure described earlier: the purpose of this chapter is to apply tests by the alternative methodology and to compare against the Wilks test to find possible advantages.

The following discussion relates to the usual slippage model in multivariate normal populations:

$$H_0: x_i \sim N_p(\mu,\Sigma), \qquad i=1,\ldots,n$$

vs

$$H_1: x_i \sim N_p(\mu,\Sigma), \qquad i\neq j,k,\ldots$$
$$x_j \sim N_p(\mu+a_j,\Sigma),$$
$$x_k \sim N_p(\mu+a_k,\Sigma),\ldots.$$

Wilks' test for declaring a set of points j,k,... of specified size as outliers requires the minimization over choices of j,k,... of the statistic

$$\frac{|A_{jk\ldots}|}{|A|} \qquad (6.1.1)$$

where the denominator and numerator are the determinants of respectively, the sum of squares and products (SSP) matrices of the sample before and after deletion of the points j,k,.... Now suppose that each one of the points j,k,... is considered as belonging to a group on its own, while the remaining points form another group. Then the within-group SSP matrix receives no contribution from the groups which consist of a single point and so is given by

the SSP matrix of the remainder of the sample, that is, $A_{jk\ldots}$. In other words, the Wilks statistic is the ratio of the determinants of the within-groups and total(A) SSP matrices. In the terminology of the multivariate analysis of variance (MANOVA) between these groups, the Wilks outlier statistic (6.1.1) is the minimum over choices of the set of potentially outlying points of Wilks' lamda statistic in the one-way MANOVA (Mardia, Kent and Bibby, 1979):

$$\Lambda = |W| \, / |T| \qquad (6.1.2)$$

where W is the within-groups, error or residual SSP (under H_1) and T is the total SSP, identical to A above. The usual definitions are

$$W = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)'$$

$$T = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})(x_{ij} - \bar{x})'$$

where there are m groups, with n_i observations $x_{i1},\ldots,x_{in_i}$ in group i with mean $\bar{x}_i$, the overall mean being $\bar{x}$. T is often written as B+W, where B is called the between-groups or hypothesis SSP matrix. B and W are often denoted by H and E respectively.

The Λ statistic (6.1.2) may be written as

$$\begin{aligned} |T^{-1}W| &= |B+W|^{-1} |W| \\ &= |I+W^{-1}B|^{-1} \\ &= \prod_{j=1}^{p} (1+\lambda_j)^{-1} \end{aligned}$$

where $\lambda_1,\ldots,\lambda_p$ are the eigenvalues of $W^{-1}B$. The distribution of Λ has already been discussed in Chapter 3.

The general idea of the union-intersection approach to

hypothesis testing in multivariate data (Roy, 1957) is that, given a problem involving the vector random variable x, it is projected onto the direction of the vector α and the equivalent univariate problem for the random variable $\alpha'x$ is solved. Then the test statistic for this univariate problem is maximized or minimized, as appropriate, over choices of α: this optimum value is the union-intersection test statistic for the multivariate problem. Applied to the one-way MANOVA, where H_0: $\mu_1=\mu_2=\ldots=\mu_m$ against H_1: not all equal, m being the number of groups, the vector random variable $x_i \sim N_p(\mu_i,\Sigma)$ is projected onto α to give the scalar random variable

$$y_i=\alpha'x_i \sim N_p(\alpha'\mu_i,\alpha'\Sigma\alpha)$$

and reduces the hypotheses to H_0: $\alpha'\mu_1=\alpha'\mu_2=\ldots=\alpha'\mu_m$ against not all equal, i.e., the univariate situation. The univariate analogue of H_0 would be tested using the analysis of variance statistic proportional to

$$\sum_{i=1}^{m} n_i(\bar{y}_i-\bar{y})^2 / \sum_{i=1}^{m} n_i s_i^2,$$

where the ith group has sample size n_i and variance s_i^2. The corresponding formula for the linear combination $y=\alpha'x$ is

$$z^2=\sum_{i=1}^{m} n_i(\alpha'(\bar{x}_i-\bar{x}))^2 / \sum_{i=1}^{m} n_i\alpha'S_i\alpha,$$

where $S_i=A_i/n_i$ is the within-group sample variance-covariance matrix. The rejection region is $\{z^2>c\}=R_\alpha$. The rejection region for the initial H_0 is then $R=\bigcup R_\alpha$, that is, at least one $H_0(\alpha)$ is rejected. Therefore if $\max_\alpha z^2>c$, then H_0 is rejected. It is straightforward to carry out the maximisation over α and show that $\lambda_i=\max_\alpha z^2$ is the largest eigenvalue of

$$\left(\sum_{i=1}^{m} n_i S_i\right)^{-1} \sum_{i=1}^{m} n_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' = W^{-1}B.$$

The test may be expressed equivalently in terms of the largest eigenvalue of other functions of B and W . For example, Morrison (1976) uses the greatest root of $(W+B)^{-1}B$, which may be written

$$\theta = \lambda_1/(1+\lambda_1).$$

Its distribution is denoted by $\theta(p,n-m,m-1)$, where n-m are the error degrees of freedom and m-1 are the hypothesis degrees of freedom. As with the Λ statistic, the distribution of $\theta(p,r,s)$ can sometimes be transformed exactly to F - in fact, this can be done for $\theta(1,r,s)$ and $\theta(p,r,1)$ (Mardia, Kent and Bibby, 1979). Otherwise, its distribution is complicated, and there appear to be no simple approximations.

If percentage points of this distribution can be obtained, Bonferroni bounds can be constructed for the use of this statistic as an outlier test statistic by taking its maximum value over all choices of sets of points tested as outliers. The construction of such bounds is discussed in the next section.

In general, the likelihood ratio and union-intersection methodologies give different statistics. In simple cases, however, they lead to the same statistic. One such situation is the comparison between two groups, since then the hypothesis matrix B (that is, the between-group SSP matrix) has rank one (since it is calculated from two group means), so that $W^{-1}B$ has just one non-zero eigenvalue, λ_1, and both Λ and θ are functions of this alone. As the two-group comparison underlies the test for a single outlier, the two approaches give the same result in that case. In testing for more than one

outlier, however, the analogous situation is the MANOVA between 3 or more groups, in which case the approaches do result in different tests. (This assumes that the outlying points are not all from the same distribution, that is, not all slippages are the same - otherwise, the problem is a comparison between two groups again.) The test derivation and comparisons in the following sections will therefore be on the basis of testing for two or more outliers, with unequal slippages in the mean.

6.2. Bonferroni bounds and their accuracy

Construction of a Bonferroni bound for the percentage point for an $\alpha\%$ level two-outlier test by the union-intersection method in a sample of size n requires the $\alpha/\binom{n}{2}\%$ percentage point of the greatest root distribution. Some tables of percentage points are already available (Pearson and Hartley, 1972) but the significance levels selected for tabulation usually do not match the values $\alpha/\binom{n}{2}$ required. Percentage points can also be read from charts (Heck, 1960), but with lower accuracy, while this graphical method is unsuitable for use in a computer simulation study. Therefore, the $\alpha/\binom{n}{2}\%$ percentage points required here were specially computed by using the series expansion of Khatri (1972). This gives the distribution function F(x) of the greatest root statistic as a polynomial in (1-x), and is published as a Fortran algorithm by Venables (1975).

Although the algorithm as given only returns the value of F(x) given x, in other words the level of significance, it can be modified so that it returns the coefficients of the powers of (1-x), enabling its use to determine percentage points. The series is

$$F(x)=x^{n_1 p/2}\sum_{k=0}^{mp}{}'_{\kappa}(n_1/2)_{\kappa}(1-x)^{k}C_{\kappa}(I_p)/k!$$

where n_1 and n_2 are the degrees of freedom in B nd W respectively, there are p dimensions and $m=(n_2-p-1)/2$, which must be an integer. The inner summation $\sum'_{\kappa}$ is over all partitions κ of k; a partition is a set of $r\leq p$ integers $k_1\geq k_2\geq\ldots\geq k_r>0$ whose total is k. The coefficient $(n_1/2)_{\kappa}$ is to be evaluated from the definitions

$$(a)_{\kappa}=\prod_{i=1}^{p}(a-(i-1)/2)_{k_i}$$

and
$$(b)_j=\begin{cases}1 & j=0\\ b(b+1)\ldots(b+j-1) & j\geq 1\end{cases}$$

and $C_{\kappa}(I_p)$ is the zonal polynomial

$$C_{\kappa}(I_p)=2^{2k}(p/2)_{\kappa}\prod_{1\leq i<j\leq r}(2k_i-2k_j-i+j)/\prod_{i=1}^{r}(2k_i+p-i)!$$

Writing the series as

$$F(x)=x^{n_1 p/2}\sum_{k=0}^{mp}G(k)(1-x)^{k}$$

it is possible to compute the coefficients

$$G(k)=\sum_{\kappa}{}'(n_1/2)_{\kappa}C_{\kappa}(I_p)/k!$$

which depend on n_1, n_2 and p but not on x, and hence iteratively solve

$$F(x)=\alpha$$

for x given α. Because F(x) is expressed as a polynomial in x, it is easy to calculate its gradient and carry out a Newton-Raphson iteration.

One difficulty with Khatri's method is that it is necessary for n-p, the difference between sample size and

dimensionality, to be an even number. In the cases where this restriction meant that the required sample size could not be used in the algorithm, percentage points were calculated for adjacent usable values and a smooth curve fitted to obtain the desired percentage point by interpolation.

Specifically, suppose the test for k outliers is being carried out in a sample of size n p-dimensional vectors. Then, carrying out the MANOVA (between k+1 groups) after specifying a particular set of k points as potential outliers, the SSP matrices B and W have Wishart distributions

$$B \sim W_p(\Sigma, k)$$
and
$$W \sim W_p(\Sigma, n-k-1)$$

independently of B. Consequently θ, the greatest eigenvalue of $(B+W)^{-1}B$ is the greatest root statistic θ with distribution

$$\theta(p, n-k-1, k)$$

(Mardia, Kent, and Bibby, 1979). It is usual to assume that n-k-1>p in defining this distribution: otherwise, the identity

$$\theta(p,r,s)=\theta(s,r+s-p,p)$$

gives the distribution

$$\theta(k, n-1-p, p).$$

However, for the small values of k met in outlier tests, the former situation will usually be the applicable one, for realistic sample sizes. The restriction of Khatri's method for computing this distribution is that n-k-p-2 must be even: for k=2 outliers, this means that n-p must be even. Suppose then that the percentage point

for n=20, p=5 is required. This was obtained by finding percentage points for n=17,19,21 and 23, all for p=5, then fitting a cubic to these four points to interpolate for n=20. In certain cases, a quadratic was used, if the four points corresponding to the ones in this example could not be obtained. For example, for p=5, the requirement that the error degrees of freedom be at least equal to the number of dimensions, $n-3 \geq p$, means that n must be at least 8. Therefore, to find the percentage point for n=10, it is not possible to fit through the points for n=7,9,11 and 13, since the first of these does not exist. In such cases, a quadratic was fitted through the remaining three points.

The Bonferroni bounds computed as above are shown in parentheses in Table 6.2.1 for dimensions 2 to 5, selected sample sizes from 10 to 100 and significance levels 0.01, 0.025, 0.05, and 0.10.

Since this test is for two outliers, it may be expected that, as is the case with Wilks' test, the Bonferroni bounds will not be very good. Consequently, simulated percentage points were constructed and are also shown in Table 6.2.1 for comparison. These are averages over 5 percentage points each based on 8000 sets of data for each combination of sample size and dimensionality, even though the lack of a simple updating formula in this eigenvalue problem means that the repeated computations are very much heavier than in the determinant calculations underlying Wilks' statistic.

The same sets of simulated data were used to estimate exceedence probabilities for the Bonferroni percentage points and thus indicate the importance of the numerical differences between simulated and Bonferroni points. The exceedence probabilities are shown in Table 6.2.2. It can be seen that the Bonferroni bounds are very conservative, as predicted. For a sample size of 30, the true size of

Table 6.2.1a Simulated percentage points for union-intersection two-outlier test statistic, with Bonferroni bounds in parentheses, α=0.01.

Sample size, n	Dimensions, p 2	3	4	5
10	.95504	.98121	.99420	.99902
	(.95861)	(.98288)	(.99471)	(.99932)
15	.84382	.89260	.92805	.95264
	(.85524)	(.90079)	(.93377)	(.95789)
20	.74189	.79568	.83708	.87142
	(.75772)	(.80892)	(.84936)	(.88245)
25	.65819	.71288	.75601	.79432
	(.67725)	(.72847)	(.77045)	(.80620)
30	.58850	.64201	.68537	.72336
	(.61175)	(.66108)	(.70232)	(.73818)
40	.48784	.53781	.57688	.61141
	(.51312)	(.55739)	(.59517)	(.62871)
50	.41918	.46282	.49759	.52758
	(.44287)	(.48235)	(.51642)	(.54696)
75	.31301	.34558	.37238	.39687
	(.33238)	(.36297)	(.38970)	(.41395)
100	.25186	.27820	.29976	.32020
	(.26784)	(.29268)	(.31452)	(.33443)

Table 6.2.1b Simulated percentage points for union-intersection two-outlier test statistic, with Bonferroni bounds in parentheses, α=0.025.

Sample size, n	Dimensions, p 2	3	4	5
10	.93742	.97264	.99066	.99818
	(.94367)	(.97536)	(.99161)	(.99839)
15	.81232	.86932	.90979	.93944
	(.82849)	(.88029)	(.91836)	(.94674)
20	.70767	.76471	.81092	.85018
	(.72763)	(.78316)	(.82732)	(.86376)
25	.62312	.67992	.72742	.76820
	(.64705)	(.70128)	(.74591)	(.78410)
30	.55597	.61147	.65574	.69681
	(.58261)	(.63408)	(.67723)	(.71487)
40	.45916	.50869	.55018	.58313
	(.48695)	(.53237)	(.57120)	(.60570)
50	.39335	.43793	.47333	.50433
	(.41955)	(.45968)	(.49435)	(.52547)
75	.29337	.32436	.35269	.37728
	(.31439)	(.34513)	(.37200)	(.39640)
100	.23563	.26144	.28275	.30331
	(.25329)	(.27813)	(.29996)	(.31989)

Table 6.2.1c Simulated percentage points for union-intersection two-outlier test statistic, with Bonferroni bounds in parentheses, α=0.05.

Sample size, n	Dimensions, p 2	3	4	5
10	.91964	.96345	.98663	.99700
	(.92884)	(.96749)	(.98811)	(.99721)
15	.78401	.84832	.89335	.92708
	(.80493)	(.86190)	(.90427)	(.93631)
20	.67579	.73891	.78930	.83087
	(.70232)	(.76126)	(.80840)	(.84754)
25	.59315	.65378	.70243	.74459
	(.62075)	(.67879)	(.72547)	(.76558)
30	.52923	.58544	.63262	.67402
	(.58264)	(.61212)	(.65672)	(.69573)
40	.43542	.48619	.52714	.56188
	(.46616)	(.51241)	(.55200)	(.58727)
50	.37305	.41620	.45228	.48436
	(.40121)	(.44180)	(.47688)	(.50842)
75	.27673	.30840	.33636	.36060
	(.30043)	(.33125)	(.35820)	(.38269)
100	.22164	.24819	.26955	.28989
	(.24208)	(.26688)	(.28869)	(.30861)

Table 6.2.1d Simulated percentage points for union-intersection two-outlier test statistic, with Bonferroni bounds in parentheses, α=0.10

Sample size, n	Dimensions, p 2	3	4	5
10	.89620	.95013	.98060	.99510
	(.91004)	(.95701)	(.98312)	(.99537)
15	.75059	.82137	.87211	.91120
	(.77804)	(.84057)	(.88763)	(.92374)
20	.64149	.70935	.76358	.80760
	(.67455)	(.73700)	(.78725)	(.82924)
25	.56025	.62401	.67447	.71935
	(.59564)	(.65448)	(.70324)	(.74529)
30	.49804	.55703	.60456	.64859
	(.53409)	(.58873)	(.63476)	(.67513)
40	.40952	.46013	.50238	.53820
	(.44447)	(.49150)	(.53182)	(.56780)
50	.35000	.39321	.43012	.46188
	(.38224)	(.42323)	(.45870)	(.49063)
75	.25924	.29084	.31860	.34362
	(.28614)	(.31702)	(.34403)	(.36859)
100	.20800	.23349	.25552	.27598
	(.23066)	(.25542)	(.27719)	(.29707)

Table 6.2.2a Simulated null probability of obtaining a value of union-intersection two-outlier test statistic more than the Bonferroni approximation at α=0.01.

Sample size, n	Dimensions, p 2	3	4	5
10	.0079	.0082	.0086	.0060
15	.0071	.0069	.0073	.0066
20	.0066	.0065	.0063	.0060
25	.0058	.0061	.0052	.0062
30	.0047	.0055	.0055	.0056
40	.0043	.0056	.0052	.0051
50	.0044	.0050	.0044	.0050
75	.0047	.0047	.0048	.0043
100	.0042	.0044	.0045	.0046

Table 6.2.2b Simulated null probability of obtaining a value of union-intersection two-outlier test statistic more than the Bonferroni approximation at α=0.025.

Sample size, n	Dimensions, p 2	3	4	5
10	.0186	.0200	.0210	.0210
15	.0159	.0167	.0168	.0160
20	.0154	.0150	.0143	.0145
25	.0135	.0138	.0137	.0147
30	.0120	.0129	.0133	.0141
40	.0104	.0117	.0128	.0125
50	.0098	.0116	.0114	.0112
75	.0098	.0099	.0102	.0102
100	.0093	.0101	.0100	.0103

Table 6.2.2c Simulated null probability of obtaining a value of union-intersection two-outlier test statistic more than the Bonferroni approximation at α=0.05.

Sample size, n	Dimensions, p 2	3	4	5
10	.0355	.0388	.0392	.0453
15	.0303	.0319	.0324	.0303
20	.0284	.0278	.0273	.0276
25	.0257	.0256	.0267	.0273
30	.0229	.0247	.0244	.0262
40	.0202	.0226	.0240	.0218
50	.0188	.0223	.0223	.0215
75	.0181	.0187	.0194	.0198
100	.0177	.0186	.0187	.0190

Table 6.2.2d Simulated null probability of obtaining a value of union-intersection two-outlier test statistic more than the Bonferroni approximation at α=0.10.

Sample size, n	Dimensions, p 2	3	4	5
10	.0689	.0728	.0782	.0922
15	.0573	.0617	.0613	.0600
20	.0515	.0528	.0530	.0527
25	.0472	.0493	.0492	.0491
30	.0441	.0454	.0470	.0488
40	.0391	.0429	.0439	.0418
50	.0370	.0404	.0403	.0408
75	.0344	.0343	.0365	.0369
100	.0318	.0349	.0338	.0346

the test using the Bonferroni bound is only about half the nominal size. For a sample size of 15, the true size is about two-thirds of the nominal size. The true size decreases as sample size increases and increases slightly as dimensionality increases.

6.3 Comparison between likelihood ratio and union-intersection tests

There is no general answer as to which form of testing is better in MANOVA - otherwise one of the methods would have been discarded. There have been several comparative studies from the point of view of power and the main features of their results can be found in general texts such as Mardia, Kent and Bibby (1979), Morrison (1976) and Chatfield and Collins (1980). One result often quoted is that the union-intersection test is much the more powerful if differences between the groups in the MANOVA are nearly one-dimensional: that is, if the population means lie nearly on a straight line. In the outlier problem with slippage of the mean as the model, this situation would represent slippages of different magnitudes in the same direction. This is not an implausible structure. One example could be of data referring to a group of animals contaminated by one or two individuals at different stages of growth from the rest: if the rates of growth of different parts of the body are equal, these outlying individuals would be drawn from populations with means differing only by scale factors from the mean of the main group and so all the slippages would be in the same direction.

Given this expectation of a particular way in which the test statistic based on union-intersection may surpass the Wilks statistic in performance, the simulated data in the power studies described here include data constructed for models with slippages of this kind.

The power studies require the simulation of data under the alternative hypothesis of slippage of the mean by unequal amounts in two sample members. Two basic situations were considered. In one, the slippages were along the same axis through the origin (the mean of the generating distribution). In the second, the directions

of the two slippages were at right angles to each other. Data were generated from the distribution $N_p(0,I)$ and outliers (slippage of the mean) were simulated by adding a suitable quantity to the first and last members of each sample. For slippages along the same axis, the quantity added was plus or minus the vector with each component equal to $d/\sqrt{p}$, where d is the desired generalized distance from the origin and p the dimensionality. Three combinations were taken: first, the slippages were both with squared generalized distance equal to 30 and in the same direction; second, one squared distance was 30 and the other 15, again in the same direction; thirdly, one squared distance was 30 and the other was 15 in the opposite direction from the origin (that is, one vector was positive in each component and the other was negative). For the case of slippages at right angles to each other, both were given squared generalized distance of 30 from the origin. The first slippage had the equal component of $d/\sqrt{p}$ in each dimension. To be orthogonal to a vector x in the metric of Σ, a vector y must satisfy

$$y'\Sigma^{-1}x=0$$

which gives

$$y'x=0$$

for $\Sigma=I$. Since $x \propto 1$, this means that y should be a vector whose components add up to zero and are scaled to give the correct distance. Suitable vectors are as follows:

$$p=2 \quad y=d(1,-1)/\sqrt{2}$$

$$p=3 \quad y=d(1,-2,1)'/\sqrt{6}$$

$$p=4 \quad y=d(1,-1,1,-1)'/2$$

$$p=5 \quad y=d(1,-1,0,-1,1)'/2$$

where $d^2=30$.

In each simulated set of data, the Wilks and

union-intersection statistics were computed and compared to the simulated percentage points. The simulated percentage points (obtained for Wilks' test in Chapter 3.) provide the appropriate basis for comparison, because the conservative Bonferroni bounds are a poor approximation for either test. The results are displayed in Table 6.3.1.

Table 6.3.1 Comparison of powers of Wilks' test and the union-intersection test for two outliers: proportion of times that two outliers are declared by the union-intersection test (U), Wilks' test (W), union-intersection but not Wilks' ($U\bar{W}$) and Wilks' but not union-intersection ($W\bar{U}$).

In the following tables the squared generalized distances corresponding to the two outliers, the first and last observations in a sample, are expressed as dist1 and dist2. For the one outlier slip1=$\sqrt{dist1}/\sqrt{p}$ is the slippage added to all components of 1st observation for all dimensions, where dist1=30. For the other outlier in the same direction, slipn=$\sqrt{dist2}/\sqrt{p}$ is added to all components of the last observation for all dimensions. Dist2=15 or 30. For the opposite direction slipn=$-\sqrt{dist2}/\sqrt{p}$ is added to all components of last observation for all dimensions and dist2=15. When the outliers are at right angles to each other, the slippages added to the components of the last observation vary, slippage added to each i-component is denoted by $slipn_i$, so for

p=2, $slipn_1=\sqrt{dist2}/\sqrt{p}$, $slipn_2=-slipn_1$

p=3, $slipn_1=slipn_3=\sqrt{dist2}/\sqrt{6}$, $slipn_2=-2(slipn_1)$

p=4, $slipn_1=slipn_3=\sqrt{dist2}/2$, $slipn_2=slipn_4=-slipn_1$

p=5, $slipn_1=slipn_5=\sqrt{dist2}/2$, $slipn_2=slipn_4=-slipn_1$, $slipn_3=0$

for dist2=30.

p=2 , (i) dist1=dist2=30, slip1=slipn=3.87298
(same direction)

α	n	U	W	$U\bar{W}$	$W\bar{U}$
.01	10	13.23	12.43	3.00	2.20
	20	70.66	64.83	7.31	1.48
	30	89.23	85.86	4.13	0.76
	50	95.95	94.55	1.64	0.24
.025	10	26.06	24.91	4.65	3.50
	20	82.68	78.96	5.01	1.30
	30	94.51	92.74	2.15	0.38
	50	97.71	97.14	0.74	0.16
.05	10	40.61	38.41	5.96	3.76
	20	90.24	87.76	3.48	1.00
	30	96.83	95.86	1.29	0.33
	50	98.76	98.35	0.55	0.14
.10	10	59.06	55.28	7.10	3.31
	20	94.96	93.78	1.76	0.58
	30	98.50	98.11	0.60	0.21
	50	99.34	99.21	0.24	0.11

p=2, (ii) dist1=30, dist2=15, slip1=3.87298, slipn=2.73861
(same direction)

α	n	U	W	$U\bar{W}$	$W\bar{U}$
.01	10	8.00	7.70	1.86	1.56
	20	48.24	43.68	6.93	2.36
	30	69.98	66.84	5.45	2.31
	50	82.75	80.50	3.71	1.46
.025	10	17.15	16.28	3.61	2.74
	20	63.14	59.55	6.33	2.74
	30	80.99	78.79	4.05	1.85
	50	89.26	87.89	2.74	1.36
.05	10	28.36	27.51	4.40	3.55
	20	75.01	72.23	5.23	2.38
	30	87.61	85.99	3.04	1.41
	50	92.75	92.06	1.64	0.95
.10	10	43.29	42.29	5.25	4.25
	20	84.64	83.01	3.31	1.69
	30	92.91	92.44	1.54	1.06
	50	95.84	95.34	1.04	0.54

p=2 , (iii) dist1=30, dist2=15, slip1=3.87298, slipn=-2.73861
(opposite direction)

α	n	U	W	$U\bar{W}$	$W\bar{U}$
.01	10	12.44	11.64	2.93	2.13
	20	56.51	51.41	7.61	2.51
	30	75.10	71.95	5.26	2.11
	50	85.43	82.83	3.75	1.15
.025	10	24.19	23.48	4.03	3.31
	20	71.53	68.18	5.90	2.55
	30	84.83	82.89	3.53	1.59
	50	91.19	89.78	2.46	1.05
.05	10	38.30	36.71	5.45	3.86
	20	82.03	79.81	4.10	1.89
	30	90.64	89.20	2.39	0.95
	50	94.68	93.89	1.55	0.76
.10	10	55.69	54.00	6.06	4.38
	20	89.94	88.86	2.66	1.59
	30	95.03	94.40	1.31	0.69
	50	97.10	96.75	0.90	0.55

p=2, (iv) dist1=dist2=30, slip1=$slipn_1$=3.87298,
$slipn_2$=-3.87298
(right angles)

α	n	U	W	$U\bar{W}$	$W\bar{U}$
.01	10	29.26	61.33	8.38	32.90
	20	77.36	94.08	0.39	17.10
	30	88.79	97.20	0.11	8.53
	50	93.79	98.36	0.10	4.68
.025	10	49.43	77.90	1.00	29.48
	20	89.26	97.46	0.13	8.33
	30	94.76	98.55	0.09	3.88
	50	97.31	99.35	0.05	2.09
.05	10	67.20	87.56	0.66	21.03
	20	95.14	98.86	0.15	3.88
	30	97.54	99.29	0.13	1.88
	50	98.75	99.68	0.01	0.94
.10	10	82.58	94.51	0.43	12.36
	20	98.24	99.46	0.03	1.25
	30	99.11	99.73	0.03	0.64
	50	99.55	99.81	0.04	0.30

p=3 , (i) dist1=dist2=30, slip1=slipn=3.16228
(same direction)

α	n	U	W	$U\bar{W}$	$W\bar{U}$
.01	10	5.18	5.14	1.54	1.50
	20	51.38	44.94	9.03	2.59
	30	78.36	70.41	9.50	1.55
	50	91.78	87.35	4.90	0.48
.025	10	10.75	10.65	2.90	2.80
	20	67.58	60.19	9.75	2.36
	30	87.36	81.36	7.05	1.05
	50	95.31	92.90	2.84	0.43
.05	10	19.44	18.94	4.39	3.89
	20	78.28	72.81	7.68	2.21
	30	92.71	88.64	4.75	0.68
	50	97.55	96.05	1.89	0.39
.10	10	32.33	31.54	5.89	5.10
	20	87.24	83.53	5.30	1.59
	30	96.09	94.35	2.28	0.54
	50	99.00	98.01	1.10	0.18

p=3, (ii) dist1=30, dist2=15, slip1=3.16228, slipn=2.23607
(same direction)

α	n	U	W	$U\bar{W}$	$W\bar{U}$
.01	10	3.55	3.56	1.24	1.25
	20	30.86	27.69	6.68	3.50
	30	54.66	48.79	9.06	3.19
	50	73.20	67.79	7.74	2.33
.025	10	7.96	8.41	1.91	2.36
	20	46.04	42.41	7.75	4.13
	30	68.21	62.46	8.83	3.08
	50	81.49	78.30	5.24	2.05
.05	10	14.23	14.93	3.14	3.84
	20	59.13	55.83	8.05	4.75
	30	76.88	73.61	6.06	2.80
	50	87.83	85.60	4.14	1.91
.10	10	25.99	25.39	5.14	4.54
	20	72.81	69.63	6.78	3.59
	30	85.55	82.99	4.66	2.10
	50	92.69	91.79	2.51	1.60

p=3, (iii) dist1=30, dist2=15, slip1=3.16228, slipn=-2.23607
(opposite direction)

α	n	U	W	$U\bar{W}$	$W\bar{U}$
.01	10	4.71	5.05	1.50	1.84
	20	38.15	30.70	7.51	4.06
	30	61.70	55.34	9.45	3.09
	50	76.55	71.11	7.64	2.20
.025	10	10.81	10.84	2.93	2.95
	20	54.89	50.15	8.85	4.11
	30	73.81	68.78	7.81	2.78
	50	84.80	82.14	4.94	2.28
.05	10	19.10	19.63	4.28	4.80
	20	67.36	63.81	7.55	4.00
	30	82.18	78.95	5.90	2.68
	50	90.26	88.74	3.19	1.66
.10	10	32.64	32.61	5.99	5.96
	20	80.00	77.03	6.38	3.40
	30	89.73	87.69	3.83	1.79
	50	94.58	93.41	2.05	0.89

p=3, (iv) dist1=dist2=30, slip1=3.16228,
slipn1=slipn3=2.23607, slipn2=-4.47214.
(right angles)

α	n	U	W	$U\bar{W}$	$W\bar{U}$
.01	10	13.94	32.20	1.54	19.80
	20	57.49	85.40	0.45	28.36
	30	77.99	93.01	0.28	15.30
	50	86.63	95.64	0.29	9.30
.025	10	26.08	49.89	1.69	25.50
	20	75.36	91.76	0.41	16.81
	30	88.51	96.63	0.30	8.41
	50	93.03	97.85	0.14	4.96
.05	10	41.60	65.63	1.70	25.73
	20	86.29	95.40	0.46	9.58
	30	93.90	98.24	0.13	4.46
	50	96.40	98.89	0.06	2.55
.10	10	60.36	79.74	1.69	21.06
	20	93.05	97.80	0.26	5.01
	30	97.40	99.15	0.11	1.86
	50	98.31	99.46	0.08	1.23

p=4 , (i) dist1=dist2=30, slip1=slipn=2.73861
(same direction)

α	n	U	W	$U\bar{W}$	$W\bar{U}$
.01	10	2.18	2.11	0.76	0.70
	20	36.89	30.41	9.81	3.34
	30	67.11	55.74	13.40	2.03
	50	86.41	78.78	8.54	0.90
.025	10	5.06	4.99	1.66	1.59
	20	53.18	43.48	12.69	2.99
	30	78.38	69.48	10.61	1.71
	50	91.80	87.29	5.38	0.86
.05	10	9.64	10.14	2.56	3.06
	20	65.10	56.40	11.89	3.19
	30	85.44	78.55	8.23	1.34
	50	95.09	92.14	3.58	0.63
.10	10	18.13	18.74	4.46	5.08
	20	77.30	74.11	7.15	3.96
	30	91.53	87.24	5.48	1.19
	50	97.33	95.48	2.23	0.38

p=4, (ii) dist1=30, dist2=15, slip1=2.73861, slipn=1.93649
(same direction)

α	n	U	W	$U\bar{W}$	$W\bar{U}$
.01	10	1.74	1.89	0.55	0.70
	20	20.96	18.76	6.08	3.88
	30	40.95	34.83	10.21	4.09
	50	63.21	55.81	10.45	3.05
.025	10	4.00	4.11	1.14	1.25
	20	32.89	29.35	8.19	4.65
	30	55.50	49.63	10.69	4.81
	50	73.36	69.28	7.63	3.54
.05	10	8.13	8.45	2.38	2.70
	20	44.44	41.28	8.73	5.56
	30	66.21	61.66	9.31	4.76
	50	80.99	78.06	6.23	3.30
.10	10	15.91	16.26	4.05	4.40
	20	58.85	55.43	9.00	5.58
	30	77.30	74.11	7.15	3.96
	50	87.66	85.83	4.09	2.25

p=4, (iii) dist1=30, dist2=15, slip1=2.73861, slipn=-1.93649

(opposite direction)

α	n	U	W	$U\bar{W}$	$W\bar{U}$
.01	10	2.18	2.51	0.73	0.70
	20	25.59	23.43	6.70	4.54
	30	48.48	41.59	10.93	4.04
	50	66.46	60.14	9.16	2.84
.025	10	5.50	5.71	1.66	1.88
	20	39.64	34.99	9.70	5.05
	30	62.49	56.30	10.46	4.28
	50	76.31	71.99	7.29	2.96
.05	10	9.79	11.03	2.33	3.56
	20	51.90	47.83	9.28	5.20
	30	72.86	68.10	8.88	4.11
	50	83.85	80.18	6.34	2.66
.10	10	18.65	19.91	4.45	5.71
	20	65.44	62.25	8.66	5.48
	30	82.86	79.85	6.29	3.28
	50	89.90	87.78	4.13	2.00

p=4, (iv) dist1=dist2=30, slip1=$slipn_1$=$slipn_3$=2.73861,
$slipn_2$=$slipn_4$=-2.73861

(right angles)

α	n	U	W	$U\bar{W}$	$W\bar{U}$
.01	10	6.15	13.89	1.16	8.90
	20	42.01	74.03	0.59	32.60
	30	64.19	86.35	0.61	22.78
	50	78.64	92.45	0.30	14.11
.025	10	13.50	25.11	1.89	13.50
	20	60.58	83.21	1.03	23.66
	30	78.16	92.30	0.41	14.55
	50	87.51	96.00	0.33	8.81
.05	10	23.54	38.90	2.58	17.94
	20	73.68	90.01	0.85	17.19
	30	86.61	95.56	0.45	9.40
	50	93.16	97.61	0.28	4.73
.10	10	39.15	56.04	2.85	19.74
	20	85.25	94.70	0.61	10.06
	30	93.44	97.73	0.24	4.53
	50	96.60	98.74	0.16	2.30

p=5, (i) dist1=dist2=30, slip1=slipn=2.44949
(same direction)

α	n	U	W	$U\bar{W}$	$W\bar{U}$
.01	10	1.29	1.41	0.39	0.51
	20	25.45	16.93	11.14	2.61
	30	55.94	43.49	15.11	2.66
	50	80.30	70.14	11.56	1.40
.025	10	2.85	3.39	0.71	1.25
	20	39.05	29.68	12.85	3.48
	30	69.21	57.56	14.31	2.66
	50	87.30	79.39	9.24	1.33
.05	10	63.38	65.38	2.00	2.20
	20	51.60	42.30	13.14	3.84
	30	78.56	68.70	12.04	2.18
	50	91.51	85.94	6.71	1.14
.10	10	12.33	13.21	2.91	3.80
	20	65.93	56.68	12.99	3.74
	30	87.18	79.76	9.01	1.60
	50	95.13	91.91	4.05	0.84

p=5, (ii) dist1=30, dist2=15, slip1=2.44949, slipn=1.73205
(same direction)

α	n	U	W	$U\bar{W}$	$W\bar{U}$
.01	10	1.15	1.30	0.34	0.49
	20	11.90	9.70	4.76	2.56
	30	31.33	26.19	9.44	4.30
	50	53.43	46.34	11.24	4.15
.025	10	2.96	2.96	0.90	0.90
	20	21.14	18.66	7.10	4.63
	30	43.85	39.08	10.08	5.30
	50	64.53	59.51	10.00	4.99
.05	10	5.91	5.86	1.75	1.70
	20	32.11	28.90	8.90	5.69
	30	55.61	51.11	10.11	5.61
	50	73.76	69.85	8.21	4.30
.10	10	11.50	12.16	2.91	3.58
	20	46.38	43.19	9.74	6.55
	30	68.00	64.60	9.06	5.66
	50	82.45	79.65	6.49	3.69

p=5, (iii) dist1=30, dist2=15, slip1=2.44949, slipn=-1.73205
(opposite direction)

α	n	U	W	$U\bar{W}$	$W\bar{U}$
.01	10	1.28	1.33	0.48	0.53
	20	17.15	12.84	7.31	3.00
	30	37.33	31.21	10.45	4.34
	50	58.25	50.96	11.09	3.80
.025	10	3.43	3.53	1.14	1.24
	20	28.33	24.26	9.09	5.03
	30	51.40	44.65	11.66	4.91
	50	77.16	73.59	7.85	4.28
.05	10	6.65	7.08	1.81	2.24
	20	40.00	35.71	10.39	6.10
	30	62.84	57.09	10.76	5.01
	50	77.16	73.59	7.85	4.28
.10	10	12.63	13.91	2.79	4.08
	20	54.51	50.76	10.24	6.49
	30	74.36	70.44	8.65	4.73
	50	85.18	83.03	5.60	3.45

p=5,(iv) dist1=dist2=30, slip1=2.44949, $slipn_1=slipn_5=2.73861$,
$slipn_2=slipn_4=-slipn_1$, $slipn_3=0$.
(right angles)

α	n	U	W	$U\bar{W}$	$W\bar{U}$
.01	10	2.99	5.10	0.75	2.86
	20	29.21	56.65	1.45	28.89
	30	50.80	78.50	0.95	28.65
	50	70.95	88.80	0.68	18.53
.025	10	6.93	11.38	1.66	6.11
	20	45.50	72.06	1.40	27.96
	30	66.43	87.00	0.89	21.46
	50	81.39	93.20	0.60	12.41
.05	10	13.25	20.43	2.43	9.60
	20	59.64	81.55	1.30	23.21
	30	78.14	91.75	0.71	14.33
	50	88.40	95.84	0.44	7.88
.10	10	23.35	34.05	3.53	14.23
	20	75.10	89.55	1.11	15.56
	30	87.78	95.59	0.55	8.36
	50	94.19	97.73	0.36	3.90

In the first situation, of slippages along the same axis, it can be seen that the power of the union-intersection test is higher than that of Wilks' test, in each of the three combinations of size and direction of slippage and for each dimensionality considered, except when the sample size is the smallest included in the study (n=10). For such a small sample size, both tests have very low power and the difference between them is small. McNemar's test confirms that there are statistically significant differences between the powers of the two tests, in most of the comparisons for n=10, and in most of these cases it is Wilks' test which is the more powerful (Table 6.3.2). This point will be returned to later.

For sample sizes of 20 and more, the advantage to the union-intersection test increases as the dimensionality increases, for a given sample size and combination of slippage sizes and directions (along the same axis). The biggest differences in the study occur for p=5 and sample sizes of 20 and 30, where the union-intersection test has a power up to 12 percentage points greater than Wilks' test.

As predicted, the union-intersection test does have higher power than Wilks' test for slippages along the same axis, and again as predicted, it can be seen that the opposite is true for slippages at right angles to each other. In this case the differences hold for all sample sizes considered and are much larger. For example, if Wilks' test had power of about 70% for slippages along the same axis, then the union-intersection test would have power not more than 75%, whereas if the union-intersection test had power of 70% for orthogonal slippages, then Wilks' test has power of around 90%. It also appears that it is rare for two outliers to be declared by the union-intersection test but not by Wilks' test.

Outlier sizes	Outlier directions	p	Level of significance 1% X^2_1	1% p	2.5% X^2_1	2.5% p	5% X^2_1	5% p	10% X^2_1	10% p
30,30	same	3	0.16	0.69	0.67	0.41	11.96	0.0005	22.43	<0.00001
		4	0.99	0.32	0.65	0.42	17.60	0.42	21.15	<0.00001
		5	6.67	0.010	58.34	<0.00001	3.72	0.054	46.67	<0.00001
30,15	same	3	0.02	0.90	18.74	0.00001	27.80	<0.00001	14.76	0.0001
		4	6.96	0.010	2.03	0.15	8.20	0.004	5.72	0.017
		5	10.55	0.001	0.00	1.00	0.26	0.61	26.86	<0.00001
30,15	oppo-	3	13.45	0.0003	0.03	0.85	12.03	0.0005	0.02	0.90
	site	4	25.11	<0.00001	4.99	0.026	104.0	<0.00001	62.49	<0.00001
		5	0.90	0.34	2.53	0.11	17.63	0.00003	62.44	<0.00001

Table 6.3.2. McNemar tests between powers of Wilks' and union-intersection tests

Since it appears that the union-intersection test holds the advantage when the slippages are along the same axis but not when they are at right angles to each other, it is interesting to see how big the angle between two slippage vectors can be before Wilks' test becomes the more powerful. This was investigated in a further simulated power study. The comparison was restricted to one case, with n=20, p=2 and both slippages having a squared generalized distance of 20 from the origin. One slippage vector was taken as $d/\sqrt{2}(1,1)'$ where $d^2=20$; the other was taken in varying positions on the perimeter of a circle of radius d, as in Figure 6.3.1. Because the vector OA is at

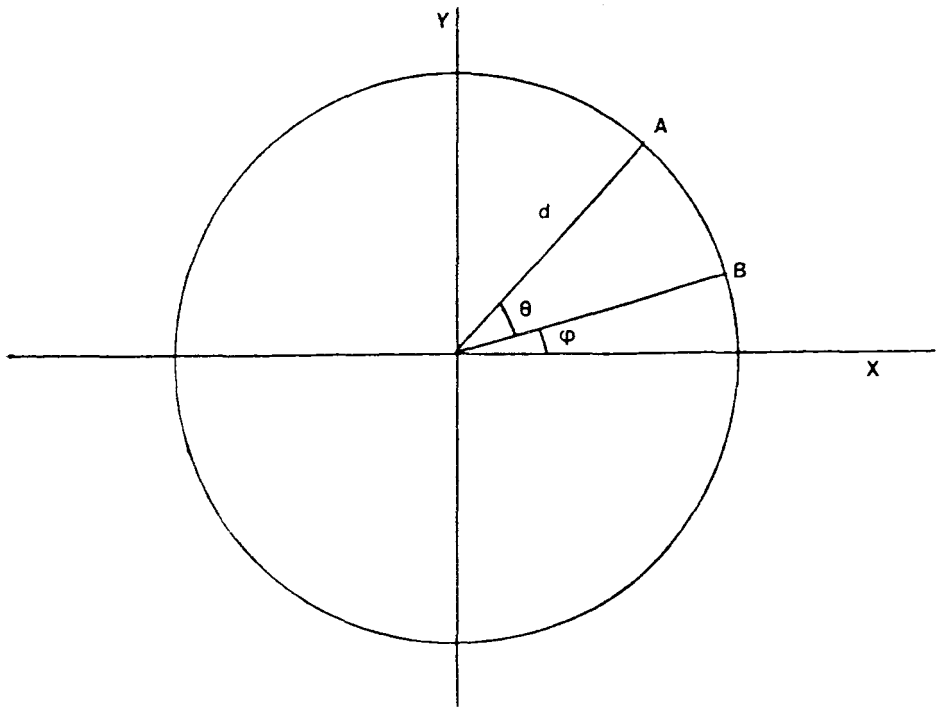


Figure 6.3.1 Positions of slippages A (fixed) and B (varying).

45° to the x axis, the angle θ between the slippage vectors is $\theta=(\pi/4)-\phi$.

Hence $$\cos\phi=\cos((\pi/4)-\theta)$$
$$=\cos(\pi/4)\cos\theta+\sin(\pi/4)\sin\theta$$

$$=(\cos\theta+\sin\theta)/\sqrt{2}$$

and

$$\begin{aligned}\sin\phi&=\sin((\pi/4)-\theta)\\&=\sin(\pi/4)\cos\theta-\cos(\pi/4)\sin\theta\\&=(\cos\theta-\sin\theta)/\sqrt{2}\end{aligned}$$

so that B should be taken as

$$d/\sqrt{2}(\cos\theta+\sin\theta,\cos\theta-\sin\theta)$$

in order for the angle between A and B to be θ.

Simulations were carried out for various values of θ, from 0° up to 180°, as shown in Table 6.3.3. Although all slippage vectors were of the same length, from the origin, the powers of both tests vary with θ. For example, Wilks' test must be more powerful when the two slippages are at right angles than when they are on the same axis, because in the latter case there is a greater probability that one extreme value in the rest of the sample can mask the two outlying points (that is, the two points with slippages added): see Figure 6.3.2. Against these varying powers, it can be seen in the Table and in Figure 6.3.3, that the relatively small power advantage to the union-intersection test applies until the angle between the slippage vectors is nearly 20°. Thereafter, the advantage to Wilks' test quickly becomes large. The union-intersection test becomes the more powerful again at about 160°.

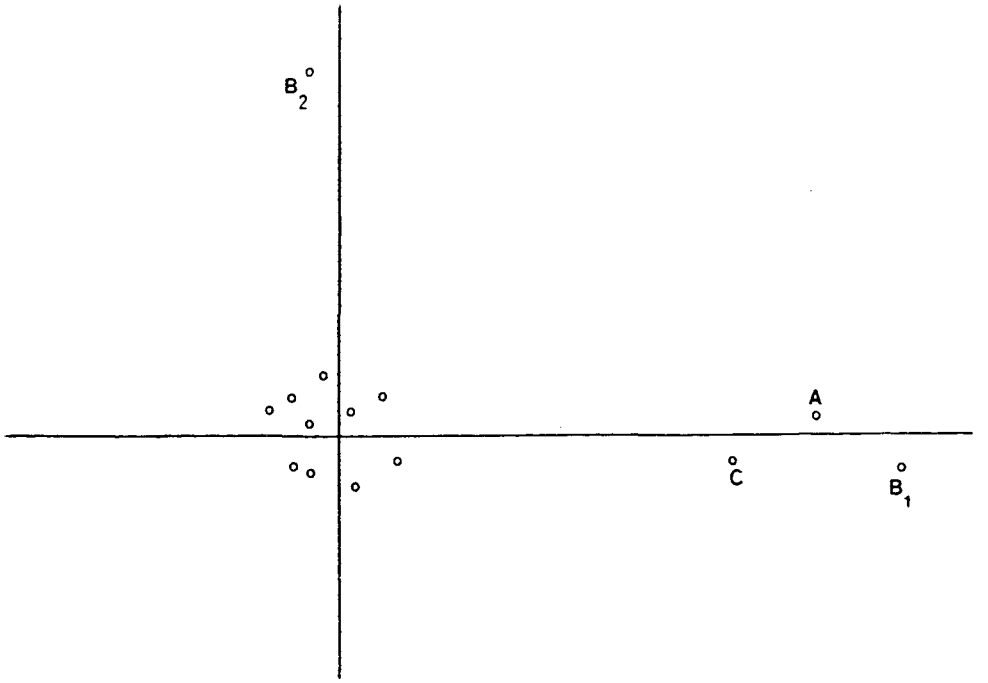


Figure 6.3.2 An extreme value from the null distribution at point C, can have a masking effect on the outlier pair AB_1 (same axis), whereas masking arises less easily for the pair AB_2 (orthogonal).

Table 6.3.3 Comparison of powers of Wilks' test and the union-intersection test for ϑ varying from 0° to 180°.

ϑ	1%		2.5%		5%		10%	
Angle	U	W	U	W	U	W	U	W
0	38.8	34.3	53.9	49.4	66.7	63.0	78.5	75.4
10	38.5	35.5	53.5	51.1	66.6	63.9	78.5	76.8
20	38.2	39.0	53.3	54.3	66.6	67.0	78.4	78.7
40	38.6	50.8	54.0	64.9	67.6	76.2	79.8	84.7
60	42.3	61.7	58.4	74.5	72.0	83.0	83.1	90.0
80	47.0	67.9	63.3	79.5	76.2	86.7	86.8	92.8
90	47.8	69.1	64.7	80.3	77.0	87.5	87.5	93.2
100	47.8	69.0	64.3	80.0	77.2	87.2	87.5	93.1
120	46.0	64.7	62.4	76.4	75.2	84.9	86.0	91.2
140	45.0	55.9	61.0	69.7	73.3	79.7	84.4	87.9
160	46.3	46.6	61.1	62.2	73.5	73.6	83.9	83.5
170	47.1	43.2	61.5	59.5	74.0	71.7	83.9	82.1
180	47.1	41.9	61.8	57.9	74.1	70.6	84.2	82.2

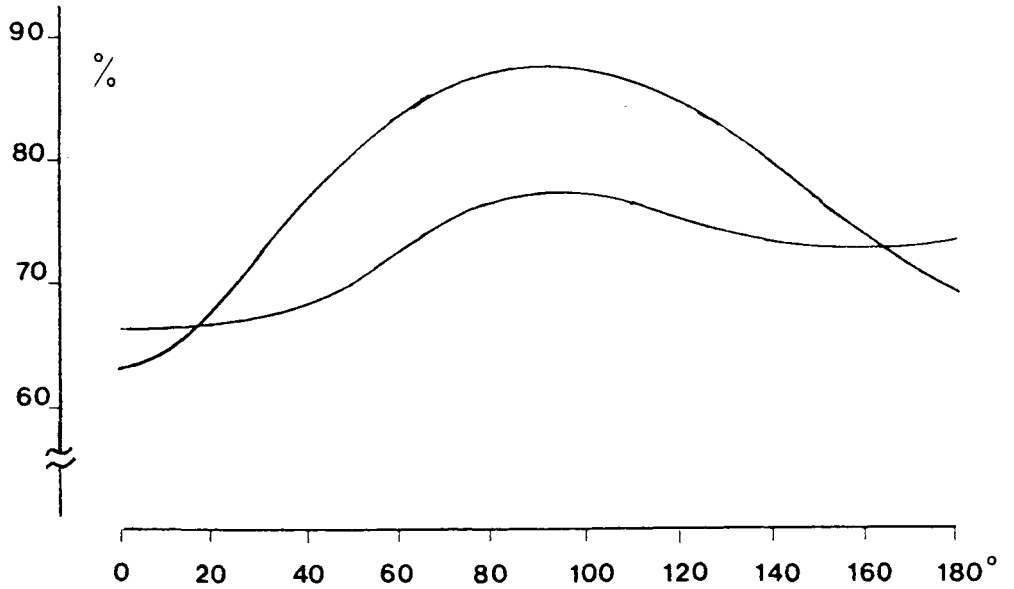


Figure 6.3.3 Powers between Wilks' test and the union-intersection test at the α=0.05 level for p=2, n=20 and for varying angles.

The fact that it appears that a quite close approach to collinearity is necessary in order for the union-intersection test to be more powerful probably explains why Wilks' test was earlier found to be often the more powerful even for slippages along the same axis, for a small sample size, n=10. Because the sample after selection of two outlying points then contains only 8 points, the position of the mean is much less accurately determined than for larger sample sizes. There is therefore extra variation, besides the variation of the positions of the generated outliers, and therefore a greater chance that there will be a substantial departure from collinearity.

To summarize the power comparison between the union-intersection and Wilks' tests, it is suggested that

there is no very good reason to prefer the former. It is true that in some circumstances, it may offer a large increase in power over Wilks' test. However, the difference between the tests is much larger when Wilks' test is superior than when the union-intersection test is superior; furthermore, Wilks' test is the more powerful for most relative positions of the two slippage vectors. Consequently, unless there is specific reason to expect that the slippages are nearly collinear so that the union-intersection test will be superior, it is recommended that Wilks' test be used, as it is generally more powerful.

One other disadvantage of the union-intersection test may be seen as follows. The test was introduced here along the lines of a MANOVA between 3 groups (with 1, 1 and n-2 members). Rejecting the null hypothesis in the MANOVA implies that not all three means (of the populations from which these samples have been drawn) are equal. It does not say that all three differ from each other: it could be that two are the same and the third differs. In this sense, the union-intersection two-outlier test might not seem to really be a test for two outliers. However, if one of the two selected points is not distinct from the main body of the sample and the significant result is due to the more extreme isolation of the other point, then this is just the usual "swamping" effect as occurs with other multiple outlier tests, including Wilks'. A way to avoid swamping is offered by proper sequential application of outlier tests for different numbers of outliers. In principle this could be done for the union-intersection test as it was for Wilks' test. However, the heavy computations required seem to be prohibitive. The same applies to extending the union-intersection test to any other situation, such as for three outliers. The amount of computation will always be very much heavier than for equivalent uses of Wilks' test and seems unlikely to be worthwhile.

In concluding this chapter, it may be remarked that, although union-intersection construction of outlier test statistics seems not to have been considered before, Fieller (1976, 1989) has discussed a property of one-dimensional projection of a multivariate sample. He shows that the value of Wilks' statistic for observation x_k in the full p dimensions is the same as its value in the univariate projected sample, when the projection is onto the direction of the eigenvector of $S^{-1}(x_k-\bar{x})(x_k-\bar{x})'$. He calls this direction the outlier-displaying component for that observation. The direction corresponding to the maximum is the outlier-projecting component for the sample: it holds all the information on one outlier. There is no directly equivalent result for two outliers (unless they are assumed to have the same slippage), which is the case considered in this chapter.

CHAPTER 7

OUTLIER TESTS WITH STRUCTURED COVARIANCE MATRICES

7.1 Introduction

Methods of multivariate statistical analysis are often presented as having the purpose of examining the structure of data. Sometimes this may refer to the relationships within the cases (observational units), as in most cluster analyses, but more often it refers to relationships between the variables. Whenever the multivariate normal distribution applies to the variables, their inter-relationships will be described by the correlation matrix (which, together with the means and variances, is sufficient for the multinormal distribution). Investigating the structure of a set of multivariate data therefore often means investigating the structure of a correlation matrix, either as a purely exploratory analysis (no structure has been hypothesized beforehand) or as a confirmatory analysis (a particular structure has been proposed and is to be tested). Textbooks on multivariate analysis consequently include several methods for investigating and testing particular structures. One example is provided by factor analysis. The observed variates x are assumed to be linearly related to a set of unobserved variates y via a matrix of coefficients (factor loadings) Λ:

$$x=\Lambda y+\varepsilon$$

where ε is an error term. Applying this model means that the covariance matrix Σ of x is being represented as

$$\Sigma=\Lambda\Phi\Lambda'+\Psi$$

where Φ (often taken to be I) is the matrix of correlations between the unobserved factors y and $\Psi=\mathrm{var}(\varepsilon)$ is diagonal.

Although the factor analysis model is widely used,

there are other much simpler structures which may be presented for illustration. One is the equicorrelation structure

$$\Sigma = \sigma^2 R = \sigma^2 \begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \rho & \ldots & \rho \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \rho & \rho & \ldots & 1 \end{pmatrix}$$

This is important in the analysis of repeated measures data, since it is a sufficient condition for a straightforward analysis of variance (treating the data as a split-plot design) to give statistics with exact F-distributions. In a genuine split-plot design, this covariance structure arises because of randomization, but it is less likely to apply to measurements repeated in time, which cannot be randomized (Rowell and Walters, 1976).

A related form of correlation structure may be more appropriate when the variables form a time series, since a measurement may be expected to be most highly correlated with those closest to it in time, instead of equally correlated with all others. If a first-order autoregressive model

$$X_{t+\tau} = \rho^{|\tau|} X_t + \varepsilon_{t+\tau}$$

describes the process $\{X_t\}$, then the correlation between $X_{t+\tau}$ and X_t is $\rho^{|\tau|}$, so that if observations are made at times $t_1, t_2, \ldots, t_p$ with $\tau_i = t_{i+1} - t_i$, then the correlation matrix is

$$R = \begin{pmatrix} 1 & \rho^{\tau_1} & \rho^{\tau_1+\tau_2} & \dots & \rho^{\sum_{i=1}^{p-1}\tau_i} \\ & 1 & \rho^{\tau_2} & \dots & \rho^{\sum_{i=1}^{p-2}\tau_i} \\ & & \dots & & \dots \\ & & \dots & 1 & \rho^{\tau_{p-1}} \\ & & & & \rho^{\sum_{i=1}^{p-1}\tau_i} \\ & & \dots & & 1 \end{pmatrix}$$

Morrison (1976, 9.11) discusses this model and also processes which lead to the same covariance structure, including a Wiener stochastic process where

$$X_{t+1} = X_t + Y_{t+1}$$

and successive increments Y_t, Y_{t+1} are uncorrelated. This is a model for Brownian motion or any other process where the outcome can be thought of as the sum of independent contributions. A Guttman scaling model, where the variables can be thought of as occupying positions along a continuum, also gives rise to this correlation structure.

All of the structures mentioned above have practical importance in describing multivariate data structures. But of course none is applicable if there is no structure at all, in other words if the variates are independent. This is the simplest structure : R=I. In many applications it is obvious on sight that there are significant correlations, so independence would not be tested. However, there are plenty of fields of research where high correlations are never achieved, often because the measuring instruments are very imprecise, as may be the case with scales in many sociological and psychological applications. For this reason, the factor analysis procedure in the widely used SPSS package - which is commonly used with just this kind of data - includes a test for complete independence of the variates.

A more common variation of independence arises when only certain parts of the correlation matrix are zero, and particularly important is the case when a block, rather than specified elements, is zero. This happens as follows.

A frequent aim of statistical analyses is to demonstrate an association. In the context of two normally-distributed random variables, this may be done by testing the null hypothesis of zero correlation between them. It is not unusual to come across similar hypotheses in data of higher dimensionality, where, however, the correlations are zero between sets of variates rather than just pairs. For one example, in a multivariate regression analysis, the overall test of significance examines whether all the dependent variables are independent of all the predictors. For another, in canonical correlation analysis, linear combinations are found within each set of variates so that the correlation between the two combinations, one from each set, is maximized; further combinations may then be derived with maximal correlations subject to orthogonality to preceding linear combinations. The test of significance for the first, maximized correlation, against the null hypothesis of zero correlation, tests whether all the variables of one set are uncorrelated with all those of the other set. Formally, suppose that the p-dimensional random vector x follows the multivariate normal distribution $N_p(\mu,\Sigma)$ and can be partitioned into two sets of variates of dimension p_1 and p_2 ($p_1+p_2=p$) with covariance matrix correspondingly partitioned as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

The test is for $\Sigma_{12}=\Sigma'_{21}=0$.

Morrison (1976, p.254) suggests that investigation of

block structures of this kind is interesting when the variates divide into two groups, so that the variates within each group are characterized by a common feature which distinguishes them from the other group. He gives an example where the measurements are the responses in a stress experiment, with one group consisting of physiological observations (such as blood pressure and skin conductivity) and the other consisting of behavioural measurements. The purpose would be to see if there were any connection between the physiological and behavioural data.

In this chapter, a slightly different problem is examined, in which it is assumed that the block structure holds and that there may be outliers distorting this picture. In testing for outliers in a simple random sample of data from this distribution, two alternatives will be considered. These are Wilks' test, which treats the data simply as p-dimensional random vectors with Σ unrestricted, and a test which will be developed specifically to incorporate the information that Σ has this block structure. This new test will, like Wilks' test, be based on maximum likelihood for slippage of the mean. The main purpose of this comparison, which will be made for a single outlier only, is to demonstrate the kind of increase in power that might be obtained by applying knowledge of the covariance structure.

7.2 A Wilks-type statistic when Σ has block structure

The derivation of a new test statistic, utilizing information on the block structure, employs the same calculation as leads to the usual Wilks statistic. Given independent p-dimensional random vectors x_i $(i=1,\ldots,n)$ and the hypotheses

$$H_0: x_j \sim N_p(\mu,\Sigma), \quad j=1,\ldots,n$$

and $$H_i: x_j \sim N_p(\mu,\Sigma), \quad j\neq i$$
$$x_i \sim N_p(\mu+a,\Sigma)$$

for known i, but unknown μ,a and Σ, then the likelihood ratio for H_0 against H_i reduces to $|A_i|/|A|$, as discussed before. Now consider how the likelihood changes when Σ has the block structure. Under H_0, the likelihood is proportional to

$$|\Sigma|^{-n/2}\exp\{(-1/2)\sum_{j=1}^{n}(x_j-\mu)'\Sigma^{-1}(x_j-\mu)\}$$
$$=|\Sigma_{11}|^{-n/2}|\Sigma_{22}|^{-n/2}\exp[(-1/2)\sum_{j=1}^{n}\{(x_{j1}-\mu_1)'\Sigma_{11}^{-1}(x_{j1}-\mu_1)$$
$$+(x_{j2}-\mu_2)'\Sigma_{22}^{-1}(x_{j2}-\mu_2)\}]$$

where $x_j=(x_{j1},x_{j2})'$ when partitioned in the same way as μ and Σ. By writing this as

$$|\Sigma_{11}|^{-/2}\exp\{(-1/2)\sum_{j=1}^{n}(x_{j1}-\mu_1)'\Sigma_{11}^{-1}(x_{j1}-\mu_1)\}$$
$$\cdot|\Sigma_{22}|^{-n/2}\exp\{(-1/2)\sum_{j=1}^{n}(x_{j2}-\mu_2)'\Sigma_{22}^{-1}(x_{j2}-\mu_2)\}$$

the likelihood is seen to be the product of two independent terms, each with the same structure as the simple likelihood for the unrestricted case. The same reasoning applies under H_i, and hence the likelihood ratio is

$$\frac{|A_{1i}|}{|A_1|}\cdot\frac{|A_{2i}|}{|A_2|}$$

where subscripts 1 and 2 denote SSP's for the first and second sets of variates.

Any one such ratio has as null distribution the product of the two independent betas

$$B\{(n-p_1-1)/2, p_1/2).B\{(n-p_2-1)/2, p_2/2)\}$$

which has p.d.f.

$$p(w) = \int_w^1 f(y)g(w/y)y^{-1}dy, \quad 0 \le w \le 1 \qquad (7.2.1)$$

where f and g are the density functions of the two beta variables. That is,

$$p(w) \propto w^{(n-p_2-3)/2} \int_w^1 y^{(p_2-p_1)/2-1}(1-y)^{p_1/2-1}(1-w/y)^{p_2/2-1}dy \qquad (7.2.2)$$

where $p_2 \ge p_1$ (otherwise, the roles of f and g in (7.2.1) should be reversed). The constant of proportionality is

$$\frac{\Gamma\{(n-1)/2)\}^2}{\Gamma\{(n-p_1-1)/2\}\Gamma\{(n-p_2-1)/2\}\Gamma(p_1/2)\Gamma(p_2/2)} \qquad (7.2.3)$$

The integral (7.2.2) does not seem to have a general solution, but in certain special cases it can be solved straightforwardly. Two cases will be taken here, for illustration. Firstly, suppose $p_1=p_2=2$. Then the integrand in (7.2.2) reduces to y^{-1} and the density becomes

$$p(w) = -\{(n-3)^2/4\}.w^{(n-5)/2}.\ln w, \quad 0 \le w \le 1$$

with distribution function

$$P(w) = w^{(n-3)/2}\{1-(n-3)/2.\ln w\}, \quad 0 \le w \le 1 \qquad (7.2.4)$$

Secondly suppose $p_1=2$, $p_2=4$. The integrand in (7.2.2) becomes (1-w/y) and the density is

$$p(w) = (n-3)^2(n-5)/8.w^{(n-7)/2}.\{1-w+w.\ln w\}$$

with distribution function

$$P(w)=w^{(n-5)/2}\,[(n-3)^2-(n-5)w\{n-1-(n-3)\ln w\}]/4 \qquad (7.2.5)$$

Knowing these distributions permits the construction of Bonferroni tests for outliers when the test statistic

$$\min_i \frac{|A_{1i}|}{|A_1|}\cdot\frac{|A_{2i}|}{|A_2|} \quad , \quad i=1,\ldots,n$$

is employed. As usual, for the α% Bonferroni bound the α/n% points of the distributions (7.2.4) and (7.2.5) are taken. This is easily done in a few program lines on a microcomputer, solving

$$P(w)= \alpha/n$$

by Newton-Raphson iteration (Appendix I). Some values are given in Table 7.2.1. These bounds can now be used in testing in

Table 7.2.1a Percentage points for Bonferroni test using block matrix structure, $p_1=p_2=2$:

	α			
n	0.01	0.025	0.05	0.10
10	.07150	.09572	.11968	.15007
20	.30841	.34744	.38056	.41722
30	.46134	.49713	.52630	.55748
50	.62614	.65346	.67509	.69762
100	.78474	.80105	.81370	.82665

Table 7.2.1b Percentage points for Bonferroni test using block matrix structure, $p_1=2$, $p_2=4$:

	α			
n	0.01	0.025	0.05	0.010
10	.02401	.03511	.04696	.06306
20	.22620	.25891	.28720	.31909
30	.38514	.41829	.44564	.47522
50	.56719	.59414	.61564	.63822
100	.74884	.76556	.77861	.79202

comparison to the ordinary Wilks tests for 4 dimensions (case $p_1=p_2=2$) and 6 dimensions (case $p_1=2$, $p_2=4$). For this purpose Wilks' Bonferroni table for one outlier (1963) was extended for p=6, n=10, 20, 30, 50, 100, as given in Table 7.2.2.

Table 7.2.2 Percentage points for Bonferroni test using Wilks' statistic for one outlier.

	α			
n	0.01	0.025	0.05	0.10
10	.00375	.00692	.01103	.01760
20	.19286	.22433	.25188	.28328
30	.36380	.39721	.42488	.45491
50	.55712	.58448	.60634	.62932
100	.74571	.76261	.77580	.78936

7.3 Power comparisons between the tests

The powers of the ordinary Wilks test and the new test incorporating the information on block structure were compared by simulation study. Multivariate normal samples, of the required dimension $p=p_1+p_2$ and of chosen size n, were generated with the block structure $\Sigma_{12}=0$ for chosen Σ_{11} and Σ_{22} and mean zero. A chosen vector was then added to the first member of the sample, representing a slippage of the mean. Both the Wilks statistic and the new statistic were computed and compared to their respective Bonferroni bounds at the same significance level. This procedure was repeated 8000 times for each combination of p and n, for each chosen Σ and slippage. For each combination of p and n, two types of slippage were used, and two distances with each type. The first type consisted of slippage in each dimension, represented by adding a multiple of the p-dimensional unit vector $\beta 1'$ to the first member of the sample. The multiplier β was chosen so that the squared generalized (Mahalanobis) distance of the size of slippage, in the metric of the

block Σ, namely

$$d^2=\beta^2 1'\Sigma^{-1}1$$

was equal to 30 or 15. The second type consisted of slippage affecting only the p_1 dimensions making up the first block of variables. This was represented by adding a vector λu, where u consisted of p_1 ones followed by p_2 zeros, with the multiplier chosen so that

$$\lambda^2 u'\Sigma^{-1}u=30 \text{ or } 15.$$

The computation of β and λ is given in Appendix II. The required values can be computed easily, because when Σ is block diagonal

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$$

then so is its inverse

$$\Sigma^{-1}= \begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix}$$

The inverse of a 2x2 block can be computed easily

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} (1-\rho^2)^{-1}$$

and the 4x4 block (in the case $p_1=2$, $p_2=4$) could also be inverted analytically, because the equicorrelation form

$$\begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}$$

was used; its inverse has elements

$$\frac{1+2\rho}{(1-\rho)(1+3\rho)}$$

along the diagonal and

$$\frac{-\rho}{(1-\rho)(1+3\rho)}$$

in all off-diagonal positions.

Tables 7.3.1 and 7.3.2 display results for Bonferroni levels of significance α=0.01, 0.025, 0.05 and 0.10 for the two types of slippage and the two distances, for cases $p_1=p_2=2$ (Table 7.3.1) and $p_1=2$, $p_2=4$ (Table 7.3.2). The following conclusions are clear:

(i) the modified test is the more likely to declare that an outlier is present, for both types and amounts of slippage, for all sample sizes;

(ii) the advantage to the modified test, which is very large for small samples, decreases as n increases; the advantage falls away faster as n increases for the second type of outlier than for the first type;

(iii) the modified test is less likely to declare an outlier with the second type of slippage than with the first type, especially in samples of size 10, 20 and 30. The performance of the unmodified test is the same for both types of outlier.

(iv) the degrees of difference between the tests are broadly similar for both sizes of slippages considered;

(v) Wilks' test declares an outlier in a small percentage of cases when the modified test does not.

The first two conclusions would be expected from the nature of the difference between the two tests. They differ only in that one uses the extra information that certain parameters of the model, namely the correlations in Σ_{12}, can be set equal to zero, instead of having to be estimated. As less information has therefore to be lost to the estimation of parameters, it follows that the modified test must be the more sensitive, on average, to

real differences. The increased sensitivity will be greatest when the sample is small, for then the efficiency of the estimation is lowest and so the imposition of the constraint has the greatest quantitative effect. Point (v) holds however because the argument of increased sensitivity applies only on the average and not to each particular sample.

The dependence on type of outlier, mentioned in (iii), is also as would be expected. The performance of Wilks' test depends only on the generalized distance, as reproduced by these simulations. With the second type of outlier, the second set of variables is actually irrelevant to the problem, so that their inclusion should not (if n is big enough) affect the test. Since their inclusion should not affect the test, the way they are treated should also not have any effect: in other words, the modified and unmodified tests should be the same, for large enough n, for the second type of outlier. The tables confirm that there is little difference at n=50, or even at n=30. Such difference as there is, is again due to the need to estimate fewer parameters in the modified test, which has a larger impact when the sample size is small. With the first type of outlier, on the other hand, both sets of variables are relevant to the problem.

The chief conclusion of this illustrative study is that, since such a large increase in power can be obtained in small- to medium-sized samples by using a more suitable outlier test statistic than the standard Wilks statistic, it is very much worthwhile to construct these alternative statistics. In the remainder of this chapter, statistics for some other particular structured matrices will be considered.

Table 7.3.1 Power comparison between the Wilks test and the Wilks test modified to incorporate block structure information. Case $p_1=p_2=2$, with

$$\Sigma_{11} = \begin{pmatrix} 1 & .4 \\ .4 & 1 \end{pmatrix} \qquad \Sigma_{22} = \begin{pmatrix} 1 & -.4 \\ -.4 & 1 \end{pmatrix}$$

Outlier type 1:

slip=$\sqrt{dist}/\sqrt{2\{1/(1+\rho_1)+1/(1+\rho_2)\}}$, is the slippage added to each component where ρ_1 and ρ_2 are the correlations corresponding to the first and second blocks of the variance-covariance matrix. Dist is the squared generalized distance. In this case $\rho_1=.4$ and $\rho_2=-.4$. Slippage for (i) dist=30 is 2.50998 and for (ii) dist=15 is 1.77482 in each component.

Outlier type 2:

Slip=$\sqrt{dist(1+\rho_1)/2}$ is the slippage added to the first two components of 1st observation which for (i) dist=30 is 4.58258 and (ii) dist=15 is 3.24037.

a) α=0.01

% of times an outlier detected by:

n	Mod. Wilks	Wilks	Mod. Wilks only	Wilks only
(i)	Sq. gen. distance=30, outlier type 1.			
10	35.98	7.18	30.29	1.49
20	63.49	39.14	26.18	1.83
30	70.19	54.48	17.30	1.59
50	75.21	66.91	10.03	1.73
100	76.01	72.10	5.28	1.36
(ii)	Sq. gen. distance=30, outlier type 2.			
10	15.58	6.76	11.28	2.46
20	46.01	38.61	10.55	3.15
30	58.83	54.55	7.01	2.74
50	68.44	66.04	4.15	1.75
100	72.68	71.95	2.19	1.46
(iii)	Sq. gen. distance=15, outlier type 1.			
10	10.75	2.64	9.29	1.18
20	18.98	9.80	10.95	1.78
30	21.55	14.38	9.06	1.89
50	23.48	18.25	7.21	1.99
100	22.59	20.05	4.46	1.93
(iv)	Sq. gen. distance=15, outlier type 2.			
10	6.11	3.13	4.54	1.55
20	12.61	10.01	5.13	2.53
30	16.91	14.56	4.34	1.99
50	20.06	18.31	3.56	1.81
100	19.84	18.96	2.21	1.34

b) α=0.025

% of times an outlier detected by:

n	Mod. Wilks	Wilks	Mod. Wilks only	Wilks only
(i)	Sq. gen. distance=30, outlier type 1.			
10	51.63	15.18	38.46	2.01
20	74.75	54.14	22.38	1.76
30	79.78	66.71	14.63	1.56
50	83.13	76.76	7.61	1.25
100	82.84	79.85	4.14	1.15
(ii)	Sq. gen. distance=30, outlier type 2.			
10	28.19	14.44	17.63	3.88
20	59.88	53.04	10.11	3.28
30	70.99	66.99	6.11	2.11
50	78.04	76.30	3.51	1.78
100	80.70	80.05	1.70	1.05
(iii)	Sq. gen. distance=15, outlier type=1.			
10	18.53	6.45	14.54	2.46
20	28.30	17.53	13.46	2.69
30	30.50	22.69	10.51	2.70
50	32.66	27.35	7.88	2.56
100	30.49	28.25	4.53	2.29
(iv)	Sq. gen. distance=15, outlier type=2.			
10	11.78	6.74	8.00	3.88
20	21.50	17.45	7.33	3.28
30	25.78	22.60	5.80	2.63
50	28.25	27.03	3.76	2.54
100	27.15	26.68	2.26	1.79

c) α=0.05

% of times an outlier detected by:

n	Mod. Wilks	Wilks	Mod. Wilks only	Wilks only
(i)	Sq. gen. distance=30, outlier type=1.			
10	63.51	25.18	40.75	2.41
20	82.39	65.20	18.79	1.60
30	86.05	75.88	11.45	1.28
50	88.31	83.23	6.18	1.09
100	87.36	85.10	3.34	1.08
(ii)	Sq. gen. distance=30, outlier type=2.			
10	40.59	24.34	20.99	4.74
20	70.36	64.16	9.46	3.26
30	79.24	76.63	4.71	2.10
50	83.74	82.78	2.49	1.53
100	85.78	85.13	1.49	0.84
(iii)	Sq. gen. distance=15, outlier type=1.			
10	27.44	11.51	19.61	3.69
20	36.85	25.85	14.41	3.41
30	39.29	31.31	11.36	3.39
50	41.11	35.75	8.44	3.08
100	37.96	35.61	4.98	2.63
(iv)	Sq. gen. distance=15, outlier type=2.			
10	19.05	12.34	11.29	4.58
20	30.76	26.31	8.33	3.88
30	33.84	31.53	6.31	4.00
50	36.84	34.83	4.64	2.63
100	35.04	34.13	3.15	2.24

d) α=0.10

% of times an outlier detected by:

n	Mod. Wilks	Wilks	Mod. Wilks only	Wilks only
(i)	Sq. gen. distance=30, outlier type=1.			
10	75.84	39.68	38.75	2.59
20	88.46	76.30	13.56	1.40
30	91.19	84.26	7.96	1.04
50	92.16	88.98	4.18	0.99
100	91.24	89.78	2.39	0.93
(ii)	Sq. gen. distance=30, outlier type=2.			
10	55.58	39.16	22.31	5.90
20	81.20	75.85	7.81	2.46
30	86.29	84.24	3.75	1.70
50	89.10	88.23	1.98	1.10
100	89.90	89.54	1.21	0.85
(iii)	Sq. gen. distance=15, outlier type=1.			
10	39.53	21.09	24.09	5.65
20	48.08	37.00	15.25	4.18
30	49.79	42.66	11.11	3.99
50	51.23	46.39	8.40	3.56
100	47.38	44.91	5.36	2.90
(iv)	Sq. gen. distance=15, outlier type=2.			
10	29.56	22.43	14.13	6.99
20	41.68	37.41	9.26	0.50
30	45.54	42.74	7.04	4.24
50	46.80	45.38	5.06	3.64
100	44.39	43.64	3.61	2.86

Table 7.3.2 Power comparison between the Wilks test and the Wilks test modified to incorporate block structure information. Case $p_1=2$, $p_2=4$, with

$$\Sigma_{11}=\begin{pmatrix} 1 & .4 \\ .4 & 1 \end{pmatrix} \qquad \Sigma_{22}=\begin{pmatrix} 1 & -.2 & -.2 & -.2 \\ -.2 & 1 & -.2 & -.2 \\ -.2 & -.2 & 1 & -.2 \\ -.2 & -.2 & -.2 & 1 \end{pmatrix}$$

Outlier type 1 :

Slippage added to 1st observation of sample is $slip=\sqrt{dist}/\sqrt{\{2/(1+\rho_1)+4/(1+3\rho_2)\}}$ where ρ_1 and ρ_2 are the corresponding correlations of the first and second blocks of the variance-covariance matrix. In this case $\rho_1=.4$ and $\rho_2=-.2$. Dist is the squared generalized distance. This gives slippage (i) for dist=30, slip=1.62019 and (ii) for dist=15, slip=1.14564 in each component.

Outlier type 2 :

$Slip=\sqrt{dist(1+\rho_1)/2}$ is the slippage added to the first two components of 1st observation. For (i) dist=30, slippage is slip=4.58258 and for (ii) dist=15, it is slip=3.24037.

a) $\alpha=0.01$

% of times an outlier detected by:

n	Mod. Wilks	Wilks	Mod. Wilks only	Wilks only
(i)	Sq. gen. distance=30, outlier type 1.			
10	10.53	1.81	9.79	1.08
20	38.79	21.40	19.88	2.49
30	51.49	38.61	15.00	2.12
50	60.40	52.31	10.20	2.11
100	64.03	60.53	5.25	1.75
(ii)	Sq. gen. distance=30, outlier type 2.			
10	6.40	1.70	6.01	1.31
20	29.58	20.54	13.33	4.29
30	43.83	37.63	10.59	4.39
50	56.13	51.85	7.09	2.81
100	61.78	60.15	3.79	2.16
(iii)	Sq. gen. distance=15, outlier type 1.			
10	3.43	1.13	3.21	0.91
20	9.24	5.10	5.50	1.36
30	12.80	9.21	5.45	1.86
50	14.18	11.86	3.98	1.66
100	15.01	13.81	2.59	1.39
(iv)	Sq. gen. distance=15, outlier type 2.			
10	2.50	1.48	2.28	1.25
20	7.46	4.85	4.35	1.74
30	10.25	7.93	4.51	2.19
50	12.65	11.18	3.69	2.21
100	14.50	13.71	2.36	1.58

b) α=0.025

% of times an outlier detected by:

n	Mod.Wilks	Wilks	Mod. Wilks only	Wilks only
(i)	Sq. gen. distance=30, outlier type 1.			
10	19.63	4.83	17.25	2.45
20	52.18	33.69	21.46	2.98
30	63.24	50.99	14.44	2.19
50	70.65	63.56	8.88	1.79
100	73.39	70.28	4.59	1.48
(ii)	Sq. gen. distance=30, outlier type 2.			
10	12.88	4.43	11.43	2.98
20	43.55	32.75	15.38	4.58
30	56.39	50.74	10.25	4.60
50	66.94	63.95	6.35	3.36
100	70.53	69.49	3.21	2.18
(iii)	Sq. gen. distance=15, outlier type 1.			
10	7.18	3.08	6.41	2.31
20	15.78	10.16	8.29	2.68
30	20.09	14.79	8.05	2.75
50	21.74	18.28	5.79	2.33
100	22.10	19.83	3.95	1.68
(iv)	Sq. gen. distance=15, outlier type 2.			
10	5.49	3.31	4.63	2.45
20	13.93	9.71	7.25	3.04
30	16.80	14.29	6.04	3.53
50	19.95	18.25	4.88	3.18
100	20.98	20.44	2.91	2.38

c) $\alpha=0.05$

% of times an outlier detected by:

n	Mod. Wilks	Wilks	Mod. Wilks only	Wilks only
(i)	Sq. gen. distance=30, outlier type 1.			
10	30.21	8.95	25.28	4.01
20	62.80	45.46	20.55	3.21
30	71.76	61.53	12.45	2.21
50	77.41	71.99	7.16	1.74
100	79.59	77.01	3.91	1.34
(ii)	Sq. gen. distance=30, outlier type 2.			
10	21.35	8.38	17.75	4.78
20	55.39	45.29	15.15	5.05
30	66.71	61.20	9.65	4.14
50	75.48	72.98	5.38	2.88
100	77.35	76.24	3.23	2.11
(iii)	Sq. gen. distance=15, outlier type 1.			
10	12.74	6.30	10.71	4.28
20	23.51	15.51	11.71	3.71
30	27.76	22.19	9.31	3.74
50	29.94	25.70	7.20	2.96
100	28.56	26.83	4.16	2.43
(iv)	Sq. gen. distance=15, outlier type 2.			
10	10.29	6.25	8.35	4.31
20	20.58	15.55	9.63	4.60
30	24.69	21.68	7.74	4.73
50	27.75	25.78	5.89	3.91
100	27.55	27.03	3.66	3.14

d) α=0.10

% of times an outlier detected by:

n	Mod. Wilks	Wilks	Mod. Wilks only	Wilks only
(i)	Sq. gen. distance=30, outlier type 1.			
10	44.64	16.76	33.55	5.68
20	73.61	58.66	17.84	2.89
30	80.00	72.08	10.10	2.18
50	84.10	80.03	5.63	1.55
100	85.28	83.33	3.15	1.20
(ii)	Sq. gen. distance=30, outlier type 2.			
10	34.30	16.25	25.04	6.99
20	67.80	58.70	13.81	4.71
30	76.03	72.39	7.58	3.94
50	82.66	80.80	4.23	2.36
100	83.76	83.29	2.48	2.00
(iii)	Sq. gen. distance=15, outlier type 1.			
10	21.38	12.51	15.91	7.05
20	33.55	25.14	13.81	5.40
30	38.11	32.49	10.30	4.68
50	39.66	35.41	7.88	3.63
100	38.00	35.76	5.24	3.00
(iv)	Sq. gen. distance=15, outlier type 2.			
10	18.13	11.78	13.81	7.46
20	30.80	25.20	12.08	6.48
30	34.75	31.26	9.01	5.53
50	38.11	35.30	7.05	4.24
100	37.28	36.66	4.50	3.89

7.4 Testing for one outlier when Σ is the equicorrelation matrix

One motivation for interest in the equicorrelation matrix

$$\begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ & & \dots & \\ \rho & & \dots & 1 \end{pmatrix}$$

was mentioned in § 7.1. Another arises frequently in sociological and psychological research, where scales of measurement are often set up which consist of the sum of scores on a number of related items which are all answered in the same way (such as "strongly agree" to "strongly disagree" on a 5-point scale). One question about such a scale is whether it is a reliable measurement of whatever it is that it does measure, in the sense that a ruler provides a reliable measurement of length because it will give the same answer when applied to the same object under the same conditions; if the scale is reliable, then the researcher can go on to study its validity - whether the thing that it measures is what he would like it to measure. A basic coefficient assessing reliability is Cronbach's alpha (e.g. Carmines and Zeller, 1979). This is usually computed as

$$\alpha = N\bar{\rho}/\{1+(N-1)\bar{\rho}\}$$

where the scale consists of N items and $\bar{\rho}$ is the average of all the N(N-1)/2 inter-item correlations. The derivation of this involves the assumption that the items are parallel measures of the same concept, all possessing exactly the same properties, including equal means and variances, and hence having equal correlations with each other. In the equicorrelation matrix thus assumed, the maximum likelihood estimator of the common correlation ρ is just $\bar{\rho}$: the proof of this will now be seen in the derivation of an outlier test statistic for data supposed to follow the equicorrelation model.

A two-stage maximum likelihood statistic will be found for testing sample homogeneity against an alternative of a single outlier with slippage in the mean. The hypotheses are

$$H_0 : \quad x_i \sim N_p(\mu,\Sigma) \qquad i=1,\ldots,n$$

$$H_1 : \quad x_i \sim N_p(\mu,\Sigma) \qquad i\neq j$$
$$\quad x_j \sim N_p(\mu+a,\Sigma)$$

where μ, a and j are unknown, and $\Sigma=\sigma^2 R$ has the equicorrelation form with unknown σ^2 and ρ.

The log-likelihood under H_0 has already been given as equation (3.1.2). Apart from a constant, it is

$$l(\mu,V)=\frac{n}{2}\ln|V|-\frac{n}{2}\mathrm{tr}(VS)-\frac{n}{2}(\bar{x}-\mu)'V(\bar{x}-\mu) \qquad (7.4.1)$$

where $V=\Sigma^{-1}$ and $nS=\sum_i (x_i-\bar{x})(x_i-\bar{x})'$

As before, taking derivatives of the log-likelihood maximized over μ at $\mu=\bar{x}$ (whereupon the third term vanishes),

$$\frac{\partial l}{\partial V} = \frac{n}{2}\left[\frac{\partial \ln|V|}{\partial V}-\frac{\partial \mathrm{tr}(VS)}{\partial V}\right]$$

$$= \frac{n}{2}(2\Sigma-\mathrm{diag}\Sigma-2S+\mathrm{diag}S) \qquad (7.4.2)$$

using results from Mardia, Kent and Bibby (1979) quoted in § 3.1. Now in this particular problem,

$$V=\alpha(I+\beta J)$$

where $\alpha(\rho,\sigma^2)=\sigma^{-2}(1-\rho)^{-1}$ and $\beta(\rho)= -\rho\{1+(p-1)\rho\}^{-1}$. In order to obtain $\partial l/\partial\alpha$ and $\partial l/\partial\beta$, we first prove the following lemma.

Lemma : If the matrix $Q=Q(\theta)$ with θ scalar, then

$$\frac{\partial l}{\partial \theta} = \mathrm{tr}\left(\frac{\partial l}{\partial Q} \frac{\partial Q'}{\partial \theta} \right)$$

Proof: For matrix $Q=(q_{ij})$, $\partial l/\partial Q$ means the matrix $(\partial l/\partial q_{ij})$. Now

$$\frac{\partial l}{\partial \theta} = \sum_i \sum_j \frac{\partial l}{\partial q_{ij}} \frac{\partial q_{ij}}{\partial \theta}$$

$$= \sum_i \sum_j \left(\frac{\partial l}{\partial Q} \right)_{ij} \left(\frac{\partial Q}{\partial \theta} \right)_{ij}$$

$$= \sum_i \sum_j \left(\frac{\partial l}{\partial Q} \right)_{ij} \left(\frac{\partial Q'}{\partial \theta} \right)_{ji}$$

$$= \mathrm{tr}\left(\frac{\partial l}{\partial Q} \frac{\partial Q'}{\partial \theta} \right)$$

Applying this lemma to $V=\alpha(I+\beta J)$ in (7.4.2),

$$\frac{\partial l}{\partial \alpha} = \mathrm{tr}\left(\frac{\partial l}{\partial V} \frac{\partial V}{\partial \alpha} \right) \quad \text{because V is symmetric}$$

$$= \mathrm{tr}\left\{ \frac{\partial l}{\partial V} (I+\beta J) \right\}$$

$$= \mathrm{tr}\left(\frac{\partial l}{\partial V} \right) + \beta \mathrm{tr}\left(\frac{\partial l}{\partial V} J \right)$$

$$= \frac{n}{2} \mathrm{tr}\Sigma - \frac{n}{2} \mathrm{tr}S + \beta \mathrm{tr}\left(\frac{\partial l}{\partial V} J \right) \qquad (7.4.3)$$

and

$$\frac{\partial l}{\partial \beta} = \mathrm{tr}\left(\frac{\partial l}{\partial V} \frac{\partial V}{\partial \beta} \right)$$

$$= \alpha \mathrm{tr}\left(\frac{\partial l}{\partial V} J \right) \qquad (7.4.4)$$

Since $\partial l/\partial \alpha = 0$ and $\partial l/\partial \beta = 0$ at the maximum likelihood estimates,

$$\mathrm{tr}\left(\frac{\partial l}{\partial V} J \right) = 0 \quad \text{from (7.4.4), and}$$

$$\frac{n}{2} (\mathrm{tr}\hat{\Sigma} - \mathrm{tr}S) = 0 \quad \text{from (7.4.3)}$$

giving

$$\frac{n}{2}(p\hat{\sigma}^2-trS) = 0$$

from which

$$\hat{\sigma}^2 = (1/p)trS.$$

Returning to (7.4.4), at the m.l.e.

$$tr\left(\frac{\partial l}{\partial V} J\right) = 0$$

which gives

$$tr\{2\hat{\Sigma}J-(diag\hat{\Sigma})J-2SJ+(diagS)J\}=0$$

$$2p\{1+(p-1)\hat{\rho}\}\hat{\sigma}^2-p\hat{\sigma}^2-2\sum_i\sum_j s_{ij}+trS=0$$

$$p\{1+(p-1)\hat{\rho}\}\hat{\sigma}^2 = \sum_i\sum_j s_{ij} = \sum_{i\neq j}\sum s_{ij}+trS$$

so finally

$$\hat{\rho}=\frac{1}{p(p-1)\hat{\sigma}^2}\sum_{i\neq j}\sum s_{ij}$$

Going back to the log-likelihood maximized over μ and writing $V=\sigma^{-2}R^{-1}$,

$$l= -\frac{np}{2}\ln\sigma^2-\frac{n}{2}\ln|R|-\frac{n}{2}\sigma^{-2}tr(R^{-1}S)$$

it can immediately be seen, by differentiating with respect to σ^2, that

$$\hat{\sigma}^2=tr(R^{-1}S)/p$$

so that the maximized log-likelihood is

$$l_0 = -\frac{np}{2}\ln\hat{\sigma}^2-\frac{n}{2}\ln|R| - \frac{np}{2}$$

under H_0.

Under H_1 for specified j, the log-likelihood is (omitting the same constant)

$$l(\mu,V)= \frac{n}{2}\ln|V|-\frac{(n-1)}{2}tr(VS_j)-\frac{n}{2}tr\{V(\bar{x}_j-\mu)(\bar{x}_j-\mu)'\}$$

$$-\frac{1}{2}(x_j-\mu-a)'V(x_j-\mu-a)$$

so that, once m.l.e.'s $\hat{\hat{\mu}}=\bar{x}_j$ and $\hat{\hat{a}}=x_j-\hat{\hat{\mu}}$ have been inserted, the log-likelihood differs from the null case only in replacing S by $(n-1)S_j/n$. Consequently

$$\hat{\sigma}_j^2 = \frac{n-1}{np}\,tr(S_j)$$

$$\hat{\rho}_j = \frac{n-1}{np(p-1)\hat{\sigma}_j^2}\sum\sum_{i\neq k}(S_j)_{ik}$$

using the j subscript to show that point x_j was omitted. With these estimates, the maximized log-likelihood is

$$l_1= -\frac{np}{2}\ln\hat{\sigma}_j^2 - \frac{n}{2}\ln|\hat{R}_j|$$

and hence the likelihood ratio λ, for given j is,

$$\lambda^{2/n}= \left(\frac{\hat{\sigma}_j^2}{\hat{\sigma}^2}\right)^p \frac{|\hat{R}_j|}{|\hat{R}|} \qquad (7.4.5)$$

Using the two-stage method, the minimum value of this over all j=1,...,n gives the outlier test statistic for unknown j, called EC for equicorrelation

$$EC= \min_j \left(\frac{\hat{\sigma}_j^2}{\hat{\sigma}^2}\right)^p \frac{|\hat{R}_j|}{|\hat{R}|}$$

It can be seen that it has a form analogous to Wilks' statistic, which is based on $|A_j|/|A|$ and hence is proportional to $|\hat{\Sigma}_j|/|\hat{\Sigma}|$. This is the ratio of determinants of the m.l.e's of the covariance matrix under the two hypothesis, which is the same structure as (7.4.5).

It does not seem possible to find the distribution of the above test statistic EC for an outlier from the equicorrelation model. In order to investigate its performance the following simulation studies were undertaken. Firstly, simulated percentage points were obtained under the null hypothesis. Secondly, these percentage points were used to examine the power of the EC test in comparison to the power of Wilks' test, which can be applied to the same data but does not utilize the extra information that the covariance matrix has the equicorrelation structure.

Simulated 1, 2.5, 5 and 10% critical values are shown in Table 7.4.1 (a)-(d). The entries in the tables are based on 30,000 simulated samples from $N_p(0,R)$ with the given combination of n and p=2; 18,000 simulations for p=3; 16,000 for p=4 and 14,000 for p=5. The different numbers of simulations arose because data under the null hypothesis were generated for different values of ρ in the range $(\rho_0(p),0.9)$, where

$$\rho_0(p) = \frac{-1}{p-1}$$

is the lowest possible value of ρ at which the equicorrelation matrix ceases to be positive definite. This was done to confirm that the percentage points do not depend on ρ.

Simulated powers are shown in Table 7.4.2 (a)-(d). In this case, each entry is based on 6,000 runs, and is the average of three lots of 2,000 simulations each for different values of ρ (again, to confirm that results were independent of ρ). A simple outlier was generated by adding the quantity $\lambda\underset{\sim}{1}$ to the first member of the sample, where λ was chosen so that the generalized distance of the slippage from the mean (the origin) was 20 in the metric of equicorrelation covariance matrix R:, the computation

is shown in Appendix II. Wilks' statistic and the EC statistic were both evaluated on the same data.

In examining the powers it should first be noted that in the case p=2 there is only one correlation, so the imposition of the equicorrelation structure actually refers to imposing the single condition of equality between the two variances. The number of parameters to be estimated is always 2 in the equicorrelation matrix for any number of dimensions, in comparison to p(p+1)/2 in the unrestricted matrix:

p	p(p+1)/2
2	3
3	6
4	10
5	15

From this, it is obvious that for a given sample size the power of the new EC test statistic will decline much less steeply as p increases than will the power of Wilks' test, so that the difference in powers of the tests will increase quite sharply from the rather small difference which should exist at p=2 (where a single restriction is imposed). This is borne out by the tables. In testing at the 5% level in a sample of size n=20, the power of the new procedure is only 4.7% greater than that of Wilks' statistic (0.467 to 0.447) for p=2, increasing to 18.1% for p=3, 31.1% for p=4 and 46.0% for p=5 (0.285 to 0.195).

The conclusion to be drawn from the data in the tables is that utilizing the information on equicorrelation, by using the EC test statistic, makes a substantial difference to the power of the single outlier test in small and moderate samples, say up to at least n=30 for the range of dimensions considered. Hence it is of practical importance to exploit this information wherever possible. Since it would probably be necessary in some

applications to confirm the applicability of the equicorrelation hypothesis, a test for this structure should also be carried out. Following the derivation above, it can be shown that a likelihood ratio test for equicorrelation against an unrestricted alternative is (Wilks, 1946; Morrison, 1976)

$$L = |\tilde{\Sigma}|/(\hat{\sigma}^2)^p|\hat{R}|$$

where $\tilde{\Sigma}$ is the unrestricted m.l.e. Then -ln L times the factor

$$n-1-p(p+1)^2(2p-3)/\{6(p-1)(p^2+p-4)\}$$

is asymptotically X^2 with $[p(p+1)/2]-2$ degrees of freedom: this is the usual asymptotic result, incorporating a correction due to Box (1949, 1950). This test could be applied to the reduced sample obtained by omitting the suspected outlier x_j and significance levels would not be affected if x_j is a genuine outlier.

Table 7.4.1(a) Simulated 1% points for two-stage maximum likelihood test for a single outlier from the equicorrelation model.

Sample size	Dimensions p			
n	2	3	4	5
10	0.1670	0.1315	0.1007	0.0801
20	0.4216	0.3750	0.3263	0.2961
30	0.5601	0.5141	0.4722	0.4433
50	0.7003	0.6673	0.6359	0.6038
100	0.8286	0.8046	0.7872	0.7685

Table 7.4.1(b) Simulated 2.5% points for two-stage maximum likelihood test for a single outlier from the equicorrelation model.

Sample size n	Dimensions p: 2	3	4	5
10	0.2078	0.1658	0.1306	0.1027
20	0.4674	0.4151	0.3678	0.3328
30	0.5983	0.5507	0.5084	0.4737
50	0.7270	0.6922	0.6603	0.6310
100	0.8438	0.8212	0.8026	0.7838

Table 7.4.1 (c) Simulated 5% points for two-stage maximum likelihood test for a single outlier from the equicorrelation model.

Sample size n	Dimensions p: 2	3	4	5
10	0.2489	0.1992	0.1581	0.1247
20	0.5052	0.4479	0.4014	0.3617
30	0.6291	0.5796	0.5370	0.5038
50	0.7479	0.7123	0.6806	0.6525
100	0.8564	0.8348	0.8151	0.7974

Table 7.4.1 (d) Simulated 10% points for two-stage maximum likelihood test for a single outlier from the equicorrelation model.

Sample size n	Dimensions p: 2	3	4	5
10	0.2979	0.2373	0.1908	0.1533
20	0.5477	0.4891	0.4379	0.3969
30	0.6628	0.6106	0.5707	0.5341
50	0.7704	0.7345	0.7038	0.6760
100	0.8685	0.8469	0.8289	0.8115

Table 7.4.2 (a) Comparison between simulated powers of EC test statistic for a single outlier from the equicorrelation model and Wilks' statistic: α=0.01

Sample size n	Dimensions p	Proportion of times that an outlier is declared: EC test	Wilks' test	Only EC test	Only Wilks' test
10	2	0.1417	0.0942	0.0640	0.0165
	3	0.1137	0.0503	0.0803	0.0170
	4	0.0885	0.0268	0.0768	0.0152
	5	0.0747	0.0165	0.0695	0.0113
20	2	0.2580	0.2258	0.0490	0.0168
	3	0.1953	0.1658	0.0553	0.0257
	4	0.1682	0.1060	0.0817	0.0195
	5	0.1238	0.0632	0.0810	0.0203
30	2	0.2942	0.2890	0.0267	0.0215
	3	0.2272	0.2112	0.0450	0.0290
	4	0.1863	0.1403	0.0667	0.0207
	5	0.1635	0.1017	0.0828	0.0210
50	2	0.3202	0.3070	0.0262	0.0130
	3	0.2498	0.2388	0.0352	0.0242
	4	0.2215	0.1832	0.0555	0.0172
	5	0.1785	0.1417	0.0585	0.0217
100	2	0.3107	0.3102	0.0107	0.0102
	3	0.2552	0.2405	0.0287	0.0140
	4	0.2162	0.1965	0.0350	0.0153
	5	0.1623	0.1485	0.0325	0.0187

Table 7.4.2 (b) Comparison between simulated powers of EC test statistic for a single outlier from the equicorrelation model and Wilks' statistic: α=0.025

Sample size n	Dimensions p	Proportion of times that an outlier is declared: EC test	Wilks' test	Only EC test	Only Wilks' test
10	2	0.2403	0.1855	0.0850	0.0302
	3	0.1985	0.1098	0.1220	0.0333
	4	0.1660	0.0632	0.1333	0.0305
	5	0.1362	0.0427	0.1210	0.0275
20	2	0.3715	0.3337	0.0595	0.0217
	3	0.3050	0.2482	0.0892	0.0323
	4	0.2587	0.1843	0.1110	0.0367
	5	0.2072	0.1233	0.1170	0.0332
30	2	0.4037	0.3845	0.0385	0.0193
	3	0.3282	0.2975	0.0588	0.0282
	4	0.2758	0.2283	0.0805	0.0330
	5	0.2525	0.1790	0.1037	0.0302
50	2	0.4308	0.4188	0.0298	0.0178
	3	0.3573	0.3325	0.0480	0.0232
	4	0.3012	0.2693	0.0562	0.2433
	5	0.2622	0.2212	0.0680	0.0270
100	2	0.4070	0.4045	0.0147	0.0122
	3	0.3408	0.3353	0.0247	0.0192
	4	0.2947	0.2790	0.0395	0.0238
	5	0.2382	0.2233	0.0408	0.0260

Table 7.4.2 (c) Comparison between simulated powers of EC test statistic for a single outlier from the equicorrelation model and Wilks' statistic: $\alpha=0.05$

Sample size n	Dimensions p	Proportion of times that an outlier is declared: EC test	Wilks' test	Only EC test	Only Wilks' test
10	2	0.3438	0.2915	0.0905	0.0382
	3	0.2818	0.1865	0.1477	0.0523
	4	0.2482	0.1148	0.1833	0.0500
	5	0.2228	0.0837	0.1842	0.0450
20	2	0.4673	0.4465	0.0497	0.0288
	3	0.4075	0.3450	0.1042	0.0417
	4	0.3553	0.2712	0.1273	0.0432
	5	0.2845	0.1950	0.1392	0.0497
30	2	0.5008	0.4850	0.0388	0.0230
	3	0.4215	0.3880	0.0687	0.0352
	4	0.3630	0.3163	0.0905	0.0438
	5	0.3298	0.2590	0.1147	0.0438
50	2	0.5157	0.5090	0.0270	0.0203
	3	0.4442	0.4338	0.0425	0.0322
	4	0.3835	0.3568	0.0628	0.0362
	5	0.3417	0.2985	0.0777	0.0345
100	2	0.4938	0.4832	0.0208	0.0102
	3	0.4300	0.4157	0.0360	0.0217
	4	0.3720	0.3650	0.0380	0.0310
	5	0.3210	0.3017	0.0530	0.0337

Table 7.4.2 (d) Comparison between simulated powers of EC test statistic for a single outlier from the equicorrelation model and Wilks' statistic: α=0.10

Sample size n	Dimensions p	Proportion of times that an outlier is declared: EC test	Wilks' test	Only EC test	Only Wilks' test
10	2	0.4743	0.4290	0.0888	0.0435
	3	0.4130	0.2955	0.1780	0.0605
	4	0.3635	0.2155	0.2223	0.0743
	5	0.3292	0.1630	0.2435	0.0773
20	2	0.5908	0.5663	0.0550	0.0305
	3	0.5228	0.4658	0.1025	0.0455
	4	0.4723	0.3887	0.1358	0.0522
	5	0.4142	0.3055	0.1727	0.0640
30	2	0.6167	0.5980	0.0403	0.0217
	3	0.5373	0.5050	0.0745	0.0422
	4	0.4815	0.4270	0.1043	0.0498
	5	0.4387	0.3638	0.1275	0.0527
50	2	0.6228	0.6123	0.0277	0.0172
	3	0.5480	0.5383	0.0477	0.0380
	4	0.4888	0.4615	0.0672	0.0398
	5	0.4467	0.4047	0.0495	0.0915
100	2	0.5910	0.5843	0.0225	0.0158
	3	0.5248	0.5222	0.0315	0.0288
	4	0.4787	0.4643	0.0485	0.1342
	5	0.4212	0.4038	0.0553	0.0380

CHAPTER 8

RESIDUALS AND INFLUENCE IN THE MULTIVARIATE LINEAR MODEL

8.1 Introduction

The examination of structured data is also the topic of this final chapter, since it deals with the relationship of a response vector to a vector of predictors via a linear model. Specifically, the model considered is

$$Y=XB+U \qquad (8.1.1)$$

where the nxq matrix Y holds n independent observations of the q-dimensional response vector, the nxp matrix X holds the corresponding observations of the p-dimensional vector of predictors, B is a pxq matrix of coefficients and U is an nxq matrix of random disturbances. X usually includes a column of ones. After replacing B by an estimate $\hat{B}$, the matrix $\hat{U}=Y-X\hat{B}$ holds the residuals: the ith row of this matrix, $\hat{u}_i'$, contains the residuals for the ith case on each of the q response dimensions. The topic of residuals from the univariate linear model (q=1) has been studied extensively (Cook and Weisberg, 1982). The basic purpose of examining residuals is to assess the adequacy and appropriateness of the model; this may include identification of outlying values, but a proper examination of residuals should assess the entire set and not just some extreme values. Along with the study of residuals, there has also been a lot of attention to the question of influence. This has been touched on already - see Figure 1.2.1 and the relevant discussion in the text of § 1.2.

In this chapter, these ideas are applied to the multivariate linear regression problem. Although ordinary least squares estimates of regression coefficients are the

same in the multivariate analysis as in q separate univariate analyses, so that the residuals for a particular response dimension are the same whether this is analyzed separately or together with the other responses, there are obvious reasons for carrying out the multivariate analysis to consider simultaneously the different response variables. One is the possibility that the residual for one response variable in a particular case may not seem to be out of the ordinary in relation to other residuals for that response, but only in relation to the residuals for other responses on the same case. Another is that interest may lie in specifically multivariate aspects of the data. For example, in the problem that prompted this investigation, the main item of interest was the matrix of inter-correlations between five indicators of pollution from sampling stations in the Aegean Sea. This was calculated as the matrix of correlations between the residuals from the regressions of the indicators on covariates including temperature and pH of the seawater. Correlations are particularly vulnerable to distortion by outlying values (Gnanadesikan & Kettenring, 1972), so examination of the multivariate residuals to protect against this was essential.

In the following two sections, multivariate residuals and influence measures are presented. Section 4 outlines an application to illustrate the usefulness of the methodology.

8.2 Residuals

It is useful to start by recalling some results from the univariate linear model as may be found in Cook and Weisberg or many other sources. Write the model as $y=X\beta+\varepsilon$ where y is the nx1 vector of n independent observations of the dependent variable, X is the nxp matrix of predictors, usually including a column of 1's, β is the px1 vector of regression coefficients and ε is the nx1 vector of

residuals, with variance $\text{var}(\varepsilon)=\sigma^2 I$. Estimation of β is usually carried out by ordinary least squares, which is the same as maximum likelihood when the normal distribution is assumed for ε. The estimator is

$$\hat{\beta}=(X'X)^{-1}X'y.$$

The observed least-squares residuals are

$$\begin{aligned} e&=y-\hat{y}=y-X\hat{\beta} \\ &=(I-X(X'X)^{-1}X')y \\ &=(I-V)y \end{aligned}$$

where $V=X(X'X)^{-1}X'$, is often called the "hat matrix". Substituting $y=X\beta+\varepsilon$, this expression for e reduces to

$$e=(I-V)\varepsilon$$

From this, it can be seen immediately that the variance of the observed residuals is

$$\text{Var}(e)=\sigma^2(I-V)$$

In particular, the variance of the residual for case i is $\sigma^2(1-v_{ii})$, where v_{ii} is the ith diagonal element of V. Therefore, in general, these residuals do not have the same variance and this inequality must be removed before comparing residuals between cases. One may do this by constructing the internally studentized residuals

$$r_i=e_i/\{\hat{\sigma}(1-v_{ii})^{1/2}\} \qquad (8.2.1)$$

where $\hat{\sigma}^2$ is the usual residual mean square, or the externally studentized residuals

$$t_i=e_i/\{\hat{\sigma}_{(i)}(1-v_{ii})^{1/2}\} \qquad (8.2.2)$$

where $\hat{\sigma}^2_{(i)}$ is the residual mean square obtained from fitting the regression to all cases except case i. The two versions are related by

$$t_i^2=r_i^2(n-p-1)/(n-p-r_i^2) \qquad (8.2.3)$$

which indicates that outlying residuals will appear even more widely separated from the rest on the scale of t_i than of r_i. This is one reason for preferring the externally studentized form; another is that its distribution is a very familiar one, since t_i follows the t distribution with n-p-1 degrees of freedom, whereas $r_i^2/(n-p)$ follows the Beta distribution with parameters 1/2 and (n-p-1)/2.

Equivalent results will now be developed for the multivariate linear model Y=XB+U. Row i of U is u_i' with covariance matrix $\Sigma=(\sigma_{jk})$. The ordinary least squares estimator is, analogously to the univariate case,

$$\hat{B}=(X'X)^{-1}X'Y$$

(for example, Mardia, Kent and Bibby, 1979). Hence, the observed residuals work out as

$$\hat{U}=(I-V)U \qquad (8.2.4)$$

by the same matrix manipulations as in the univariate case, with $V=X(X'X)^{-1}X'$ as before. It is easy to derive the variances and covariances of elements of U from first principles. Rewriting (8.2.4) for an individual element,

$$\hat{u}_{ij}=\sum_k \alpha_{ik}u_{kj}$$

where the summation range is 1 to n and $A=(\alpha_{ik})=I-V$. Hence

$$\operatorname{cov}(\hat{u}_{ij},\hat{u}_{il})=\operatorname{cov}(\sum \alpha_{ik}u_{kj},\ \sum \alpha_{ik}u_{kl})$$

$$=\sum \alpha_{ik}^2 \operatorname{cov}(u_{kj},u_{kl})$$

because the independence of different cases implies that $\text{cov}(u_{kj},u_{ml})=0$ for $m\neq k$. Hence

$$\text{cov}(\hat{u}_{ij},\hat{u}_{il})=\sigma_{jl}\sum_k \alpha_{ik}^2$$

Now $$\sum_k \alpha_{ik}^2=[(I-V)^2]_{ii}$$

and $$(I-V)^2=I-2V+V^2$$
$$=I-V$$

because, as can easily be checked from its definition, V is idempotent, that is $V^2=V$. Therefore

$$\text{cov}(\hat{u}_{ij},\hat{u}_{il})=(1-v_{ii})\sigma_{jl} \qquad (8.2.5)$$

For the case l=j, this means

$$\text{var}(\hat{u}_{ij})=(1-v_{ii})\sigma_{jj}$$

which is, of course, exactly the univariate regression result (as given earlier) for the variable of dimension j, with variance $\sigma_{jj}=\sigma_j^2$. Rewriting (8.2.5) in vector form,

$$\text{var}(\hat{u}_i)=(1-v_{ii})\Sigma \qquad (8.2.6)$$

where $\hat{u}_i'$ is the ith row of $\hat{U}$ and holds the observed residuals for case i on the q responses.

The question now is how to examine these residuals. One obvious way is to reduce each residual vector $\hat{u}_i$ to a scalar, which can conveniently be done by taking the quadratic form

$$\hat{u}_i'\{\text{var}(\hat{u}_i)\}^{-1}u_i$$

From (8.2.6), the expression is

$$R_i^2=\hat{u}_i'\hat{\Sigma}^{-1}\hat{u}_i/(1-v_{ii}) \qquad (8.2.7)$$

where $\hat{\Sigma}$ is the usual mean residual sums of squares and

products matrix, $\hat{U}'\hat{U}/(n-p)$. This expression is a matrix equivalent of (8.2.1). An externally studentized version analogous to (8.2.2) is

$$T_i^2=\hat{u}_i'\hat{\Sigma}_{(i)}^{-1}\hat{u}_i/(1-v_{ii}) \qquad (8.2.8)$$

where $\hat{\Sigma}_{(i)}$ is the mean residual SSP matrix from the linear model fitted to the data after deletion of case i.

Another version of residuals from the multivariate linear model can be found in Gnanadesikan (1977, § 6.4). This is $\hat{u}_i'S^{*-1}\hat{u}_i$, where S^* is a robust estimate of the covariance matrix of residuals. As it lacks the standardizing factor $1-v_{ii}$, this would not offer an adequate basis for assessing the regression model except in the case when the v_{ii} were all similar.

Distributions for both forms of residuals, (8.2.7) and (8.2.8), can be found under the assumption of a normal distribution for the random disturbances. The residual SSP matrix from the regression with case i omitted follows the Wishart distribution

$$\hat{U}_{(i)}'\hat{U}_{(i)} \sim W_q(\Sigma,n-p-1)$$

(Mardia, Kent and Bibby, 1979). Also, the observed residual $\hat{u}_i$ from the full regression has the distribution

$$\hat{u}_i(1-v_{ii})^{-1/2} \sim N_q(0,\Sigma)$$

Hence the residual T_i^2 defined in (8.2.8) is distributed proportionally to Hotelling's T^2 distribution (Mardia, Kent and Bibby, 1979, § 3.5), so

$$(n-p-q)T_i^2/\{q(n-p-1)\} \sim F_{q,n-p-q}$$

The distribution of R_i^2 is found from the relationship between R_i^2 and T_i^2 and between the F and Beta

distributions. From the usual updating formula, it can easily be shown that

$$T_i^2 = R_i^2 \, (n-p-1)/ \, (n-p-R_i^2)$$

which is exactly the same as the result (8.2.3) for the univariate residuals. It follows that

$$R_i^2/(n-p) \sim B(q/2,(n-p-q)/2)$$

Thus both common forms of univariate residual can readily be extended to the multivariate case. T_i^2 might again be preferred to R_i^2 because of its exaggeration of the separation of outliers, but the advantage of easy reference to a very familiar distribution no longer applies. As in the univariate case, however, the joint distribution of residuals is intractable and so the emphasis would anyway be on informal, chiefly graphical, methods rather than formal significance testing. Possibilities include Q-Q plots with simulated envelopes superimposed (Cook & Weisberg, 1982, section 2.3.4), as illustrated in the example below.

Other useful forms of residual might be developed. A referee for a published version of this material (Caroni, 1987) remarked that the multivariate residual vectors could be examined in ways other than by reducing to a distance. One possibility could be to rotate to principal axes, although strictly speaking this is not applicable because of the correlations between the vectors. In fact, a robust principal components analysis (Campbell, 1980, using the GENSTAT macro from Matthews, 1984, as mentioned in chapter 2) had been applied to the example of section 8.4 below, and was helpful in understanding in what way the outlying point identified there differed from the rest.

8.3 Influence

Cook & Weisberg (1982, Chapter 3) review methods of examining the influence of a case, or group of cases, on the univariate regression in the sense of the effect on the estimate of deleting the case or cases from the data. A basic measure is the sample influence curve, which here is proportional to $\hat{\beta}-\hat{\beta}_{(i)}$ where $\hat{\beta}_{(i)}$ is the estimate of $\hat{\beta}$ with case i deleted. As usual, one way of assessing this vector would be by reducing it to a scalar as a distance in some norm. There are various choices of norm, including that introduced by Cook (1977) which results in measures

$$D_i=(\hat{\beta}_{(i)}-\hat{\beta})'(X'X)(\hat{\beta}_{(i)}-\hat{\beta})/(p\hat{\sigma}^2)$$

This choice is motivated by analogy to the confidence ellipsoids for $\hat{\beta}$. To convert to a familiar scale, if the value of D_i equals the $100(1-\alpha)\%$ percentile of the F distribution with p and n-p degrees of freedom, the effect of the deletion of case i can be described as moving the estimate to the edge of a $100(1-\alpha)\%$ confidence ellipsoid. This device is employed in the BMDP regression program P9R. Arguing along exactly the same lines in the multivariate case leads to a sample influence curve proportional to $\hat{B}-\hat{B}_{(i)}$. This is a pxq matrix and so its assessment is not easy. Critchley (1985), in the context of examining the matrix of principal component scores, suggests that a norm such as $\{tr(A'A)\}^{1/2}$ could be used: in this application, A would be $\hat{B}-\hat{B}_{(i)}$.

In the same work, however, he develops influence curves separately for each sample eigenvector, rather than attempting to examine them together as a matrix. Following this line means investigating particular vectors extracted from $\hat{B}-\hat{B}_{(i)}$. Choosing a column of this matrix is equivalent to looking at the regression coefficients for the univariate regression, so requires only the established univariate theory. Another choice, peculiar

to the multivariate problem, is to take a row of the matrix. This corresponds to the regression coefficients for all response variables on a particular predictor and was a natural choice in all the applications tried so far, in which either there was only one predictor or only one predictor was of real interest to the experimenter. Let $\hat{\beta}_j'$ be row j of $\hat{B}$. Then, under normality assumptions,

$$g_{jj}^{-1/2}(\beta_j-\hat{\beta}_j) \sim N_q(0,\Sigma)$$

where g_{jj} is the jth diagonal element of $(X'X)^{-1}$ (Mardia, Kent and Bibby, 1979). This indicates how to construct a distance measure similar to Cook's, based on confidence ellipsoid analogies. Specifically

$$D_i=(\hat{\beta}_{j(i)}-\hat{\beta}_j)'(g_{jj}\hat{\Sigma})^{-1}(\hat{\beta}_{j(i)}-\hat{\beta}_j)(n-p-q+1)/\{q(n-p)\}$$

where $\hat{\beta}_{j(i)}$ is the estimate of β_j after deletion of case i, and this measure can be converted to a percentile of the F distribution with q and n-p-q+1 degrees of freedom.

Other versions of influence measurement, particularly replacing $\hat{\Sigma}$ by $\hat{\Sigma}_{(i)}$, could easily be developed, analogously to the choices listed in Table 3.5.4 of Cook and Weisberg. Graphical aids to the assessment of the set of all D_i for i=1...n could again include plots with simulated envelopes, as in Atkinson (1981).

Since the publication of the above material (Caroni, 1987), another paper has appeared on the topic of residuals from the multivariate linear model (Hossain & Naik, 1989). These authors also give the two forms of multivariate residual (8.2.7) and (8.2.8), as their equation (9), and present several influence measures by writing down matrix expressions equivalent to some of the measures which have been suggested for the linear model with a single response variable. They do not discuss the relative merits of the measures. The emphasis in their

example is on detecting influential observations, using cut-off values.

8.4 An Example

In order to illustrate the above methods and the circumstances in which they are useful, an example will now be described. The data (Table 8.4.1) come from a study of foetal development. Various dimensions were measured on the jaw bones of 9 foetuses and from these were calculated nine angles indicating alignment of the jaw bones. The only covariate recorded was the age of the foetus. The univariate regressions of angle against age are summarized in Table 8.4.2.

Table 8.4.1 Measurements of nine angles Y1...Y9 on jaws of 19 foetuses, with age in weeks.

Case	Age	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9
1	16	139.33	84.70	87.18	134.81	156.79	182.18	172.55	95.39	60.97
2	17	141.31	59.02	61.19	123.43	143.45	175.92	178.54	81.58	42.72
3	18	132.51	96.25	133.57	135.39	172.73	207.92	159.41	88.77	75.34
4	19	139.33	101.91	123.85	126.92	161.19	191.89	144.58	105.08	86.31
5	20	133.33	94.54	88.32	128.03	146.99	180.03	155.21	111.22	61.32
6	21	128.60	77.21	83.39	113.53	137.77	171.46	143.61	102.20	63.69
7	22	139.74	96.79	125.76	152.03	170.25	214.81	185.25	87.50	61.27
8	23	135.25	95.23	121.43	141.05	151.97	189.69	169.70	104.40	80.69
9	24	132.09	89.29	123.18	120.21	170.55	206.67	151.21	79.08	63.76
10	25	119.42	97.77	89.38	139.64	166.58	204.71	162.38	90.72	30.92
11	26	126.52	87.22	88.80	133.75	152.17	197.92	160.79	93.78	35.65
12	27	128.67	98.51	131.22	159.03	167.22	217.03	182.53	89.15	58.03
13	30	111.35	89.76	107.92	153.60	147.31	183.19	177.13	97.86	64.25
14	31	136.02	105.90	125.37	148.10	163.26	201.77	161.30	97.22	61.74
15	32	136.61	99.24	121.66	150.23	163.90	205.85	169.85	89.10	52.61
16	33	119.20	95.34	99.18	153.42	169.26	199.66	182.83	79.03	34.98
17	34	133.76	99.51	113.58	139.68	181.09	210.50	156.96	70.69	37.09
18	35	130.11	84.43	129.56	162.08	181.39	213.71	181.89	54.34	48.60
19	36	125.72	108.09	140.71	131.50	161.52	195.17	156.43	97.10	76.89

Table 8.4.2 Univariate regressions on age for data of Table 8.4.1

Dependent angle	Regression coefficients: Constant	Age	$F_{1,17}$	p	R^2	$\hat{\sigma}$
Y_1	145.65	-0.569 (0.263)*	4.70	0.045	0.216	7.224
Y_2	72.71	0.775 (0.374)	4.31	0.053	0.053	10.270
Y_3	71.48	1.507 (0.729)	4.28	0.054	0.201	20.029
Y_4	108.66	1.190 (0.422)	7.93	0.012	0.318	11.612
Y_5	138.87	0.873 (0.406)	4.62	0.046	0.214	11.157
Y_6	172.11	0.982 (0.459)	4.57	0.047	0.212	12.622
Y_7	155.72	0.396 (0.482)	0.67	0.423	0.038	13.257
Y_8	113.83	-0.917 (0.447)	4.20	0.056	0.198	12.295
Y_9	76.70	-0.737 (0.576)	1.64	0.218	0.088	15.833

* standard error in parentheses

The regression was clearly statistically significant for the fourth angle, non-significant for the seventh and ninth angles and round about the 5% level for the remainder. Q-Q normal plots of the residuals from each separate regression were drawn, as illustrated in Figure 8.4.1 for the second angle. The residuals are listed in Table 8.4.3.

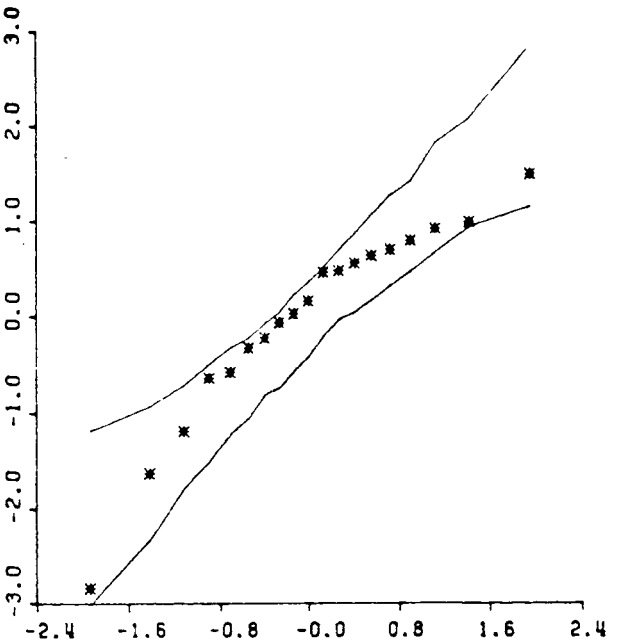


Figure 8.4.1 Normal probability plot for internally studentized residuals r_i from regression of second angle on age, with simulated 95% envelope.

A 95% envelope from 200 simulations of pseudo-randomly normally distributed residuals with the same structure is superimposed (Atkinson, 1981); this serves the purpose of indicating what shape of plot is acceptable, since these are not 19 independent observations as required for probability plotting. As the observed residuals fall within the envelope there is no evidence of violation of assumptions. In particular, the point at the bottom left of the plot does not seem to be excessively large numerically, even though this was the most extreme of all residuals from the nine univariate regressions. On the basis of these and other examinations of the regression fits, it seems that all is well with these univariate regressions.

The multivariate regression of the nine-dimensional response variable against age resulted in a large residual for case 3, with $T_3^2=169.5$ ($F_{9,8}=9.42$, $p=0.00217$): all

residuals are given in Table 8.4.3. Using a Bonferroni upper bound for the significance of this considered as the maximum of a set of 19 values, gives p=0.00217x19=0.041. Figure 8.4.2 shows the same point falling outside the 95% envelope from 200 simulations of a probability plot for residuals (Atkinson, 1981), confirming that it is unlikely to be from the same normal distribution as the rest.

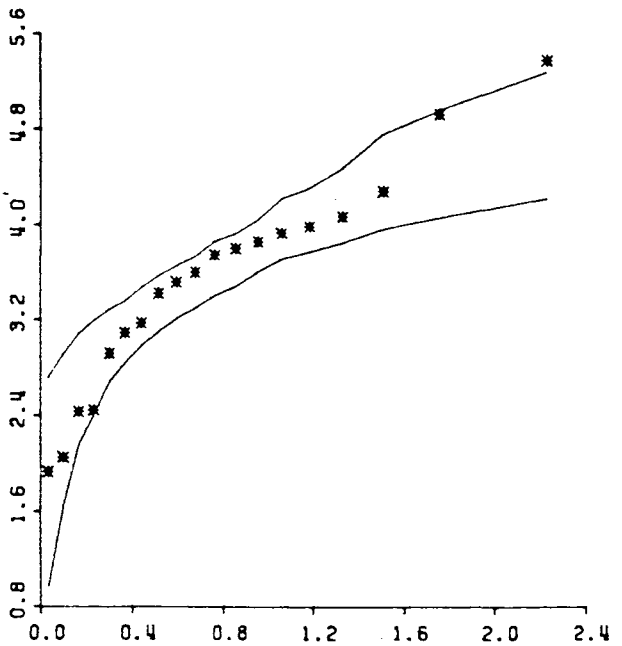


Figure 8.4.2 Probability plot for externally studentized residuals T_i^2 with simulated 95% envelope.

(Because the F Q-Q plot has a very awkward scale for plotting, this plot was constructed by mapping the F-deviates onto normal deviates. If

$$p=F(x)=\Phi(y)$$

where F is the F distribution function and Φ the normal distribution function, then

$$y=\Phi^{-1}(F(x))$$

gives probability plotting positions for a normal Q-Q

plot.) The residuals for case 3 in the nine univariate regressions did not have unusual values. However, the recorded value for the third angle in this observation was the second largest in the data set and it appears that, although not excessively large in itself (the externally studentized univariate residual is 2.04), it violated the pattern of the data since it was not accompanied by particularly large values of other angles which were positively correlated with the third angle. Therefore this suspicious data value was only to be seen by employing the multivariate residual analysis developed here.

Influence on regression coefficients could be assessed by the method of Section 8.3 because attention focussed on one particular row of the coefficient matrix, namely the coefficients of age. The constant terms, forming the first row of the matrix, were not of interest. In fact calculation of the distance measure defined above showed that no point had undesirably high influence on the regression coefficients and, in particular, the point with the large residual did not even have the largest influence.

Case	Dimension 1	2	3	4	5	6	7	8	9	Multi-variate
1	.416	-.044	-.453	.664	.381	-.481	.867	-.328	-.267	24.32
2	.795	-3.811	-2.146	-.500	-1.000	-1.117	1.351	-1.529	-1.529	29.75
3	-.419	1.001	2.040	.480	1.869	1.614	-.270	-.736	.799	169.49
4	.650	1.557	1.281	-.388	.534	.092	-1.555	.739	1.664	19.62
5	-.131	.636	-.688	-.391	-.874	-.975	-.657	1.382	-.041	4.25
6	-.725	-1.215	-1.031	-1.953	-1.965	-1.886	-1.694	.637	.158	16.32
7	.948	.698	1.101	1.606	1.142	1.885	1.721	-.507	.050	91.71
8	.376	.460	.779	.436	-.634	-.399	.370	.979	1.405	4.69
9	.015	-.198	.789	-1.572	.989	.892	-1.094	-1.071	.301	22.89
10	1.818	.556	-1.016	.106	.531	.645	-.243	-.014	-1.908	18.00
11	-.603	-.554	-1.131	-.506	-.859	.023	-.394	.309	-1.467	12.72
12	-.222	.476	.976	1.704	.430	1.563	1.274	.007	.078	25.89
13	-3.012	.618	-.445	.820	-1.753	-1.580	.739	.976	.624	28.02
14	1.177	.930	.365	.224	-.244	-.062	-.517	1.008	.511	6.54
15	1.379	.171	.099	.309	-.267	.189	.114	.387	-.032	6.48
16	-1.142	-.299	-1.189	.494	.147	-.399	1.142	-.383	-1.187	11.77
17	1.126	.044	-.482	-.872	1.233	.419	-.995	-1.054	-.992	10.03
18	0.654	-1.740	.283	1.118	1.189	.617	1.019	-2.935	-.154	32.60
19	0.085	.800	.823	-2.099	-.869	-1.088	-1.146	1.531	2.047	41.64

Table 8.4.3. Externally studentized residuals from univariate and multivariate regressions

CHAPTER 9

CONCLUSIONS AND SUGGESTIONS FOR FURTHER RESEARCH

The main purpose of this thesis has been to study the field of outlier detection methodology with multivariate data. This has involved detailed examination of the few existing methods, and the development of alternatives and of new methods for particular problems. The emphasis has been heavily on the multivariate normal distribution (and on the model of slippage in the mean); other multivariate distributions are not met very frequently in practice, while their treatment from the point of view of outlier detection is very difficult, as shown by Barnett's first steps.

"Existing methods" effectively means Wilks' test. Unfortunately, while the existence of accurate Bonferroni approximations to the percentage points makes this a good test for one outlier, the situation is less satisfactory for two or more (Chapter 3). Simulated percentage points can be obtained, so that the test becomes accurate for the situations covered in the simulation study. However, it is surely unusual for a hypothesis test for, say, two outliers versus none to be properly justified. It is for this reason that the procedure for sequential application of Wilks' statistic, developed here in Chapter 4, is valuable. It enables the choice of the number of outliers to be accounted for within the framework of the test.

Rohlf's test is an "existing method", but one that seems to be cited more than used. It avoids the question of the number of outliers, because it is not a test for a specific number. However the analysis here (Chapter 6) shows that Rohlf's test is not a good one. One objection is that the approximations suggested by Rohlf are very

inaccurate. This would not necessarily prevent its use as a graphical procedure, but the performance of the method in the presence of more than one outlier turns out to be poor, since it may be unable to declare the correct number of outliers in situations where it seems that any worthwhile method should not have any difficulty. The desirable modification would involve robust estimation of the sample covariance matrix. However, this can be a useful analysis in its own right, as in Campbell's method, and so Rohlf's method then appears to be redundant. There is no reason for Rohlf's method to continue to be suggested in the literature as a potential alternative to Wilks' test.

Wilks' statistic can be derived by a likelihood ratio analysis, among other methods. This is only one of the standard techniques for testing multivariate hypotheses. The other is union-intersection: application of this methodology is considered here in Chapter 5. The results of the study of the two-outlier problem confirm that the outlier test based on union-intersection can be more powerful than Wilks' test, depending on the configuration of the outlier slippages. However, the advantage usually lies with Wilks' statistic, so this will be more useful in general.

Likelihood ratio provides the basis for the tests suggested in Chapter 7, applicable to multivariate normal data where the covariance matrix has a specified structure. (Wilks' test makes use of no specification.) It is generally true in statistics that a more powerful analysis can be obtained by incorporating knowledge of this kind into the analysis. The results here show the considerable extent of this improvement in these particular problems. Consequently these are methods which can be recommended for use, so long as the assumption of the covariance structure is justified. The analysis of

residuals from the multivariate general linear model, considered in Chapter 8 here, is also an analysis of a structured problem (although in this case the structure refers to the mean rather than the covariance of the observations). The results here are useful because there will always be regression problems where it is necessary to take the multidimensional view.

Taken as a whole, the results in this thesis tend to confirm the value of Wilks' statistic - either in its conventional form or applied sequentially - for general use, since Rohlf's method appears to be unsatisfactory and the union-intersection alternative does not offer sufficient advantages to compensate for its greater computational complexity. However, the results also show that one can do considerably better than using Wilks' ordinary statistic in problems where some structure behind the multivariate normal covariance can be specified. One line for further research which can be suggested is therefore the development of outlier detection statistics for other multivariate structures which arise in practice. A related point is the development of influence measures for particular structures: this has been touched on in this thesis only in respect of influence in the multivariate general linear model.

Further needs for future research can be seen from the emphasis of this thesis on outlier detection in multivariate normal data. Little has been said about what to do once outliers have been found. Alternative actions suggested in the literature include use of an alternative model, such as a mixture model - which could be mixtures of normals - or an entirely different distribution (skew, or longer-tailed than the normal) in relation to which the supposed outliers no longer appear to be extreme. The difficulty in these options is the same as the difficulty of taking any other distribution than the multivariate

normal as the null distribution: that is, the analysis tends to be very difficult if not impossible. Some progress in the area of non-normal distributions would be desirable in extending the range of multivariate data problems which can be treated.

APPENDIX I : Newton-Raphson iteration

The Newton-Raphson method is a simple iterative procedure for the numerical solution of an equation. It can be applied to solve a system of equations for several unknowns, but the applications in this thesis are all for a single unknown so the corresponding details will be given here.

A simple geometrical illustration (Figure AI.1) explains the method; algebraically this is equivalent to a linear approximation from a Taylor series expansion. Suppose that

$$f(x)=0$$

is to be solved for x, and an initial guess is $x=x_0$. The slope of the curve $y=f(x)$ at x_0 is $f'(x_0)$, so that in triangle ABC in the figure

$$\tan \vartheta = f'(x_0)$$

where CB is tangent to $y=f(x)$ at $x=x_0$.

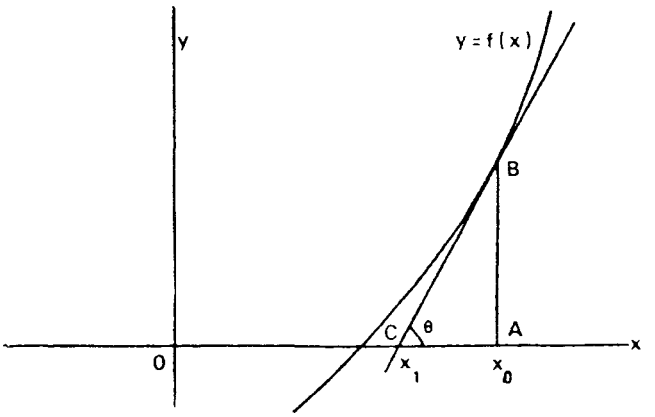


Figure A.I.1 Illustration of Newton-Raphson method

But from the sides of the same triangle

$$\tan \vartheta = \frac{f(x_0)}{x_0 - x_1}$$

where x_1 is the coordinate of point C where the tangent meets the x-axis. Equating these two expressions and re-arranging

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \qquad \text{(A.I.1)}$$

Now, as seen in the figure, x_1 is closer than x_0 to the solution. Repeated application of (A.I.1) will therefore, under certain conditions, provide a sequence of values converging towards the solution; in other words, the Newton-Raphson iteration scheme consists of applying the iterative scheme

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \qquad n=0,1,2,...$$

to generate a sequence x_0, x_1, ... converging to the solution of $f(x)=0$. In practice, the solution will be taken to be the first x_k for which

$$|f(x_k)| < \delta$$

where δ is a pre-determined constant, such as 10^{-9}.

In this thesis, the method was applied to obtain Bonferroni percentage points. This requires the solution of equations such as

$$g(r)=\alpha/n$$

where α is the desired Bonferroni significance level, n

the sample size and g a probabilty density function. The above method is then applied to solve f(r)=0 where

$$f(r)=g(r)-\alpha/n$$

APPENDIX II: The construction of slippages

II.1 General

In § 2.2, it was shown that Wilks' lamda statistic is a monotonic function of generalized distance. Specifically, equation (2.2.5) gave the relationship

$$\Lambda_n = 1 - \frac{n}{n-1}(x_n-\bar{x})'A^{-1}(x_n-\bar{x})$$

for testing point x_n; equation (2.2.6) then gave the equivalent expression

$$\Lambda_n^{-1}=1+T_n^2/(n-2)$$

where

$$T_n^2 \propto (x_n-\bar{x}_n)'A_n^{-1}(x_n-\bar{x}_n) \qquad \text{(II.1.1)}$$

with $\bar{x}_n$, A_n calculated after omitting x_n. Thus Λ_n is proportional to the generalized distance of the point being tested from the remainder of the sample.

This indicates that the appropriate way to generate a slippage of the mean in the multivariate problem is to fix the generalized distance of the slipped mean vector from the original. Therefore, given the hypotheses

$$H_0: \; x_i \sim N_p(\mu,\Sigma) \qquad i=1,\ldots,n$$

$$H_1: \; x_i \sim N_p(\mu,\Sigma) \qquad i\neq j$$
$$\quad x_j \sim N_p(\mu+a,\Sigma)$$

the slippage a should be defined so that

$$a'\Sigma^{-1}a = d^2 \qquad \text{(II.1.2)}$$

where d^2 is the desired squared distance. (Values of d^2=15 or 30 are used at most points of this thesis.) The

vector a can then be chosen in any convenient way to satisfy this equation; the closeness of results for different choices is illustrated in § II.3. For example, a simple choice for a is a constant α times the vector 1 consisting of ones. The necessary α is given by

$$\alpha^2=d^2/1'\Sigma^{-1}1$$

The quantity in the denominator is the sum of all the elements of Σ^{-1}. In the special case $\Sigma=\sigma^2 I$, this is $p\sigma^{-2}$, so that

$$\alpha=d\sigma/\sqrt{p}$$

This is the quantity which would be added to each element of, say, the last member of each simulated sample, in a power study with a single outlier.

Use of the result (II.1.2) makes it possible to ensure that equivalent slippages are being used, either when different directions of slippage are being used with the same Σ (as in the study of the union-intersection statistic in Chapter 6) or when different Σ's are being used. This facilitates the comparison of results. It is actually possible to obtain identical results in the simulations with different Σ's, as follows.

In the method of generating multivariate normal data used by the IMSL routine GGNSM employed in this research, suppose that vectors x from $N(0,\Sigma)$ are required. The first step is to construct the Cholesky factorization of Σ; that is, the lower triangular matrix L is obtained satisfying

$$\Sigma=LL'$$

Vectors z are then generated from the uncorrelated multivariate normal distribution $N(0,I)$, which requires only the same methods as are employed for the generation

of univariate normal deviates. Finally, the desired vectors x are obtained by transforming

$$x=Lz$$

since

$$V(x)=LV(z)L'=LIL'=LL'=\Sigma$$

Consequently, if vectors $x \sim N_p(0,\Sigma_1)$ and $y \sim N_p(0,\Sigma_2)$ are being generated, and the same seed is used in each sample so that the same sequence of vectors z from $N_p(0,I)$ is being used,

$$x=L_1 z$$

so that

$$z=L_1^{-1}x$$

but

$$y=L_2 z$$
$$=L_2 L_1^{-1}x$$

Hence if the slippage a is used in the sample of x vectors, then applying the slippage $L_2L_1^{-1}a$ to the corresponding vector in the sample of y vectors will mean that all details of the two samples are the same. For example, it is trivial to check that generalized distances as in (II.1.1) are identical.

The above result is illustrated below in § II.3 for the case of block Σ.

II.2 Slippages in the equicorrelation model

In the equicorrelation model, with all variances taken as equal to unity,

$$\Sigma=(1-\rho)I+\rho J$$

where J is the matrix whose every element is equal to unity. The inverse is

$$\Sigma^{-1}=(1-\rho)^{-1}[I-\rho\{1+(p-1)\rho\}^{-1}J]$$

so that a generalized distance is

$$\begin{aligned} d^2&=a'\Sigma^{-1}a \\ &=(1-\rho)^{-1}[a'a-\rho\{1+(p-1)\rho\}^{-1}a'Ja] \end{aligned}$$

This equation must be solved to find a suitable vector a for chosen d^2, in a simulation with given ρ. In the case where $a=\alpha 1$ will be used,

$$d^2=(1-\rho)^{-1}[p\alpha^2-\rho\{1+(p-1)\rho\}^{-1}\alpha^2p^2]$$

and this simplifies to give

$$\alpha=d\sqrt{\frac{\{1+(p-1)\rho\}}{p}}$$

as the quantity to be added to each element of a prespecified member of the sample.

II.3 Slippages in the block structure model

When Σ has block structure, so does its inverse and this simplifies the computation of slippages corresponding to a desired d^2 in (II.1.2). Special cases are employed in the simulations in Chapter 7:

a) the blocks have equicorrelation structure; and
b i) the slippages are equal in each component, $\beta 1$, or
b ii) equal in each component corresponding to the first block and zero for the remainder: $\lambda(1,1,\ldots,0,0,\ldots)'$.

As in II.1, in the case (b i)

$$\beta^2=d^2/1'\Sigma^{-1}1$$

where the denominator is the sum of all elements in Σ^{-1}, which is

$$p/\{1+(p-1)\rho\}$$

for an equicorrelation matrix. Hence for the case of two equicorrelated blocks, with dimensions p_1 and p_2 and correlations ρ_1 and ρ_2,

$$1'\Sigma^{-1}1 = \frac{p_1}{1+(p_1-1)\rho_1} + \frac{p_2}{1+(p_2-1)\rho_2}$$

For $p_1=p_2=2$

$$\beta=d\left\{\frac{2}{1+\rho_1}+\frac{2}{1+\rho_2}\right\}^{-1/2} \qquad \text{(II.3.1)}$$

(outlier type 1 in Table 7.3.1), and for $p_1=2$, $p_2=4$:

$$\beta=d\left\{\frac{2}{1+\rho_1}+\frac{4}{1+3\rho_2}\right\}^{-1/2}$$

(outlier type 1 in Table 7.3.2).

For the case (b ii), when the slippage is only in the components affecting the first block,

$$\lambda^2= d^2/1'\Sigma_{11}^{-1}1$$

so that

$$\lambda=d\left\{\frac{2}{1+\rho_1}\right\}^{-1/2} \qquad \text{(II.3.2)}$$

for the case $p_1=2$: this is outlier type 2 in Tables 7.3.1 and 7.3.2.

Some extra simulations were run for the block covariance structure, to illustrate the similarity of results obtained for different choices of a satisfying (II.1.2) with the same Σ. The result in § II.1 showing how to obtain identical results from different Σ's is also

applied here. The necessary calculations are easy to carry out, since the Cholesky factorization matrix also has block structure, and the factorization of $\Sigma=LL'$ of

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

is given by

$$\begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix}.$$

It can then easily be worked out that given the two block matrices

$$\Sigma_1 = \left(\begin{array}{cc|c} 1 & \rho_{11} & \\ \rho_{11} & 1 & \\ \hline & & \begin{array}{cc} 1 & \rho_{12} \\ \rho_{12} & 1 \end{array} \end{array}\right)$$

and

$$\Sigma_2 = \left(\begin{array}{cc|c} 1 & \rho_{21} & \\ \rho_{21} & 1 & \\ \hline & & \begin{array}{cc} 1 & \rho_{22} \\ \rho_{22} & 1 \end{array} \end{array}\right)$$

the matrix $L_2 L_1^{-1}$ is

$$\begin{pmatrix} 1 & 0 \\ \rho_{21}-\rho_{11}\sqrt{\frac{1-\rho_{21}^2}{1-\rho_{11}^2}} & \sqrt{\frac{1-\rho_{21}^2}{1-\rho_{11}^2}} \end{pmatrix} \Bigg| \begin{pmatrix} 1 & 0 \\ \rho_{22}-\rho_{12}\sqrt{\frac{1-\rho_{22}^2}{1-\rho_{12}^2}} & \sqrt{\frac{1-\rho_{22}^2}{1-\rho_{12}^2}} \end{pmatrix}$$

Some results are presented in Table II.3.1, for simulated powers of Wilks' ordinary statistic. Each line of the table is generated from 8000 simulations, starting from the same seed for the pseudorandom generator.

The slippage 3(1,1,1,1)' in the last line of the table corresponds to $d^2=30.857142$. The slippage in the penultimate line was computed using the Cholesky factorization to give identical results for different Σ. Other slippages a were then computed to satisfy $a'\Sigma^{-1}a=d^2$; in particular, the slippage in the first line corresponds to (II.3.1) and that in the third line to (II.3.2). It can be seen how small are the diferences in results between the different choices.

Table II.3.1 Simulated power of Wilks' statistic in the presence of one outlier: block covariance matrix with $p_1=p_2=2$, n=50. Slippage vector added to first member of the sample.

ρ_1	ρ_2	slippage vector	% of times outlier declared at level 1%	5%
.4	-.4	2.54558(1,1,1,1)'	69.5375	84.8625
.4	-.4	(3,3,1.8,2.74955)'	68.1500	84.2625
.4	-.4	3.04256(0,0,1,1)'	69.3375	84.5125
.4	-.4	(0,0,0,5.909167)'	68.9875	84.3625
.4	-.4	(3,3,3,1.54955)'	68.9125	84.6750
.4	0	3(1,1,1,1)'	68.9125	84.6750

BIBLIOGRAPHY

ANDERSON, T.W. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

ATKINSON, A.C. (1981). "Robustness, transformations and two graphical displays for outlying and influential observations in regression". *Biometrika,* **68**, 13-20.

BACON-SHONE, J. and FUNG, W.K. (1987). "A new graphical method for detecting single and multiple outliers in univariate and multivariate data." *Appl. Statist.*, **36**, 153-162.

BARNETT, V. (1976). "The ordering of multivariate data" (with Discussion). *J. Roy. Statist. Soc.*, A, **139**, 318-354.

BARNETT, V. (1979). "Some outlier tests for multivariate samples". *South African Statist. J.*, **13**, 29-52.

BARNETT, V. (1983). "Reduced distance measures and transformation in processing multivariate outliers". *Austral. J. Statist.*, **25**, 1-12.

BARNETT, V. and LEWIS, T. (1978). *Outliers in Statistical Data.* Wiley, Chichester.

BARNETT, V. and LEWIS, T. (1984). *Outliers in Statistical Data.* (2nd ed). Wiley, Chichester.

BECKMAN, R.J. and COOK, R.D. (1983). "Outlier.....s" (with Discussion). *Technometrics,* **25**, 119-163.

BERNARDO, J.M. (1976). "Algorithm AS103: Psi (digamma) function". *Appl. Statist.*, **25**, 315-317.

BOX, G.E.P. (1949). "A general distribution theory for a class of likelihood criteria". *Biometrika,* **36**, 317-346.

BOX, G.E.P. (1950). "Problems in the analysis of growth and wear curves". *Biometrics,* **6**, 362-389.

CAMPBELL, N.A. (1978). "The influence function as an aid in outlier detection in discriminant analysis". *Appl. Statist.,* **27**, 251-258.

CAMPBELL, N.A. (1980). "Robust procedures in multivariate analysis. I. Robust covariance estimation". *Appl. Statist.*, **29**, 231-237.

CARMINES, E.G. and ZELLER, R.A. (1979). *Reliability and Validity Assessment.* Sage, Beverly Hills.

CARONI, C. (1987). "Residuals and influence in the multivariate linear model". *The Statistician*, **36**, 365-370.

CHATFIELD, C. and COLLINS, A.J. (1980). *Introduction to Multivariate Analysis.* Chapman and Hall, London.

COLLETT, D. and LEWIS, T. (1976). "The subjective nature of outlier rejection procedures". *Appl. Statist.,* **25**, 228-237.

COOK, R.D. (1977). "Detection of influential observations in linear regression". *Technometrics,* **19**, 15-18.

COOK, R.D. and WEISBERG, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall, New York.

COX, D.R. and LEWIS, P.A.W. (1966). *The Statistical Analysis of Series of Events.* Methuen, London.

CRITCHLEY, F. (1985). "Influence in principal components analysis". *Biometrika,* **72**, 627-636.

DANIEL, C. (1960). "Locating outliers in factorial experiments". *Technometrics,* **2**, 149-156.

DAVID, H.A. (1981) *Order Statistics.* (2nd ed). Wiley, New York.

DAVID, H.A., HARTLEY, H.O. and PEARSON, E.S. (1954). "The distribution of the ratio, in a single normal sample, of range to standard deviation". *Biometrika*, **41**, 482-493.

DEVLIN, S.J., GNANADESIKAN, R. and KETTENRING, J.R. (1975). "Robust estimation and outlier detection with correlation coefficients". *Biometrika,* **62**, 531-545.

DIXON, W.J. (1950). "Analysis of extreme values". *Ann. Math. Statist.*, **21**, 488-506.

DIXON, W.J. (1951). "Ratios involving extreme values". *Ann. Math. Statist.*, **22**, 68-78.

FERGUSON, T.S. (1961). "On the rejection of outliers". *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* Vol. 1, 253-287.

FIELLER, N.R.J. (1976). *Some Problems Related to the Rejection of Outlying Observations.* Unpublished Ph.D. thesis, University of Hull.

FIELLER, N.R.J. (1989). "Outliers in multivariate data". Paper delivered at European Course in Advanced Statistics "Robustness in Statistics: Theory and Applications", Gunzburg, F.R.G.

FOULDS, G.A. (1976). *The Hierarchical Nature of Personal Illness.* Academic Press, London.

GENTLEMAN, J.F. and WILK, M.B. (1975). "Detecting outliers. II. Supplementing the direct analysis of residuals". *Biometrics,* **31**, 387-410.

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

GNANADESIKAN, R. and KETTENRING, J.R. (1972). "Robust estimates, residuals and outlier detection with multiresponse data." *Biometrics*, **28**, 81-124.

GOWER, J.C. and ROSS, G.J.S. (1969). "Minimum spanning trees and single linkage cluster analysis". *Appl. Statist.*, **18**, 54-64.

GRUBBS, F.E. (1950). "Sample criteria for testing outlying observations"., *Ann. Math. Statist.*, **21**, 27-58.

GRUBBS, F.E. (1969). "Procedures for detecting outlying observations in samples". *Technometrics,* **11**, 1-21.

GRUBBS, F.E. and BECK, G. (1972). "Extension of sample sizes and percentage points for significance tests of outlying observations". *Technometrics*, **14**, 847-854.

GUMBEL, E.J. (1960). "Bivariate exponential distributions". *J. Amer. Statist. Ass.*, **55**, 698-707.

GUPTA, S.S. (1960). "Order statistics from the gamma distribution". *Technometrics*, **2**, 243-262. Correction *Technometrics*, **2**, 523.

GUTTMAN, I. (1973). "Care and handling of univariate or multivariate outliers in detecting spuriosity - a Bayesian approach". *Technometrics*, **15**, 723-738.

HAWKINS, D.M. (1974). "The detection of outliers in multivariate data using principal components". *J. Amer. Statist. Ass.*, **69**, 340-344.

HAWKINS, D.M. (1980a). *Identification of Outliers*. Chapman and Hall, London.

HAWKINS, D.M. (1980b). "Critical values for identifying outliers". Letter to the Editor, *Appl. Statist.*, **29**, 95-96.

HEALY, M.J.R. (1968). "Multivariate normal plotting". *Appl. Statist.*, **17**, 157-161.

HECK, D.L. (1960). "Charts of some upper percentage points of the distribution of the largest characteristic root". *Ann. Math. Statist.*, **31**, 625-642.

HOSSAIN, A. and NAIK, D.N. (1989). "Detection of influential observations in multivariate regression". *J. Appl. Statist.*, **16**, 25-37.

IRWIN, J.O. (1925). "On a criterion for the rejection of outlying observations". *Biometrika*, **17**, 238-250.

JOHNSON, N.L. and KOTZ, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. Wiley, New York.

JOHNSON, A.J. and WICHERN, D.W. (1982). *Applied Multivariate Statistical Analysis.* Prentice-Hall, Englewood Cliffs, N.J.

JOLLIFFE, I.T. (1982). "A note on the use of principal components in regression". *Appl. Statist.*, **31**, 300-303.

KENDALL, M.G. AND BUCKLAND, W.R. (1960). *A Dictionary of Statistical Terms.* Oliver and Boyd, London.

KHATRI, C.G. (1972). "On the exact finite series distribution of the smallest or the largest root of matrices in three situations". *J. Mult. Anal.*, **2**, 201-207.

KIMBER, A.C. (1982). "Tests for many outliers in an exponential sample". *Appl. Statist.*, **31**, 263-271.

KIRCH, D.G., BIGELOW, L.B. and WYATT, R.J. (1985). "The interpretation of plasma haloperidol concentrations". *Arch. Gen. Psychiatry,* **42**, 838-839.

McMILLAN, R.G. (1971). "Tests for one or two outliers in normal samples with unknown variance". *Technometrics,* **13**, 87-100.

MARDIA, K.V. (1962). "Multivariate Pareto distributions". *Ann. Statist.*, **33**, 1008-1015.

MARDIA, K.V, KENT, J.T. and BIBBY, J.M. (1979). *Multivariate Analysis.* Academic Press, London.

MARONNA, R.A. (1976). "Robust M-estimators of multivariate location and scatter". *Ann. Statist.*, **1**, 51-67.

MATTHEWS, J.N.S. (1984). "Robust methods in the assessment of multivariate normality". *Appl. Statist.*, **33**, 272-277.

MORRISON, D.F. (1976). *Multivariate Statistical Methods.* (2nd ed). McGraw-Hill, New York.

MURPHY, R.B. (1951). *On Tests for Outlying Observations.* Ph.D. thesis, University of Princeton. University Microfilms, Ann Arbor.

PAPAKOSTAS, Y., MARKIANOS, M., PAPADIMITRIOU, G. and STEFANIS, C. (1986). "Prolactin response induced by ECT and TRH". *Br. J. Psychiatry,* **148**, 721-723.

PEARSON, E.S. and CHANDRA SEKAR, C. (1936). "The efficiency of statistical tools and a criterion for the rejection of outlying observations". *Biometrika,* **28**, 308-320.

PEARSON, E.S. and HARTLEY, H.O. (1972). *Biometrika Tables for Statisticians, Vol. 2.* Cambridge University Press, Cambridge.

PRESCOTT, P. (1978). "Examination of the behaviour of tests for outliers when more than one outlier is present". *Appl. Statist.*, **27**, 10-25.

PRESCOTT, P. (1979). "Critical values for a sequential test for many outliers". *Appl. Statist.*, **28**, 36-39.

PRESCOTT, P. (1980). Letter to the Editor. *Appl. Statist.*, **29**, 205.

RAO, C.R. (1951). "An asymptotic expansion of the distribution of Wilks' criterion". *Bull. Inst. Internal. Statist.*, **33**, 177-180.

RAO, C.R. (1973). *Linear Statistical Inference and its Applications.* Wiley, New York.

ROHLF, F.J. (1975). "Generalization of the gap test for the detection of multivariate outliers". *Biometrics*, **31**, 93-101.

ROHLF, F.J. (1977). Reply to correspondence. *Biometrics*, **33**, 763-765.

ROSNER, B. (1975). "On the detection of many outliers". *Technometrics,* **17**, 221-227.

ROSNER, B. (1977). "Percentage points of the RST many outlier procedure". *Technometrics,* **19**, 307-312.

ROSNER, B. (1983). "Percentage points for a generalized ESD many-outlier procedure". *Technometrics,* **25**, 165-172.

ROSS, G.J.S. (1969). "Algorithm AS 13: Minimum spanning tree". *Appl. Statist.*, **18**, 103-104.

ROUSSEEUW, P.J. (1989). "Unmasking multivariate outliers and leverage points". Paper delivered at European Course in Advanced Statistics "Robustness in Statistics: Theory and Applications", Gunzburg, F.R.G.

ROWELL, J.G. and WALTERS, D.E. (1976). "Analysing data with repeated observations on each experimental unit". *J. agric. Sci., Camb.*, **87**, 423-432.

ROY, S.N. (1957). *Some Aspects of Multivariate Analysis.* Wiley, New York.

ROYSTON, J.P. (1983). "Some techniques for assessing multivariate normality based on the Shapiro-Wilk *W*". *Appl. Statist.*, **32**, 121-133.

SCHWAGER, S.J. and MARGOLIN, B. (1982). "Detection of multivariate normal outliers". *Ann. Statist.*, **10**, 943-954.

SIOTANI, M. (1959). "The extreme value of the generalized distances of the individual points in the multivariate normal sample". *Ann. Inst. Statist. Math. Tokyo*, **10**, 183-208.

SMITH, R.C. (1985). Reply to correspondence. *Arch. Gen. Psychiatry,* **42**, 835-838.

SMITH, R.C., BAUMGARTNER, R., MISRA, C.H., MAULDIN, M., SHVARTSBURD, A., HO, B.T. and DEJOHN, C. (1984). "Haloperidol: plasma levels and prolactin response as predictors of clinical improvement in schizophrenia, chemical *v.* radioceptor plasma level assays". *Arch. Gen. Psychiatry,* **41**, 1044-1049.

SNYDMAN, D.R., MUNOZ, A. and WERNER, B.G. (1984). "A multivariate analysis of risk factors for hepatitis B virus infection among hospital employees screened for vaccination". *Amer. J. Epidemiol.*, **120**, 684-693.

THOMPSON, W.R. (1935). "On a criterion for the rejection of observations and the distribution of the ratio of the deviation to the sample standard deviation". *Ann. Math. Statist.*, **6**, 214-219.

TIETJEN, G.L. and MOORE, R.H. (1972). "Some Grubbs-type statistics for the detection of several outliers". *Technometrics*, **14**, 583-597.

TIKU, M.L. (1975). "A new statistic for testing suspected outliers". *Commun. Statist.* A, **4**, 737-752.

VAN PUTTEN, T., MARDER, S.R. and MINTZ, J. (1985). "Plasma haloperidol levels: clinical response and fancy mathematics". *Arch. Gen. Psychiatry*, **42**, 835.

VENABLES, W.N. (1975). "Algorithm AS77: Null distribution of the largest root statistic". *Appl. Statist.*, **24**, 458-465.

WILK, M.B., GNANADESIKAN, R. and HUYETT, M.J. (1962a). "Probability plots for the gamma distribution". *Technometrics*, **4**, 1-20.

WILK, M.B., GNANADESIKAN, R. and HUYETT, M.J. (1962b). "Estimation of the parameters of the gamma distribution using order statistics". *Biometrika*, **49**, 525-545.

WILKS, S.S. (1946). "Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution". *Ann. Math. Statist.*, **17**, 257-281.

WILKS, S.S. (1963). "Multivariate statistical outliers". *Sankhya*, A, **25**, 407-426.