Patricia Calder

Thesis Submitted for the Degree of

Doctor of Philosophy

1986

Influence Functions in Multivariate Analysis

Abstract

In this thesis we derive and apply influence functions for the detection of observations in multivariate analysis which when omitted from, or added to, the data lead to substantial changes in some aspect of our analysis. Emphasis is placed on the influence functions for the eigenvalues and eigenvectors in principal component analysis, from both the covariance and correlation matrices, and correspondence analysis. Also considered are the influence functions for the bivariate, multiple and partial correlation coefficients and the eigenvalues and eigenvectors in canonical correlation analysis.

We derive algebraic expressions, in terms of the original analysis, for the theoretical influence function in all cases and it is compared with the sample influence function when this has a 'simple' algebraic form. Only limited sample expressions can be derived for the changes in the eigenvalues and eigenvectors in principal component analysis and correspondence analysis, but the functions are compared numerically when applied to datasets. Problems in assessing the influence on eigenvectors when we have close eigenvalues, due to rotation within a relatively unchanged subspace, are highlighted in both principal component analysis and correspondence analysis and are discussed.

Acknowledgements

The work in this thesis was undertaken with a grant from the Science and Engineering Research Council.

I wish to thank Dr. I.T.Jolliffe for his considerable help and encouragement throughout the three years of the project.

Dedicated to my Mother and Father

Contents

Chapter 1: Introduction

This thesis is concerned with the derivation and application of measures of influence for observations in multivariate analysis. In particular, we concentrate on the affect of observations in principal component analysis and correspondence analysis, which both involve the calculation of eigenvalues and eigenvectors, and on three types of correlation coefficient.

An influential observation is one whose deletion from, or addition to, the dataset leads to an unusually 'large' change in some aspect of our analysis. Snedecor and Cochran (1967, p157) advised that one should check the affect of an outlier, i.e. an extreme or atypical observation, by comparing the results when the outlier is both included and omitted from the analysis. It is possible that an outlier need not be influential, and if we perform an influence analysis on all the observations in the dataset we can reveal the 'outliers that matter' without carrying out procedures for detecting outliers. This is particularly useful in multivariate analysis where the detection of outliers is difficult since an outlier need not reveal itself when we look at the variables individually or in pairs (for example, by looking at two dimensional plots of the variables). It is also possible for an influential observation not to be an outlier. If we find influential observations, we would not usually remove them completely from the analysis unless, for example, they were found to be recording errors. The influence analysis provides us with invaluable information on the reliability of our results and interpretations, and can further our understanding of the structure of the data we are analysing. It is for the above reasons that the influence techniques have received wide attention in the recent literature, although most applications have been confined to regression analysis.

Two comprehensive books have been written on the topic of influence in

regression analysis by Belsley, Kuh and Welsch (1980) and Cook and Weisberg (1982). Various statistics have been proposed for detecting influential observations depending on what part of the analysis one is interested in. For example, the Cook statistic is defined as

$$D_i = \frac{1}{p} r_i^2 \frac{h_{ii}}{1 - h_{ii}} \tag{1.1}$$

where r_i is the studentised residual, $h_{ii} = \underline{x}'_i (X'X)^{-1} \underline{x}_j$ is ith diagonal element of the hat matrix, X is the set of regressor variables and p is the number of regressor variables. The Cook statistic is a scalar measure of influence based on the vector of changes in the regression coefficients $\hat{\underline{\beta}}$ when the ith case is deleted. The Cook statistic can be written as,

$$D_i = \frac{(\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(-i)})'(X'X)(\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(-i)})}{p\,\hat{\sigma}^2} \tag{1.2}$$

where $\hat{\underline{\beta}}_{(-i)}$ is the vector of regression coefficients when the ith observation is removed. The choice of (1.2) to provide a scalar measure from the changes in the regression coefficients is that it can be linked to the confidence ellipsoid for $\hat{\underline{\beta}}$. Alternative ways of forming a scalar measure are discussed in Cook and Weisberg (1982,§ 3.5). We obtain (1.1) from (1.2) by deriving an expression for $\hat{\underline{\beta}}_{(-i)}$ involving the original $\hat{\underline{\beta}}$ and other terms from the analysis of the full dataset. This means we do not need to repeat our analysis a further n times corresponding to the deletion of each observation in turn. Beckman and Cook (1983) note of (1.1), ' it is clear that outlying cases (r_i^2 large) need not be influential if h_{ii} is sufficiently small.....Conversely, a non outlying case may be highly influential if h_{ii} is sufficiently large ', (we have a large h_{ii} if the X variables lie in a remote part of the factor space).

The sample influence curve for $\hat{\underline{\beta}}$ is defined as

$$SIC(\underline{x}, \hat{\underline{\beta}}) = (n-1)(\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(-i)}) \tag{1.3}$$

The reason for the multiple of (n-1) will be discussed below. We prefer to use a scalar measure like (1.1) to (1.4) so that we can rank our observations by their influence on all the coefficients simultaneously.

Multivariate statistical analyses such as principal component analysis (PCA) and correspondence analysis (CA) involve the calculation of eigenvalues and eigenvectors. Except for small p, the eigenvalues and eigenvectors do not have algebraic expressions and, even when they do, we find we cannot derive expressions for the perturbed sample eigenvalues and eigenvectors in terms of the original eigenvalues $\hat{\lambda}_k$ and eigenvectors $\underline{\hat{\alpha}}_k$. However, we can use the theoretical influence function defined by Hampel (1974) to give an asymptotic expression for the influence of points on our eigenvalues and eigenvectors involving only terms from the original problem. Substituting the sample equivalents, $\hat{\lambda}_k$ for λ_k and $\underline{\hat{\alpha}}_k$ for $\underline{\alpha}_k$, into the theoretical expressions enables us to use the theoretical influence curve to investigate influence in samples. This means we do not need to recalculate our eigenvalues and eigenvectors for each observation omitted as would be required for the sample influence curve. This will be discussed further below. The definition of the theoretical influence curve is as follows. Let $\underline{y}$ be a p-variate random vector with cumulative distribution function $F(\underline{y})$ and $\underline{\theta}$ is a vector of parameters which can be expressed as a functional of $F(\underline{y})$. If F is perturbed to become $(1-\epsilon)F + \epsilon\delta_{\underline{x}}$, where $\delta_{\underline{x}}$ is the cumulative distribution function of a random variable which can only take the single value $\underline{x}$, and $\bar{\underline{\theta}}$ is the corresponding perturbed parameter then,

$$TIC(\underline{x}, \underline{\theta}) = \lim_{\epsilon \downarrow 0} \frac{\bar{\underline{\theta}} - \underline{\theta}}{\epsilon} \qquad (1.4)$$

Under suitable conditions $\bar{\underline{\theta}}$ can be expanded in a Taylor series (see Rey (1983,§ 2.4) for further details) such that,

$$\underline{\tilde{\theta}} = \underline{\theta} + \epsilon \underline{c}_1 + \frac{\epsilon^2}{2} \underline{c}_2 + \ldots\ldots \tag{1.5}$$

The theoretical influence function for θ is thus $TIC(\underline{x}, \underline{\theta}) = \underline{c}_1$, the coefficient of ϵ in this expansion. Hence, only terms up to $o(\epsilon)$ need to be retained when calculating (1.4). This is an advantage in any situation where our parameter has a complicated expression. As desired for the sample influence curve discussed above, $TIC(\underline{x}, \underline{\theta})$ will only involve terms from the original analysis. The theoretical function is in fact a right-hand derivative and thus we can apply the product rule to give, for example,

$$TIC(\underline{x}, \underline{U}'\underline{V}) = TIC(\underline{x}, \underline{U}')\underline{V} + \underline{U}'TIC(\underline{x}, \underline{V}) \quad . \tag{1.6}$$

This can also be seen by using (1.5) and letting,

$$\underline{\tilde{U}}' = \underline{U}' + \epsilon \underline{U}_1' + o(\epsilon^2)$$

where $o(\epsilon^2)$ denotes terms of order ϵ^2 and smaller.

$$\underline{\tilde{V}} = \underline{V} + \epsilon \underline{V}_1 + o(\epsilon^2) \quad . \quad .$$

Multiplying these together gives,

$$\underline{\tilde{U}}'\underline{\tilde{V}} = \underline{U}'\underline{V} + \epsilon \left[\underline{U}'\underline{V}_1 + \underline{U}'_1\underline{V} \right] + o(\epsilon^2) \quad ,$$

and substituting into (1.4) gives (1.6).

Taking $\epsilon = -1/n-1$ and $F = \hat{F}$, where $\hat{F}$ is the empirical distribution function based on a set of observations $\underline{x}_1 \ldots\ldots \underline{x}_n$ we have

$$(1-\epsilon)\hat{F} + \epsilon \delta_{\underline{x}} = \hat{F}_{(-i)} \quad ,$$

the cumulative distribution function with the ith point missing. Substituting this into (1.4) we obtain the definition of the sample influence curve, when the ith point is removed, as

$$SIC_{(-i)}(\underline{x}_i, \hat{\underline{\theta}}) = (n-1)(\hat{\underline{\theta}} - \hat{\underline{\theta}}_{(-i)}) \quad . \tag{1.7}$$

Conversely, the sample influence curve for adding in an extra observation is defined as,

$$SIC_{(n+1)}(\underline{x}_{n+1}, \hat{\underline{\theta}}) = (n+1)(\hat{\underline{\theta}}_{(n+1)} - \hat{\underline{\theta}}) \quad , \tag{1.8}$$

where $\hat{\underline{\theta}}_{(n+1)}$ is the estimated parameter with the extra observation included. From expression (1.4) and (1.5) we can see that the theoretical influence curve provides a first order approximation to the sample curves when the sample equivalents are substituted in. There are two empirical influence curves that can be defined from the substitution of the sample equivalents into the theoretical expressions. The first is called the empirical influence curve and it is obtained by substituting the sample c.d.f., $\hat{F}$, for F in the influence curve. This is what we have described above and will be used throughout this thesis. However, from the definition of the theoretical influence function, when evaluated at the point $\underline{x}_j$ we are in fact considering the addition of $\underline{x}_j$ given it is already in the analysis. An informal approach throughout this thesis justifies the use of this empirical curve for estimating the deleted sample curve; as do the good comparisons it gives with this sample curve. The alternative empirical curve is called the deleted empirical curve and is obtained by substituting $\hat{F}_{(-i)}$ into the theoretical influence function. The perturbed distribution is thus $(1-\epsilon)\hat{F}_{(-i)} + \epsilon\delta_{\underline{x}} = \hat{F}$ so this curve measures the effect of adding in the point $\underline{x}_j$ given it is not initially in the analysis. This results in n different empirical curves as each will be expressed in terms of the n possible deleted datasets. We wish to avoid the calculation of the n separate analyses so we would need to write the terms from the deleted model in terms of those from the full analysis. However, this is difficult since it involves the parameters that we are wanting the influence expressions for. Critchley (1985) examines the two types of empirical curves for principal component analysis on the covariance matrix. If $EIC_{(-i)}$ denotes the deleted empirical curve for either the eigenvalues or eigenvectors he finds

$$EIC_{(-i)} = EIC_i + o(1/(n-1)) \quad ,$$

where EIC_i is the empirical curve evaluated at the ith case. The higher order

terms involve the second order terms c_2 (and higher) from the expansion in (1.5) which we may not wish to calculate. We find in many analyses that the empirical tends to underestimate the actual sample change and the deleted empirical overestimates. An example of this can be seen on page 212.

The theoretical influence function for the bivariate correlation coefficient was derived by Mallows in some unpublished work and used by Devlin, Gnanadesikan and Kettenring (1975) to detect outliers with respect to bivariate correlation. Chernick (1983) derived the theoretical influence function for the multiple correlation coefficient of y on (x_1, x_2). Campbell (1978) looked at influence functions in discriminant analysis, where it was applied to the Mahalanobis distance, the discriminant means and the discriminant function coefficients, when one of the distributions is perturbed. The theoretical influence function has been derived for a variety of statistics in multivariate analysis by Radhakrishnan and Kshirsagar (1981). This involves work on the eigenvalues and eigenvectors from a symmetric matrix, with application to PCA based on the covariance matrix. Critchley (1985) also derives theoretical influence functions for the eigenvalues and eigenvectors in PCA based on the covariance matrix. His expression for $TIC(x, \alpha_k)$ appears different to that of Radhakrishnan and Kshirsagar due to the different ways of dealing with the singular matrix $(\Sigma - \lambda_k I)$, see § 3.5 and § 3.6 for further details. Other 'influence' techniques in PCA which do not use the influence curves above have been examined by Krzanowski (1984) and Benassini (1985). There is also work by Escofier and Le Roux (1976), recorded in Greenacre (1984), which gives upper bounds for the angle between the original and perturbed eigenvectors when observations are omitted. This is mainly for correspondence analysis but it can be applied to other statistical methods that involve eigenvectors. Most of these references will be discussed

further in the relevant sections. The theoretical influence curve was originally derived for use in robust estimation. Huber (1981) says it is 'perhaps the most useful heuristic tool of robust statistics'. We wish our estimator to have a bounded influence curve, so that the effect of extreme points cannot exceed some value. This has lead to various proposals for robust estimators. The influence curve can also be used to derive the asymptotic variance of the estimators. See Huber (1981) for many applications of the influence curve in robust estimation.

In Chapter 2 we shall examine the influence functions for the bivariate, multiple and partial correlation coefficients. If possible it is preferable to obtain an expression for the sample influence curve. This chapter will show when this is possible and when it is better to use the theoretical influence function. When both sample and theoretical curves have an algebraic expression they will be compared, and for each correlation coefficient we have numerical comparisons of the deleted sample and empirical influence functions.

The theoretical influence functions for the eigenvalues, eigenvectors and component scores, in a principal component analysis, from both the covariance and correlation matrices are derived and discussed in Chapter 3. Contour plots of the influence functions are presented for $p = 2$ and $p = 3$. Multiple case deletion will also be considered. Chapter 4 is concerned with the practical application of the influence curves derived in Chapter 3. A suitable scalar measure for the change in the eigenvectors will be proposed and we will compare the 'actual sample change' with the empirical divided by $(n-1)$. The 'actual sample change' is found by numerically finding the perturbed eigenvalues and eigenvectors for each observation omitted (although we can make use of formulae for the change in the covariance matrix when a point is

omitted). A number of problems arise in principal component analysis when assessing the influence of observations, due to the eigenvectors switching in order or rotating within a relatively unchanged subspace. The theoretical and sample curves behave differently in such situations and this behaviour will be explained. Critical levels for the percentage change in an eigenvalue, for both the covariance and correlation matrix, are found by simulation and are discussed. Finally, we consider influence in detail for three datasets.

In Chapter 5 we derive the theoretical influence functions for the eigenvalues and eigenvectors in canonical correlation analysis and in correspondence analysis. Three types of perturbation are considered in correspondence analysis. The first is when we add a single observation so that a cell of a two way contingency table is incremented by one. The theoretical influence functions for this are found as special cases of the influence functions for canonical correlation analysis. Secondly, we consider the deletion of a row from a contingency table and lastly adding into cells for a multiway correspondence analysis. In Chapter 6 influence in correspondence analysis is examined by application to real datasets. Canonical correlation analysis is not discussed in Chapter 6 since most of the points which could be illustrated are the same as for other eigenvector methods. Furthermore, the first type of influence considered in correspondence analysis is a special case of canonical correlation analysis, and the multiple correlation coefficient, discussed in § 2.4 is another special case.

Chapter 7 discusses the usefulness of the different influence functions and influence in general, in light of the work in the thesis.

Chapter 2: Influence Functions for Correlation Coefficients

2.1. Introduction

We shall introduce the different influence functions, and the relationships between them, by looking at their applications to the covariance matrix and to three types of correlation coefficient. We shall obtain the sample and theoretical influence functions for the bivariate and multiple correlation coefficients. The theoretical expressions will be seen to be simplified versions of the sample curves. If we form the empirical curve by substituting the sample equivalents into the theoretical expression we can observe that the empirical will tend to underestimate the sample influence curve for omitting an observation, (which is what we are normally interested in), and over-estimate the sample influence curve for adding an observation. The theoretical influence function for the partial correlation coefficient is derived in § 2.5 and we shall discuss why it would be difficult to obtain a simple algebraic expression for the sample curve. We will usually consider both types of sample curves, and it will be shown how the empirical curve can be used to approximate either curve. For all the correlation coefficients we will numerically compare the deleted sample curve with the empirical.

2.2. Influence Functions for the Covariance Matrix

2.2.1. Sample Influence Function

Let $\tilde{S}$ denote the perturbed covariance matrix when we add an extra observation $\underline{x}_{n+1}$ then

$$\tilde{S} = \frac{n-1}{n} S + \frac{1}{n+1}(\underline{x}_{n+1} - \underline{\bar{x}})(\underline{x}_{n+1} - \underline{\bar{x}})' \qquad (2.2.1)$$

where $\underline{\bar{x}}$, S and n are the mean, covariance matrix and the sample size without the extra observation included.

Proof of (2.2.1)

Let $\underline{\bar{x}}^*$ be the mean with the extra observation included. Then,

$$n\tilde{S} = \sum_{j=1}^{n+1} (\underline{x}_j - \underline{\bar{x}}^*)(\underline{x}_j - \underline{\bar{x}}^*)'$$

$$= \sum_{j=1}^{n} \underline{x}_j \underline{x}_j' + \underline{x}_{n+1}\underline{x}_{n+1}' - (n+1)\underline{\bar{x}}^* \underline{\bar{x}}^{*\prime} \quad . \tag{2.2.2}$$

Substituting

$$\underline{\bar{x}}^* = \frac{n\,\underline{\bar{x}} + \underline{x}_{n+1}}{n+1}$$

into (2.2.2) we have

$$n\tilde{S} = \sum_{j=1}^{n} \underline{x}_j \underline{x}_j' + \frac{n}{n+1}\underline{x}_{n+1}\underline{x}_{n+1}' - \frac{n^2}{n+1}\underline{\bar{x}}\,\underline{\bar{x}}' - \frac{n}{n+1}\underline{\bar{x}}\,\underline{x}_{n+1}' - \frac{n}{n+1}\underline{x}_{n+1}\underline{\bar{x}}'$$

Putting in $n\,\underline{\bar{x}}\,\underline{\bar{x}}'$ and taking out again gives

$$n\tilde{S} = \sum_{j=1}^{n} \underline{x}_j \underline{x}_j' - n\,\underline{\bar{x}}\,\underline{\bar{x}}' + \frac{n}{n+1}(\underline{x}_{n+1}\underline{x}_{n+1}' + \underline{\bar{x}}\,\underline{\bar{x}}' - \underline{x}_{n+1}\underline{\bar{x}}' - \underline{\bar{x}}\,\underline{x}_{n+1}')$$

which is the same as (2.2.1). Subtracting the original covariance matrix and multipling by (n+1) gives the sample influence curve

$$SIC_{(n+1)}(S, \underline{x}_{n+1}) = -\frac{n+1}{n}S + (\underline{x}_{n+1} - \underline{\bar{x}})(\underline{x}_{n+1} - \underline{\bar{x}})' \quad . \tag{2.2.3}$$

If we wish to delete an observation $\underline{x}_i$, say, from the existing dataset then a proof similar to the above gives

$$S_{(-i)} = \frac{n-1}{n-2}S - \frac{n}{(n-1)(n-2)}(\underline{x}_i - \underline{\bar{x}})(\underline{x}_i - \underline{\bar{x}})' \tag{2.2.4}$$

and, noting that for the deleted sample curve we subtract the perturbed from the original and multiply by $(n-1)$, we have,

$$SIC_{(-i)}(S, \underline{x}_i) = -\frac{n-1}{n-2}S + \frac{n}{n-2}(\underline{x}_i - \underline{\bar{x}})(\underline{x}_i - \underline{\bar{x}})' \tag{2.2.5}$$

where $\underline{\bar{x}}$, S and n are based on the full dataset involving $\underline{x}_i$. Expressions (2.2.3) and (2.2.5) are very similar only differing in the functions of n and in the order of subtraction of the original and perturbed covariance matrices. It is due to the similarity of these expressions that we can convert the empirical

formulae based on adding points to deal with influence when a point is removed. We will return to this in § 2.2.2.

2.2.2. Theoretical and Empirical Influence Functions

The perturbed population covariance matrix is given by Campbell (1978) and results in the influence function

$$TIC(\underline{x},\Sigma) = -\Sigma + (\underline{x}-\underline{\mu})(\underline{x}-\underline{\mu})' \quad . \tag{2.2.6}$$

Substituting S for Σ and $\underline{\bar{x}}$ for $\underline{\mu}$ we obtain

$$EIC(\underline{x},S) = -S + (\underline{x}-\underline{\bar{x}})(\underline{x}-\underline{\bar{x}})' \tag{2.2.7}$$

which only differs from (2.2.3) or (2.2.5) in the appropriate functions of n. However, the functions of n in the sample curves are of $o(1)$ so we see that (2.2.7) would provide a good approximation to (2.2.3) or (2.2.5). This means that the theoretical expression could be useful in describing influence in samples when the an algebraic expression for the corresponding sample curve is not possible.

As pointed out by the external examiner, it would have been rather more appropriate to use Ω instead of S in much of what follows, where

$$\Omega = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})' = (n/(n-1))S \quad .$$

This makes little difference to most of the work but changes the functions of n involved slightly. See in particular § 4.4 where adjustments to the functions of n are considered.

In Chapter 1 we noted that the theoretical influence curve is an asymptotic result that provides a first order approximation to the sample results. Consequently, We would obtain the same expression as the empirical if we considered the sample curve as $n \to \infty$ as the higher order are not dependent upon n. This approach is used for some of the influence expressions in correspondence analysis. The sample influence for S and the empirical are perhaps two of closest expressions one will obtain. It is because the theoretical and sample expressions for the covariance matrix are so similar that the theoretical influence functions for the eigenvalues and eigenvectors in principal component analysis usually approximate the sample

...ite small datasets. See Chapter 4.

2.3. Influence Functions for the Bivariate Correlation Coefficient

2.3.1 Sample Influence Function

Using (2.2.1) and omitting the subscript $n+1$ so that x_k is the kth variable of the added observation $\underline{x}_{n+1}$ we have

$$\tilde{r}_{kj} = \frac{\frac{n-1}{n}s_{kj} + \frac{1}{n+1}(x_k - \bar{x}_k)(x_j - \bar{x}_j)}{\left[\frac{n-1}{n}s_{kk} + \frac{1}{n+1}(x_k - \bar{x}_k)^2\right]^{1/2}\left[\frac{n-1}{n}s_{jj} + \frac{1}{n+1}(x_j - \bar{x}_j)^2\right]^{1/2}}$$

$$= \frac{r_{kj} + \frac{n}{(n-1)(n+1)}y_k y_j}{\left[1 + \frac{n}{(n-1)(n+1)}y_k{}^2\right]^{1/2}\left[1 + \frac{n}{(n-1)(n+1)}y_j{}^2\right]^{1/2}}$$

where y_k is the standardised x_k variable. We cannot express the above perturbed correlation coefficient as the original correlation coefficient plus an extra term representing the change. However, subtracting the original r_{kj} and noting that $\frac{1}{(n+1)}SIC_{(n+1)}(\underline{x}, r_{kj})$ is the actual change in the correlation coefficient we have

$$\frac{1}{n+1}SIC_{(n+1)}(\underline{x}_{n+1}, r_{kj}) = \tag{2.3.1}$$

$$\frac{-r_{kj}\left[1 + \frac{n}{(n-1)(n+1)}y_k{}^2\right]^{1/2}\left[1 + \frac{n}{(n-1)(n+1)}y_j{}^2\right]^{1/2} + r_{kj} + \frac{n}{(n-1)(n+1)}y_k y_j}{\left[1 + \frac{n}{(n-1)(n+1)}y_k{}^2\right]^{1/2}\left[1 + \frac{n}{(n-1)(n+1)}y_j{}^2\right]^{1/2}}$$

This expression is rather cumbersome but it would not be very time consuming to calculate. The sample influence function when we delete an observation $\underline{x}_j$ is given by

$$\frac{1}{n-1}SIC_{(-i)}(\underline{x}_j, r_{kj}) = \tag{2.3.2}$$

$$\frac{r_{kj}\left[1 - \frac{n}{(n-1)^2}y_k{}^2\right]^{1/2}\left[1 - \frac{n}{(n-1)^2}y_j{}^2\right]^{1/2} - r_{kj} + \frac{n}{(n-1)^2}y_k y_j}{\left[1 - \frac{n}{(n-1)^2}y_k{}^2\right]^{1/2}\left[1 - \frac{n}{(n-1)^2}y_j{}^2\right]^{1/2}}$$

2.3.2. Theoretical and Empirical Influence Functions

The theoretical influence function $TIC(\underline{x}, \rho_{kj})$ is quoted and used by Devlin *et al* (1975) and is derived by Mallows in some unpublished work. A derivation of this influence function is given here, this proof serves to illustrate the duality between the sample and theoretical curves ($TIC(\underline{x}, \rho_{kj})$ could be derived as a special case of the partial correlation coefficient in § 2.5.2). From (2.2.6) the perturbed population covariance matrix is

$$\tilde{\Sigma} = (1-\epsilon)\Sigma + \epsilon(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})' + o(\epsilon^2) \qquad (2.3.3)$$

Using this we have

$$\tilde{\rho}_{kj} = \frac{\sigma_{kj} + \epsilon(-\sigma_{kj} + (x_k - \mu_k)(x_j - \mu_j))}{\sigma_{kk}^{1/2}\sigma_{jj}^{1/2}\left[1 + \frac{\epsilon}{\sigma_{kk}}(-\sigma_{kk} + (x_k - \mu_k)^2)\right]^{1/2}\left[1 + \frac{\epsilon}{\sigma_{jj}}(-\sigma_{jj} + (x_j - \mu_j)^2)\right]^{1/2}}$$

One advantage of looking at the theoretical curve is that terms of $o(\epsilon^2)$ disappear, see Chapter 1, and this means we can expand out our brackets to $o(\epsilon)$ which simplifies our expressions. It also enables us to write $\tilde{\rho}_{kj}$ in terms of the original correlation coefficient plus other terms which we could not do for the two sample curves. Noting that the expansion for

$$(1 + \epsilon b)^{-1/2} = 1 - \frac{\epsilon}{2}b + o(\epsilon^2)$$

and letting y_k be the kth standardised variable of $\underline{x}$ we have,

$$\tilde{\rho}_{kj} = \left[\rho_{kj} + \epsilon(-\rho_{kj} + y_k y_j)\right]\left[1 - \frac{\epsilon}{2}(-1 + y_k^2)\right]\left[1 - \frac{\epsilon}{2}(-1 + y_j^2)\right] + o(\epsilon^2) \quad .$$

Multiplying out the brackets gives

$$\tilde{\rho}_{kj} = \rho_{kj} - \epsilon\frac{\rho_{kj}}{2}(y_k^2 + y_j^2) + \epsilon y_k y_j + o(\epsilon^2) \quad . \qquad (2.3.4)$$

Hence,

$$TIC(\underline{x}, \rho_{kj}) = -\frac{\rho_{kj}}{2}(y_k^2 + y_j^2) + y_k y_j \quad . \qquad (2.3.5)$$

Devlin *et al* (1975) give a plot of (2.3.5) for $\rho = 0.5$ and we have a plot in § 3.8.3 for $\rho = 0.2$ since the eigenvalues for a 2×2 correlation matrix are $1 \pm \rho_{12}$

i.e. $TIC_R(\underline{x}, \lambda_k) = \pm TIC(\underline{x}, \rho_{kj})$. Fig 3.8.2 is a hyperbola and we find that the lines of zero change for ρ_{kj} make a smaller angle with the x axes as ρ_{kj} decreases. Thus, if we have uncorrelated data adding the standardised points $(y_k, 0)$ or $(0, y_j)$ will not change ρ_{kj}. Conversely, if we have perfectly correlated data then adding the point (y_k, y_j), where $y_k = y_j$, will not affect ρ_{kj}. This can be seen by substituting $\rho_{kj} = 1$ and $y_k = y_j$ into (2.3.5). These are the two extreme cases that the other values of ρ will lie between.

Expression (2.3.5) is much simpler than (2.3.1), although (2.3.1) would not take much longer to calculate than the empirical version of (2.3.5). The sample and theoretical proofs show that the empirical based on the theoretical result is a first order approximation to (2.3.1). If we considered the sample influence function, obtained by multiplying (2.3.1) through by $(n+1)$, as $n \to \infty$ and ignore terms of $o(1/n)$ we would be able to expand out the brackets as we did in the theoretical. This would result in the same expression as the empirical if we approximated $n/(n-1)$ to be 1. We ignore terms to $o(1/n)$ as we are considering the influence curve not the perturbed parameter. In fact the numerator of the sample curve would give us a similar expression to the empirical, as the terms in the denominator once expanded up do not enter into the expression since they are of a higher order. The same is true if we considered (2.3.2) as $n \to \infty$.

2.3.3. Practical Application of the Functions

The empirical curve based on (2.3.5) differs more in appearance to (2.3.1) or (2.3.2) than $EIC(\underline{x}, S)$ was to $SIC(\underline{x}, S)$. The more terms we have to expand out in our theoretical derivation the greater the two curves will differ. However, as n increases, since it is an asymptotic result, the closer the influences from the two curves will become. Below we will consider an example where $n = 55$ and we shall see the comparisons are good. We shall

compare the empirical with the sample change when we delete each observation in turn from the dataset. The empirical can be used equally well to approximate (2.3.1) or (2.3.2) as these to the first order only differ slightly in the functions of n involved. Although some of the signs are different this would not affect the asymptotic results when we let $n \to \infty$. We will consider the difference in the empirical approximation to the two sample curves below.

When we compare the empirical with deleted sample curve we will in fact compare the actual change in the parameter, in this case given by the L.H.S. of (2.3.2), with the empirical divided by $(n-1)$. This is equivalent to taking $\epsilon = 1/(n-1)$. When we use the empirical like this we will refer to it as the 'estimated change'. The data used below are taken from Barnett and Lewis (1984,p 262). It consists of two variables, the age and salary of electrical engineers and the bivariate correlation between the variables is 0.67. In Table 2.2.1 are the three most influential observations, ranked by their sample influence, recorded as the actual change, and the corresponding estimated change.

Table 2.2.1
Ranked Influences for the Sample Influence Function and Corresponding Empirical Value

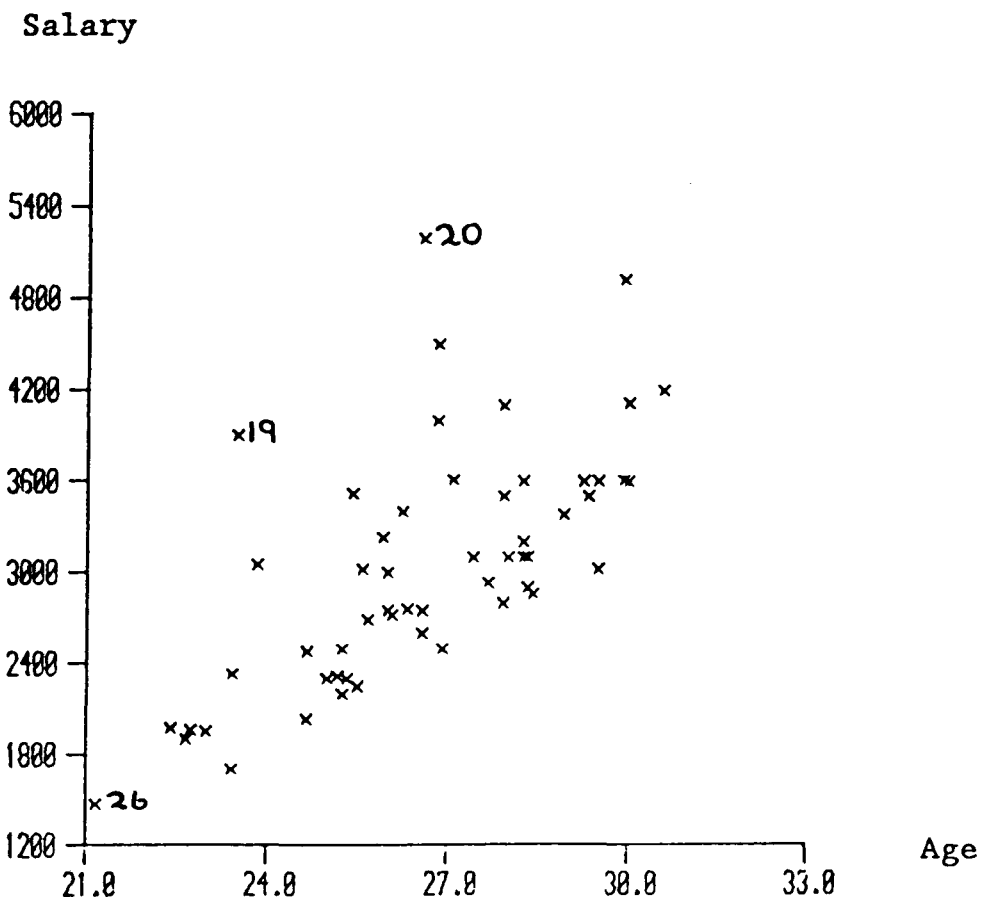
Obsn.	actual change	estimated change
20	-0.053	-0.046
19	-0.049	-0.046
26	0.031	0.027

As we subtract the perturbed parameter from the original a negative change means the parameter has increased and vice versa. Apart from two switches in rank for two observations whose influences were close, it was found, over the 55 observations, that the rankings by actual and estimated change were identical. From Table 2.2.1 we see that the estimated change consistently underestimates the sample change, in absolute value. We would similarly

find that the estimate for adding an extra observation would overestimate the actual sample change. This is the general rule in most applications of the curve we have examined, although it need not occur for every single observation. Why this occurs here can be understood from our sample and theoretical expressions. As discussed above we could obtain the empirical expression by considering the sample curve as $n \to \infty$. If we do this for the RHS of (2.3.2) multiplied by the $(n-1)$ and only expand out the brackets on the numerator the top would become the same as the empirical (apart from differing functions on n) but we would still have the denominator terms. For expression (2.3.2) the denominator terms are less than one leading to a larger value than the empirical for the sample change. A similar argument for (2.3.1) would leave us with the empirical divided by terms which are greater than 1.

We shall now consider what type of observations have come out as the most influential. The Mahalanobis distances for the three most influential observations are 13.7, 9.2 and 6.0 respectively. These values are given by Barnett and Lewis (1984) and using critical values they conclude observation 20 to be an outlier but say 'the status of L is more questionable' where L is observation 19. However, despite the differences in their Mahalanobis distances their affects on the bivariate correlation only differ by 0.004. Both increase the correlation when they are omitted, since from the plot of the data in Fig. 2.2.1 we can see both undermine the correlation. Observation 26 decreases the correlation when it is omitted since it is enhancing the correlation. The effect of observation 26 is smaller than those for observations 20 and 19 but it is possible that this is due as much to its positioning in the plot than because it has a smaller Mahalanobis distance. To examine this, observation 26 was multiplied through by 0.75 to make it smaller, and so

Figure 2.3.1 Plot of ages and salaries of 55 electrical engineers in the U.K in 1974



more extreme (at the bottom of the plot) than it was originally. The new Mahalanobis distances and sample changes for the resulting data were

Obs	Mah. Dist.	Sample Change
20	13.5	-0.050
19	8.0	-0.041
26	17.2	0.044

Observation 26 now has the largest Mahalanobis distance but its absolute influence is still smaller than that for observation 20. This is the first example to show that the most outlying point need not be the most influential. From the theoretical contour plot we see the affect of an observation would be zero no matter how far along the zero asymptote it was, at least in population case.

2.4. Influence Functions for the Squared Multiple Correlation Coefficient

2.4.1. Sample Influence Function

The squared multiple correlation coefficient is the squared correlation between $\underline{y}$ and the fitted $\underline{\hat{y}}$s in multiple regression. It is thus the proportion of the total sums of squares (TSS) of $\underline{y}$ explained by its regression on the set of X variables.

$$R^2 = \mathrm{Corr}^2(\underline{y}, X\hat{\beta}) = \frac{S_{yx} S_{xx}^{-1} S_{xy}}{s_{yy}} \qquad (2.4.1)$$

where X has a column of 1's for the constant term and where the covariance matrix

$$S = \begin{pmatrix} s_{yy} & S_{xy} \\ S_{yx} & S_{xx} \end{pmatrix} .$$

We can re-express

$$R^2 = \frac{REG(SS)}{TSS} = 1 - \frac{RSS}{TSS}$$

where *REG(SS)* and *RSS* are the regression and residual sums of squares respectively. We will consider the influence function for R^2 when we delete the ith observation from the full dataset since we can use existing deletion formulae, and deletion is what we are interested in practice. Then

$$TSS = (n-1)\text{var}(\underline{y}) = (n-1)s_{yy}$$

so from (2.2.4)

$$TSS_{(-i)} = TSS - \frac{n}{n-1}(y_i - \bar{y})^2 \quad .$$

From Belsley *et al* (1980, p64)

$$RSS_{(-i)} = RSS - \frac{e_i^2}{1-h_{ii}}$$

where e_i is the residual $(y_i - \hat{y}_i)$ and h_{ii} is the ith diagonal element of the hat matrix $X(X'X)^{-1}X'$, where X is centred. Hence,

$$R^2_{(-i)} = 1 - \frac{RSS_{(-i)}}{TSS_{(-i)}} = 1 - \frac{\left[RSS - \frac{e_i^2}{1-h_{ii}}\right]}{\left[TSS - \frac{n}{n-1}(y_i - \bar{y})^2\right]}$$

and

$$\begin{aligned}\frac{1}{(n-1)}SIC_{(-i)}(x_j, R^2) &= R^2 - R^2_{(-i)} \\ &= \frac{\left[RSS - \frac{e_i^2}{1-h_{ii}}\right]}{\left[TSS - \frac{n}{n-1}(y_i - \bar{y})^2\right]} - \frac{RSS}{TSS} \\ &= \frac{-\frac{e_i^2}{1-h_{ii}} + \frac{n}{n-1}(1-R^2)(y_i - \bar{y})^2}{\left[TSS - \frac{n}{n-1}(y_i - \bar{y})^2\right]} \quad .\end{aligned}$$

Multiplying each side by $(n-1)$ we have,

$$SIC_{(-i)}(x_j, R^2) = \frac{\frac{n}{n-1}(1-R^2)(y_i^s)^2 - \frac{e_i^2}{(1-h_{ii})s_{yy}}}{\left[1 - \frac{n(y_i^s)^2}{(n-1)^2}\right]} \qquad (2.4.2)$$

where $y_i^s = (y_i - \bar{y})/s_{yy}^{1/2}$ and $s_{yy} = TSS/(n-1)$. Hence, y_i^s is the standardised y variable w.r.t. the sample variance of the y_s, rather than by $\hat{\sigma}^2 = RSS/(n-p)$ which is the estimated variance of the y_s specified by the regression model.

2.4.2 Theoretical and Empirical Influence Functions

The population squared multiple correlation coefficient P^2 is defined as

$$P^2 = \frac{\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{yx}'}{\sigma_{yy}} \qquad (2.4.3)$$

and is the squared correlation between $\underline{y}$ and the linear combination of the Xs which has maximum correlation with $\underline{y}$.

Chernick (1983) derived the theoretical influence curve for P^2 for two X variables by writing it in terms of the correlation coefficients

$$P^2 = \frac{\rho_{yx_1}^2 + \rho_{yx_2}^2 - 2\rho_{yx_1}\rho_{yx_2}\rho_{x_1x_2}}{1 - \rho_{x_1x_2}^2}$$

This results in a long expression for the influence curve and he notes that one could derive P^2 for any number of variables by this approach but the number of parameters would increase rapidly and the formulae would not be useful in practice. Radhakrishnan and Kshirsagar (1981) also obtain the influence function for P^2 for any number of variables, as a special case of generalised variance. There appear to be typing errors in their expression and it is not expressed in the simplest possible form. Below is a derivation of $TIC(\underline{x}, P^2)$ and we will compare this with the sample curve in (2.4.2). The influence curve is also derived in § 5.2.3, where it comes out as a special case of the canonical correlations.

$$\Sigma = \begin{pmatrix} \sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix} .$$

Then using (2.3.3) the perturbed covariance matrix is

$$\begin{pmatrix} (1-\epsilon)\sigma_{yy}^2 + \epsilon(y-\mu_y)^2 & (1-\epsilon)\Sigma_{yx} + \epsilon(y-\mu_y)(\underline{x}-\underline{\mu}_x)' \\ (1-\epsilon)\Sigma_{yx}' + \epsilon(\underline{x}-\underline{\mu}_x)(y-\mu_y) & (1-\epsilon)\Sigma_{xx} + \epsilon(\underline{x}-\underline{\mu}_x)(\underline{x}-\underline{\mu}_x)' \end{pmatrix}$$

From Campbell (1978)

$$\tilde{\Sigma}_{xx}^{-1} = (1+\epsilon)\Sigma_{xx}^{-1} - \epsilon\Sigma_{xx}^{-1}(\underline{x}-\underline{\mu}_x)(\underline{x}-\underline{\mu}_x)'\Sigma_{xx}^{-1} + o(\epsilon^2) \quad . \qquad (2.4.4)$$

We wish the perturbed form of (2.4.3); therefore, by ignoring terms of order

$o(\epsilon^2)$

$$\tilde{\Sigma}_{yx}\tilde{\Sigma}_{xx}^{-1} = \Sigma_{yx}\Sigma_{xx}^{-1} + \epsilon\left[\Sigma_{yx}\left\{\Sigma_{xx}^{-1} - \Sigma_{xx}^{-1}(\underline{x} - \underline{\mu}_x)(\underline{x} - \underline{\mu}_x)'\Sigma_{xx}^{-1}\right\}\right.$$
$$\left. + \left\{-\Sigma_{yx} + (y - \mu_y)(\underline{x} - \underline{\mu}_x)'\right\}\Sigma_{xx}^{-1}\right]$$

$$= \Sigma_{yx}\Sigma_{xx}^{-1} + \epsilon\left[-\Sigma_{yx}\Sigma_{xx}^{-1}(\underline{x} - \underline{\mu}_x)(\underline{x} - \underline{\mu}_x)'\Sigma_{xx}^{-1}\right.$$
$$\left. + (y - \mu_y)(\underline{x} - \underline{\mu}_x)'\Sigma_{xx}^{-1}\right] \quad .$$

Similarly

$$\tilde{\Sigma}_{yx}\tilde{\Sigma}_{xx}^{-1}\tilde{\Sigma}_{yx}' = \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{yx}'$$
$$+ \epsilon\left[\left\{-\Sigma_{yx}\Sigma_{xx}^{-1}(\underline{x} - \underline{\mu}_x)(\underline{x} - \underline{\mu}_x)'\Sigma_{xx}^{-1} + (y - \mu_y)(\underline{x} - \underline{\mu}_x)'\Sigma_{xx}^{-1}\right\}\Sigma_{yx}'\right.$$
$$\left. + \Sigma_{yx}\Sigma_{xx}^{-1}\left\{-\Sigma_{yx}' + (\underline{x} - \underline{\mu}_x)(y - \mu_y)\right\}\right] \quad . \qquad (2.4.5)$$

Letting $\zeta = \Sigma_{yx}\Sigma_{xx}^{-1}(\underline{x} - \underline{\mu}_x) = \underline{\beta}'(\underline{x} - \underline{\mu}_x)$ be the point on the regression line for y, then (2.4.5) becomes

$$\tilde{\Sigma}_{yx}\tilde{\Sigma}_{xx}^{-1}\tilde{\Sigma}_{yx}' = \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{yx}' + \epsilon\left[-\zeta^2 + 2(y - \mu_y)\zeta - \sigma_{yy}P^2\right] \quad .$$

Noting that

$$\frac{1}{\tilde{\sigma}_{yy}} = \frac{\left[1 + \frac{\epsilon}{\sigma_{yy}}(-\sigma_{yy} + (y - \mu_y)^2)\right]^{-1}}{\sigma_{yy}}$$

$$= \frac{\left[1 - \epsilon(-1 + \frac{(y - \mu_y)^2}{\sigma_{yy}})\right]}{\sigma_{yy}}$$

and substituting this and (2.4.5) into the perturbed form of (2.4.3) we obtain after some tidying that

$$\tilde{P}^2 = P^2 + \epsilon\frac{\left[-P^2(y - \mu_y)^2 + 2(y - \mu_y)\zeta - \zeta^2\right]}{\sigma_{yy}} \quad . \qquad (2.4.6)$$

Adding in $\epsilon\frac{(y-\mu_y)^2}{\sigma_{yy}}$ and taking out again we obtain

$$TIC(\underline{x}, P^2) = (1-P^2)(y^s)^2 - \frac{\gamma^2}{\sigma_{yy}} \qquad (2.4.8)$$

where γ is the residual $(y-\mu_y-\zeta)$, and y^s is the standardised y variable. The empirical curve is obtained by substituting the sample equivalents R^2, and e_i for P^2, and γ respectively. This gives

$$EIC(\underline{x}, R^2) = (1-R^2)(y^s)^2 - \frac{e_i}{s_{yy}}$$

The empirical and sample curve given by (2.4.2) are thus quite similar. The residual for the sample influence curve has the divisor $(1-h_{ii})$ and so the sample and empirical may differ most when h_{ii} is large i.e. $\underline{x}_i$ is remote in the factor space (see Cook and Weisberg, 1980). Like the bivariate they differ in the denominator as this is expanded up for the theoretical. The denominator in (2.4.2) is smaller than one which should lead to the estimated change generally being smaller than the actual sample change.

2.4.3. Practical Application of the Functions

We will now apply the sample and empirical influence functions to a dataset taken from Cook and Weisberg (1982,p118-122). The data are concerned with the amount of drug (y) retained in the liver of rats. There are 19 observations (rats) and three regressor variables for the body weight and liver weight of the rats and the dose they are given. Cook and Weisberg note that none of the simple regressions of y on the individual Xs are significant but when y is regressed on all three variables together there are significant regression coefficients for X_1 and X_3. They gave plots of the studentised residuals, leverage (h_{ii}) values and the Cook statistic (all of which are defined in Chapter 1) against the observation numbers. The last two, but not the studentised residual, have a large value for the third observation and when it

is omitted there are no longer significant regression coefficients. They note that the influence of the case is due to this rat receiving a higher dose for its weight than the others, see Cook and Weisberg (1982) for further details.

Table 2.4.1 gives the most influential observations by the sample change on the multiple correlation coefficient and the corresponding estimated change from the empirical.

Table 2.4.1

Most Influential Observations on the Multiple Correlation Coefficient

Obs.	Actual Change	Estimated Change
3	0.343	0.224
1	-0.104	-0.076
19	-0.093	-0.059

The three most influential observations on R^2 are the same for the two curves but the empirical underestimates the sample (clearly, we would not wish to use the empirical in a practical situation when we have as simple an expression for the sample curve). The two curves disagree most for observations 5 and 13, where the sample has smaller values, (but larger absolute value for observation 13) than the empirical. These observations have the 2nd and 3rd largest leverages, after observation 3, which gives a larger negative coefficient multiplying the residual term in the sample than empirical curve.

However, observation 3 is the most influential for both curves. The original R^2 was 0.364 and when it is omitted R^2 falls to 0.021. The 2nd and 3rd most influential observations lead to an increase in the multiple correlation coefficient when they are omitted, as do most of the other observations. This probably occurs as observation 3 which is still in there receives more weight and so greater pull in the regression due to the decrease in n. If we had just relied on the studentised residual rather than calculating

the Cook influence measure, that looks at the changes in the regression coefficients, or examining the change in R^2, the important nature of observation 3 would not have been highlighted. The Cook statistic and the change in the multiple correlation coefficient are two different measures of influence. Although they both give observation 3 as the most influential they differ in their second and third rankings. When the two coincide and when they differ is not examined here.

2.5. Influence Functions for the Partial Correlation Coefficients

2.5.1 Sample Influence Function

The partial correlation between two variables $\underline{x}_k$ and $\underline{x}_j$ given the variables in X is the correlation of the two sets of residuals from regressing $\underline{x}_k$ and $\underline{x}_j$ individually on X. If we let s_{kj} be the covariance between $\underline{x}_k$ and $\underline{x}_j$ and S_{kx} be the vector of covariances between $\underline{x}_k$ and X and similarly S_{jx}, then the partial correlation is defined as

$$r_{kj.x} = \frac{s_{kj} - S_{kx} S_{xx}^{-1} S_{jx}'}{\left[s_{kk} - S_{kx} S_{xx} S_{kx}'\right]^{1/2} \left[s_{jj} - S_{jx} S_{xx} S_{jx}'\right]^{1/2}} \quad . \qquad (2.5.1)$$

This can be re-expressed as

$$r_{kj.x} = \frac{r_{kj} - S_{kx} S_{xx}^{-1} S_{jx}' / s_{kk}^{1/2} s_{jj}^{1/2}}{\left[1 - R_k^2\right]^{1/2} \left[1 - R_j^2\right]^{1/2}} \quad , \qquad (2.5.2)$$

where R_k^2 is the squared multiple correlation coefficient for the regression of $\underline{x}_k$ on the set in X. The sample influence curves for the bivariate and squared multiple correlation coefficients when a point is deleted from the dataset are given by (2.3.2) and (2.4.2) respectively. Substituting the perturbed forms for these parameters into the perturbed expression for (2.5.2) would yield a very uninformative sample expression for the perturbed partial correlation. The square roots in (2.5.2) make the mathematics intractable in the sample case as

we can see they did to a certain extent for the bivariate correlation.

2.5.2 Theoretical and Empirical Influence Functions

We will derive the theoretical influence function for $\rho_{kj.x}$ using the definition similar to (2.5.2) which is

$$\rho_{kj.x} = \frac{\rho_{kj} - \Sigma_{kx}\Sigma_{xx}^{-1}\Sigma_{jx}' / \sigma_{kk}^{1/2}\sigma_{jj}^{1/2}}{\left[1 - P_k^{\,2}\right]^{1/2}\left[1 - P_j^{\,2}\right]^{1/2}} \quad . \tag{2.5.3}$$

This enables us to use the perturbed form for the bivariate and squared multiple correlation coefficients in § 2.3.2 and § 2.4.2 respectively. Using equations (2.3.5) and (2.4.8) for the above terms we have

$$\tilde{\rho}_{kj} = \rho_{kj} - \epsilon\frac{\rho_{kj}}{2}(y_k^{\,2} + y_j^{\,2}) + \epsilon y_k y_j \tag{2.5.4}$$

where y_k is the standardised value of the kth variable in the added point; and

$$\tilde{P}_k^{\,2} = P_k^{\,2} + \epsilon\left[(1 - P_k^{\,2})y_k^2 - \frac{\gamma_k^2}{\sigma_{kk}}\right] \quad . \tag{2.5.5}$$

If we let

$$A = \frac{\Sigma_{kx}\Sigma_{xx}^{-1}\Sigma_{jx}'}{\sigma_{kk}^{1/2}\sigma_{jj}^{1/2}}$$

then similar algebra to that for squared multiple correlation coefficient in § 2.4.2 gives

$$\tilde{A} = A\left[1 - \frac{\epsilon}{2}(y_k^2 + y_j^2)\right] + \epsilon\left(-\frac{\gamma_k}{\sigma_{kk}^{1/2}}\frac{\gamma_j}{\sigma_{jj}^{1/2}} + y_j y_k\right) \tag{2.5.6}$$

where γ_k is the standardised residual from the regression of x_k on the variables in X. If we let B denote the numerator of (2.5.3) then using (2.5.4) and (2.5.6) gives

$$\tilde{B} = B\left[1 - \frac{\epsilon}{2}(y_k^2 + y_j^2)\right] + \epsilon\frac{\gamma_k}{\sigma_{kk}^{1/2}}\frac{\gamma_j}{\sigma_{jj}^{1/2}} \quad . \tag{2.5.7}$$

For the denominator,

$$\frac{1}{\left[1-\tilde{P}_k{}^2\right]^{1/2}} = \frac{1}{\left[1-P_k{}^2\right]^{1/2}}\left[1-\epsilon(y_k^2-\frac{\gamma_k^2}{\sigma_{kk}\left[1-P_k{}^2\right]})\right]^{-1/2}$$

Ignoring terms of $o(\epsilon^2)$ and letting $\nu_k = \frac{\gamma_k}{\sigma_{kk}^{1/2}\left[1-P_k{}^2\right]^{1/2}}$, which is the residual standardised w.r.t. the partial variance, we have

$$\frac{1}{\left[1-\tilde{P}_k{}^2\right]^{1/2}} = \frac{1}{\left[1-P_k{}^2\right]^{1/2}}\left[1+\frac{\epsilon}{2}(y_k^2-\nu_k^2)\right] \quad .$$

To $o(\epsilon)$

$$\frac{1}{\left[1-\tilde{P}_k{}^2\right]^{1/2}\left[1-\tilde{P}_j{}^2\right]^{1/2}} = \frac{\left[1+\frac{\epsilon}{2}(y_k^2+y_j^2-\nu_k^2-\nu_j^2)\right]}{\left[1-P_k{}^2\right]^{1/2}\left[1-P_j{}^2\right]^{1/2}} \quad . \tag{2.5.8}$$

Multiplying (2.5.7) and (2.5.8) together and performing the necessary steps results in,

$$TIC(\underline{x}\,,\rho_{kj.x}) = -\frac{\rho_{kj.x}}{2}\left[\nu_k^2+\nu_j^2\right]+\nu_k\nu_j \quad . \tag{2.5.9}$$

The empirical curve is then formed by substituting in the sample equivalents. The form of (2.5.9) is similar to that for (2.3.5) but the standardised x variables in (2.3.5) are replaced by the two sets of residuals standardised w.r.t. their partial variance. This result makes sense since the partial correlation coefficient is the usual bivariate correlation between the two sets of residuals. However, this means that the empirical does not take into account any changes in the individual unstandardised residuals when an observation is omitted from, or included in, the analysis.Instead, the residuals are treated like a fixed set of variables as in the usual bivariate correlation coefficient.

2.5.3. Practical Application of the Functions

We shall compare the sample and empirical curves for the partial correlation coefficient for two examples from the same dataset. The dataset

was collected by Dr. B.J.T. Morgan and consists of seven anatomical measurements made on statistic students at the University of Kent. The data were collected over three years, and we will examine the data from one of the years here. These data will be used throughout Chapter 4 as well. An influence analysis for the three different years, including when they are combined together, is discussed in a paper by Calder, Jolliffe and Morgan (1986). For the dataset discussed here, $n = 33$ and it contains one clear outlier, observation 30, in the 3rd and 7th variables which were the hand and wrist measurements respectively. It appeared that these readings may have been entered in the wrong order since if we swop them around the point is no longer an outlier. Fig. (2.5.1) is a plot of variables 3 and 7. We will consider $\rho_{37.4}$ and $\rho_{34.7}$ where variable 4 is the head circumference.

Table 2.5.1 gives the largest and smallest actual (sample) changes and corresponding estimated (empirical) changes in $\rho_{37.4}$ (the actual change = $1/(n-1)$ SIC and the estimated change = $1/(n-1)$ EIC). We will not always look at the smallest changes, but it does serve to show how much the two curves agree. In fact, we often find that they agree more for the smallest changes as the empirical tends to underestimate the largest changes the larger they get. Table 2.5.2 gives the largest and smallest changes for $\rho_{34.7}$. The sample and empirical disagree on the 2nd and 3rd rankings, with the sample placing observation 30 second. However, both agree that observation 27 is the most influential on $\rho_{34.7}$ and that observation 30 which is 'highly influential' on $\rho_{37.4}$ has comparatively little affect on $\rho_{34.7}$. Also given in the tables are the influences for the observations on the bivariate correlations ρ_{37} and ρ_{34} respectively. We will now consider briefly the differences in the most influential observations on $\rho_{37.4}$ and $\rho_{34.7}$ and how these differ with the usual bivariate correlations ρ_{37} and ρ_{34}. We may use the partial correlation

coefficient if we think the high correlation between two variables is due to their dependency on a third (or possibly more) variable. Since we are introducing another variable we can find those observations most influential on the partial and bivariate may differ due to an outlier in the variable being conditioned upon. However, if an observation has a similar structure to the rest of the data but is extreme on all variables then it may be influential on the bivariate correlation coefficient but not the partial since it maybe accounted for in the regressions and so has two small residuals. The values of the partial and bivariate correlations are given below the tables. The affect of observation 30 was so great on $\rho_{37.4}$ that when it was included the correlation was negative. The bivariate correlation was also exceedingly undermined. Observation 30 is thus highly influential on both the partial and bivariate correlation. This is not surprising as the discrepancy in observation 30, which is outlying only on variables 3 and 7 is not likely to be explained in the regression of these variables on another variable. Observation 30 is not so influential on $\rho_{34.7}$ but it is still influential on the bivariate correlation ρ_{34}. The bivariate correlation increases when it is omitted as the observation is undermining the correlation (note, a negative influence refers to an increase in the parameter when an observation is removed). However, it appears that taking account of variable 7, which is the other variable that 30 was discrepant on, has helped to reduce the affect of the outlier.

Table 2.5.1

Ranked Actual Changes and Corresponding Estimated Changes for $\rho_{37.4}$

obsn.	sample	empirical	biv(sample)
30	-0.83	-0.38	-0.66
27	0.08	0.07	0.03
17	0.05	0.04	0.05
"	"	"	"
"	"	"	"
21	0.00	0.00	0.00
20	0.00	0.00	0.00

$\rho_{37.4} = -0.18$ $\rho_{37} = 0.11$

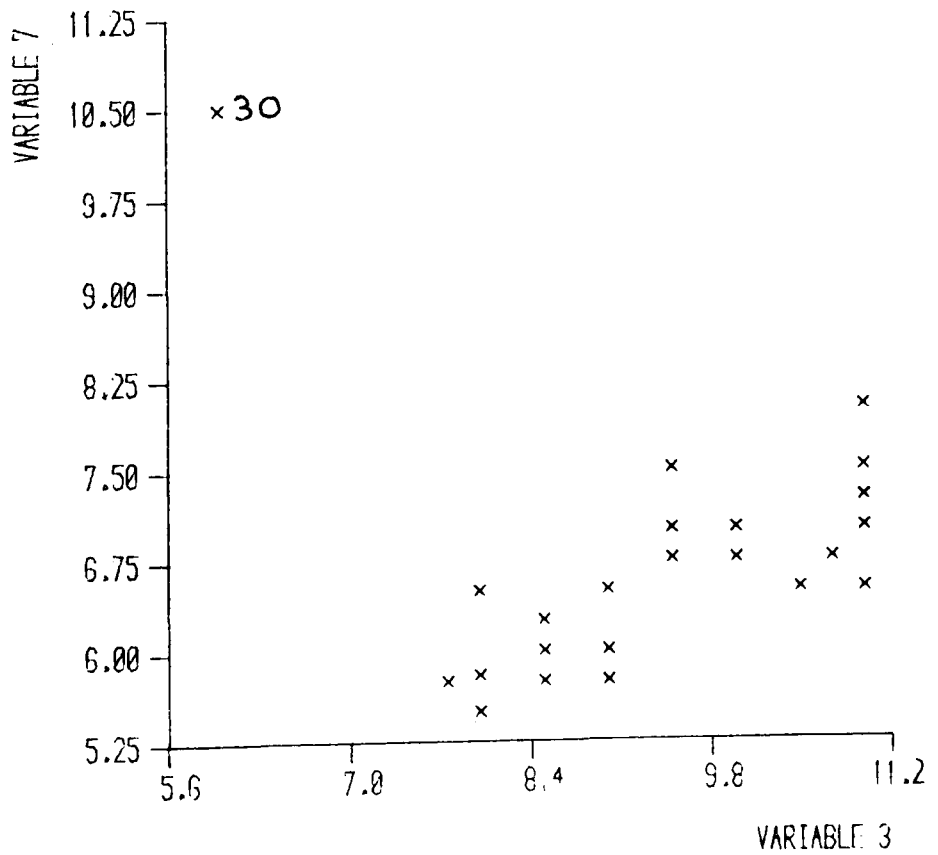
Table 2.5.2

Ranked Actual Changes and Corresponding Estimated Changes for $\rho_{34.7}$

obsn.	sample	empirical	biv(sample)
27	-0.05	-0.05	-0.03
30	0.05	0.03	-0.15
18	-0.04	-0.04	-0.03
"	"	"	"
"	"	"	"
13	0.00	0.00	0.02
6	0.00	0.00	0.00

$\rho_{34.7} = 0.54$ $\rho_{34} = 0.53$

Figure 2.5.1 Plot of variables 3 and 7 for student anotomical measurements



Chapter 3: Influence Functions in Principal Component Analysis

3.1. Introduction

Let the p vector random variable $\underline{x}$ come from a distribution with mean vector $\underline{\mu}$ and positive definite covariance matrix Σ. If the eigenvalues of Σ are $\lambda_1 > \lambda_2 > \ldots\ldots > \lambda_p$ with corresponding eigenvectors $\underline{\alpha}_1, \underline{\alpha}_2 \ldots\ldots \underline{\alpha}_p$ then we have the relationship,

$$(\Sigma - \lambda_k I)\underline{\alpha}_k = 0 \quad \text{where } \underline{\alpha}_k \neq 0 \tag{3.1.1}$$

A non-trivial solution to (3.1.1) holds only if,

$$|\Sigma - \lambda_k I| = 0$$

and expanding this determinant as a power series in λ_k we obtain the characteristic equation. The linear transformation

$$\underline{Z} = \Gamma'(\underline{x} - \underline{\mu}) \tag{3.1.2}$$

where $\Gamma = (\underline{\alpha}_1, \ldots\ldots \underline{\alpha}_p)$, forms new variables $Z_1, \ldots\ldots, Z_p$, called the principal components, such that the Z_k $k = 1, \ldots . p$ are uncorrelated with each other and,

$$\mathrm{Var}(Z_k) = \lambda_k \qquad k = 1, \ldots p \quad .$$

This means that the principal components have decreasing variance but

$$\sum_{k=1}^{p} \mathrm{Var}(Z_k) = \sum_{k=1}^{p} \lambda_k = \mathrm{tr}(\Sigma) = \sum_{k=1}^{p} \mathrm{Var}(x_k)$$

so the total variance remains the same. The k th element of $\underline{Z}$ is called the k th principal component score of the point $\underline{x}$.

Replacing Σ by the sample covariance matrix, S, we can define the sample principal components as,

$$\hat{\underline{Z}}_i = \hat{\Gamma}'(\underline{x}_j - \underline{\bar{x}}) \tag{3.1.3}$$

where $\hat{\Gamma} = (\hat{\underline{\alpha}}_1, \ldots\ldots, \hat{\underline{\alpha}}_p)$ and $\hat{\underline{\alpha}}_k$ is the k th eigenvector of S. The sample components have variances $\hat{\lambda}_1 > \ldots\ldots > \hat{\lambda}_p$ the sample covariance eigenvalues.

Principal components is thus a rotation of the original axes so that the

structure of the covariance matrix is simplified. If the first few principal components account for most of the variance in the original variables then we can reduce the dimensionality of our problem by only analysing the first few principal components. This does not mean that the principal components with small variances are never of interest. These can reveal constant relationships among our original variables, and they have been advocated for detecting multivariate outliers.

In practice it is often preferred to carry out our principal component analysis on standardised data to avoid the principal components being similar to the original variables when there are large differences in the variances. A principal component analysis on the covariance matrix of standardised variables is the same as the principal component analysis of the correlation matrix R. We will consider principal component analysis from both types of matrix. Further information on principal component analysis can be obtained from any good text book, see for example Mardia, Kent and Bibby (1979, Chapter 8).

In this chapter we will investigate the influence functions for the principal component variances, λ_k, the coefficients of the linear transformations, $\underline{\alpha}_k$, and finally the principal component scores for both the covariance and correlation matrices. The eigenvector, $\underline{\alpha}_k$ represents the contrast of the original variables accounting for the kth largest variance, λ_k. Thus, the $\underline{\alpha}_k$ are often interpretated especially for the largest and smallest λ_k. It is important to know how reliable our conclusions are and influence techniques provide this information by examining to what extent these contrasts change when points are added in or omitted. A study of influence on the eigenvalues λ_k is also important since these are used to determine how many principal components we should retain, and their relative sizes again

reveal the structure of the original variables. Most of the chapter is devoted to the theoretical influence function as only limited sample results can be derived. These sample results are considered first, in § 3.2. Even though algebraic expressions exist for the eigenvalues and eigenvectors, when p is small, we will see in § 3.3 and § 3.4 that they can only be used to derive the theoretical influence functions due to their complicated form. In § 3.5 we derive the influence functions for the eigenvalues and eigenvectors of a general symmetric matrix W. This is then applied to the covariance and correlation matrices respectively in § 3.6 and § 3.7, where the influence function for the component scores are also derived. In § 3.5 and § 3.6 we look in detail at the different ways one can express the influence function for the eigenvectors. Three approaches are considered. One method is developed in this thesis and the others were used by Radhakrishnan and Kshirsagar (1981) and Critchley (1985). The latter approach uses generalised inverses to deal with the singular matrix $(\Sigma - \lambda_k I)$ and only this method is used for the correlation eigenvectors as it leads to the simplest equations. In § 3.8 we examine and contrast the different influence functions from the covariance and correlation matrices. Contour plots for small p are presented in § 3.8.3. The section is concluded with a look at alternative influence approaches to principal component analysis by Escofier and Roux (1976) (which is outlined in Greenacre (1984)), the 'sensitivity' analysis of Krzanowski (1984) and work by Benasseni (1985) which gives bounds for the changes in the eigenvalues. In the final section we consider the theoretical influence functions when we add in more than one point. All practical applications and comparisons of the sample and empirical curves are considered in the next chapter.

3.2. Sample Influence when we Delete Specific Types of Observations in Covariance PCA

3.2.1. Observation Lying out along a Principal Component

Here we consider deleting an observation $\underline{x}_i$ whose score on the kth principal component is $\hat{Z}_{ki}$ and zero on all the other components. The principal component scores are formed from

$$\hat{\Gamma}'(\underline{x}_i - \underline{\bar{x}}) = \underline{\hat{Z}}_i$$

where $\hat{\Gamma}$ has the eigenvectors as columns. Taking $\hat{\Gamma}$ over the other side and noting that $\hat{\Gamma}'\hat{\Gamma} = I$ so that $\hat{\Gamma}^{-1} = \hat{\Gamma}'$ we have

$$(\underline{x}_i - \underline{\bar{x}}) = \hat{\Gamma}\begin{pmatrix} 0 \\ \vdots \\ \hat{Z}_{ki} \\ \vdots \\ 0 \end{pmatrix} = \underline{\hat{\alpha}}_k \hat{Z}_{ki} \quad .$$

Substituting this into (2.2.4) gives

$$S_{(-i)} = \frac{n-1}{n-2}S - \frac{n}{(n-1)(n-2)}\hat{Z}_{ki}^2 \underline{\hat{\alpha}}_k \underline{\hat{\alpha}}_k' \quad .$$

We can show that if $\underline{\hat{\alpha}}_j \;\; j = 1,....p$ is an eigenvector of S then it is also an eigenvector of $S_{(-i)}$.

$$S_{(-i)}\underline{\hat{\alpha}}_j = \left[\frac{n-1}{n-2}S - \frac{n}{(n-1)(n-2)}\hat{Z}_{ki}^2 \underline{\hat{\alpha}}_k \underline{\hat{\alpha}}_k'\right]\underline{\hat{\alpha}}_j$$

if $j \neq k$ then $\underline{\hat{\alpha}}_k'\underline{\hat{\alpha}}_j = 0$ and this gives

$$\begin{aligned} S_{(-i)}\underline{\hat{\alpha}}_j &= \frac{n-1}{n-2}S\underline{\hat{\alpha}}_j \\ &= \frac{n-1}{n-2}\hat{\lambda}_j\underline{\hat{\alpha}}_j \end{aligned}$$

Taking $\hat{\lambda}_{j(-i)} = \frac{n-1}{n-2}\hat{\lambda}_j$ gives

$$S_{(-i)}\underline{\hat{\alpha}}_j = \hat{\lambda}_{j(-i)}\underline{\hat{\alpha}}_j \quad ,$$

so $\underline{\hat{\alpha}}_j$ is an eigenvector of $S_{(-i)}$.

If $j = k$ then $\hat{\underline{\alpha}}_k' \hat{\underline{\alpha}}_k = 1$ so that

$$S_{(-i)}\hat{\underline{\alpha}}_k = \left(\frac{n-1}{n-2}S - \frac{n}{(n-1)(n-2)}\hat{Z}_{ki}^2 I\right)\hat{\underline{\alpha}}_k$$

Letting

$$\hat{\lambda}_{k(-i)} = \frac{n-1}{n-2}\hat{\lambda}_k - \frac{n}{(n-1)(n-2)}\hat{Z}_{ki}^2$$

gives

$$(S_{(-i)} - \hat{\lambda}_{k(-i)} I)\hat{\underline{\alpha}}_k = \frac{n-1}{n-2}(S - \hat{\lambda}_k I) = 0 \quad ,$$

so $\hat{\underline{\alpha}}_k$ is also an eigenvector of $S_{(-i)}$. Thus, when we delete an observation that lies along a principal component the eigenvectors remain unchanged but may switch in order according to the values of the perturbed eigenvalues. These were

$$\hat{\lambda}_{j(-i)} = \frac{n-1}{n-2}\hat{\lambda}_j \qquad j \neq k$$

$$\hat{\lambda}_{k(-i)} = \frac{n-1}{n-2}\hat{\lambda}_k - \frac{n}{(n-1)(n-2)}\hat{Z}_{ki}^2 \quad .$$

Hence,

$$(n-1)(\hat{\lambda}_j - \hat{\lambda}_{j(-i)}) = -\frac{n-1}{n-2}\hat{\lambda}_j \qquad j \neq k$$

$$(n-1)(\hat{\lambda}_k - \hat{\lambda}_{k(-i)}) = -\frac{n-1}{n-2}\hat{\lambda}_k + \frac{n}{n-2}\hat{Z}_{ki}^2 \quad . \tag{3.2.1}$$

We have not named the above as *SIC* since the changes are not for all possible values of $\underline{x}$. The sum of the perturbed eigenvalues is not the same as the sum of the original eigenvalues, (this needs to be the case for correlation eigenvalues). We will return to this point later. The same occurs if we add in an observation along an existing principal component with

$$(n+1)(\hat{\lambda}_{j(n+1)} - \hat{\lambda}_j) = -\frac{n+1}{n}\hat{\lambda}_j \qquad j \neq k$$

$$(n+1)(\hat{\lambda}_{k(n+1)} - \hat{\lambda}_k) = -\frac{n+1}{n}\hat{\lambda}_k + \hat{Z}_{k(n+1)}^2 \tag{3.2.2}$$

3.2.2. Observation lying in a Plane

The result in § 3.2.1 can be extended to show that if an observation has a zero score on the lth principal component then it will have no affect on the eigenvector, $\underline{\alpha}_l$, even though other eigenvectors may be changing. The change in the corresponding eigenvalue when the observation is removed is again

$$\hat{\lambda}_{l(-i)} = \frac{n-1}{n-2}\hat{\lambda}_l \quad .$$

We shall illustrate this by supposing $\underline{x}_i$ has non-zero principal component scores $\hat{Z}_{1i}$ and $\hat{Z}_{2i}$ and $\hat{Z}_{ji} = 0$ for $j = 3, 4, \ldots, p$. Then,

$$(\underline{x}_i - \underline{\bar{x}}) = \hat{Z}_{1i}\hat{\underline{\alpha}}_1 + \hat{Z}_{2i}\hat{\underline{\alpha}}_2$$

so,

$$S_{(-i)} = \frac{n-1}{n-2}S - \frac{n}{(n-1)(n-2)}\left[\hat{Z}_{1i}^2\hat{\underline{\alpha}}_1\hat{\underline{\alpha}}_1' + \hat{Z}_{1i}\hat{Z}_{2i}\hat{\underline{\alpha}}_1\hat{\underline{\alpha}}_2' + \hat{Z}_{1i}\hat{Z}_{2i}\hat{\underline{\alpha}}_2\hat{\underline{\alpha}}_1' + \hat{Z}_{2i}^2\hat{\underline{\alpha}}_2\hat{\underline{\alpha}}_2'\right] .$$

Thus, for any eigenvector $\hat{\underline{\alpha}}_j$ $j \neq 1,2$

$$S_{(-i)}\hat{\underline{\alpha}}_j = \frac{n-1}{n-2}S\hat{\underline{\alpha}}_j = \frac{n-1}{n-2}\hat{\lambda}_j\hat{\underline{\alpha}}_j$$

$$\Rightarrow \quad \alpha_{j(-i)} = \hat{\underline{\alpha}}_j \quad \text{and} \quad \lambda_{j(-i)} = \frac{n-1}{n-2}\hat{\lambda}_j \quad .$$

When $j = 1$ or 2 we find $\hat{\underline{\alpha}}_1$ and $\hat{\underline{\alpha}}_2$ are no longer the eigenvectors of $S_{(-i)}$. For example if $j = 1$

$$S_{(-i)}\hat{\underline{\alpha}}_1 = \frac{n-1}{n-2}S\hat{\underline{\alpha}}_1 - \frac{n}{(n-1)(n-2)}\left[\hat{Z}_{1i}^2\hat{\underline{\alpha}}_1 + \hat{Z}_{1i}\hat{Z}_{2i}\hat{\underline{\alpha}}_2\right]$$

which cannot be written in the form

$$(S_{(-i)} - \lambda_{1(-i)}I)\hat{\underline{\alpha}}_1 = 0.$$

Although these sample results are restrictive they do give us something to compare our theoretical (empirical) expressions with. These are derived in the next few sections.

3.3 Theoretical Influence Functions for the Eigenvalues and Eigenvectors from the Covariance Matrix for Small p

When $p = 2$ (or 3) the characteristic equation is a quadratic (or cubic) in λ. Solving these equations using the formulae for the roots of quadratic and cubic equations provide us with algebraic expressions for our eigenvalues. We also have expressions for the eigenvector coefficients. It is interesting to look at the theoretical proofs for $p = 2$ and 3, even though in § 2.5 we will derive the results for any p, since they will show that we could not obtain simple expressions for the sample curves even though the eigenvalues and eigenvectors have an algebraic form. This is due to the square and cubic roots that make the mathematics in the sample case intractable, as noted in the previous chapter on correlation coefficients. For the theoretical curve one can expand the roots and other terms out to $o(\epsilon)$ and ignore the higher order terms that go to zero in the definition.

We will go through the algebra for $p = 2$ in some detail but, due to the lengthy algebra needed, we shall only touch on the work for $p = 3$ in § 3.3.2

3.3.1. The Case for $p = 2$

The characteristic equation for the eigenvalues from the covariance matrix when $p = 2$ is

$$\lambda^2 - \lambda(\sigma_{11}+\sigma_{22}) + (\sigma_{11}\sigma_{22}-\sigma_{12}^2) = 0 \quad ,$$

where σ_{11} is the variance of the first variable and σ_{12} is the covariance between the two variables. This results in the formulae,

$$\lambda_1 = \frac{(\sigma_{11} + \sigma_{22}) + \sqrt{\sigma_{11}^2 + \sigma_{22}^2 - 2\sigma_{11}\sigma_{22} + 4\sigma_{12}^2}}{2} \qquad (3.3.1)$$

$$\lambda_2 = \frac{(\sigma_{11} + \sigma_{22}) - \sqrt{\sigma_{11}^2 + \sigma_{22}^2 - 2\sigma_{11}\sigma_{22} + 4\sigma_{12}^2}}{2} \quad .$$

We will only find the theoretical influence function for λ_1 and $\underline{\alpha}_1$ since the proofs for λ_2 and $\underline{\alpha}_2$ are almost identical, due to the similarity of their

expressions. We will re-express

$$\lambda_1 = \frac{B + \sqrt{C}}{2} \tag{3.3.2}$$

where

$$B = (\sigma_{11} + \sigma_{22}) \tag{3.3.3}$$

$$C = \sigma_{11}^2 + \sigma_{22}^2 - 2\sigma_{11}\sigma_{22} + 4\sigma_{12}^2 \tag{3.3.4}$$

The eigenvectors are obtained by solving, see (3.1.1.)

$$(\Sigma - \lambda_1 I)\underline{\alpha}_1 = 0 \tag{3.3.5}$$

Letting $\underline{a}$ denote the eigenvector whose first coefficient is 1, we obtain from (3.3.5), for $\sigma_{12} \neq 0$,

$$a_{12} = \frac{\lambda_1 - \sigma_{11}}{\sigma_{12}} \text{ or } a_{12} = \frac{\sigma_{12}}{\lambda_1 - \sigma_{22}} \quad , \tag{3.3.6}$$

according to whether one takes the first or second equation. The eigenvectors that are normalised such that $\underline{\alpha}'_1\underline{\alpha}_1 = 1$ can be written as

$$\alpha_{11} = \frac{1}{(1 + a_{12}^2)^{1/2}} \text{ and } \alpha_{12} = \frac{a_{12}}{(1 + a_{12}^2)^{1/2}} \quad . \tag{3.3.7}$$

Using (2.3.3) gives

$$\tilde{\sigma}_{11} = \sigma_{11} + \epsilon(-\sigma_{11} + (x_1 - \mu_1)^2) + o(\epsilon^2) \tag{3.3.8}$$
$$\tilde{\sigma}_{11}^2 = \sigma_{11}^2 + 2\epsilon(-\sigma_{11}^2 + \sigma_{11}(x_1 - \mu_1)^2) + o(\epsilon^2)$$

and similarly for $\tilde{\sigma}_{22}$ and $\tilde{\sigma}_{22}^2$.

$$\tilde{\sigma}_{11}\tilde{\sigma}_{22} = \sigma_{11}\sigma_{22} + \epsilon(-2\sigma_{11}\sigma_{22} + \sigma_{11}(x_2 - \mu_2)^2 + \sigma_{22}(x_1 - \mu_1)^2) + o(\epsilon^2)$$
$$\tilde{\sigma}_{12}^2 = \sigma_{12}^2 + 2\epsilon(-\sigma_{12}^2 + \sigma_{12}(x_1 - \mu_1)(x_2 - \mu_2)) + o(\epsilon^2) \tag{3.3.9}$$

Combining these together, as for the unperturbed C in (3.3.4), gives

$$\tilde{C} = C + \epsilon(-2C + 2D) + o(\epsilon^2)$$

where,

$$D = (\sigma_{11} - \sigma_{22})\left[(x_1 - \mu_1)^2 - (x_2 - \mu_2)^2\right] + 4\sigma_{12}(x_1 - \mu_1)(x_2 - \mu_2) \quad .$$

From (3.3.2) we are interested in $\sqrt{\tilde{C}}$ but for the theoretical influence function we can expand this out to $o(\epsilon)$ as we did for similar terms in the previous sections. This enables us to write $\tilde{\lambda}_1$ in terms of the original

eigenvalue which we could not do if we considering the sample influence function. Instead for the sample influence function we would be left with a complicated expression under the square root sign which would not be very informative.

Since,

$$\sqrt{\tilde{C}} = \sqrt{C}\sqrt{1 + \frac{2\epsilon}{C}(-C + D)} \quad + o(\epsilon^2)$$
$$= \sqrt{C} + \epsilon(-\sqrt{C} + \frac{D}{\sqrt{C}}) + o(\epsilon^2) \quad .$$

Using (3.3.3)

$$\tilde{B} = B + \epsilon(-B + (x_1 - \mu_1)^2 + (x_2 - \mu_2)^2) + o(\epsilon^2)$$

From (3.3.2)

$$\tilde{\lambda}_1 = \lambda_1 + \epsilon\left[-\lambda_1 + \frac{1}{2}\left(\frac{D}{\sqrt{C}} + (x_1 - \mu_1)^2 + (x_2 - \mu_2)^2\right)\right] + o(\epsilon^2)$$

Substituting in for D and $\sqrt{C}$ as $2\lambda_1 - (\sigma_{11} + \sigma_{22})$ gives, after some simple algebra,

$$TIC_V(\underline{x}, \lambda_1) =$$
$$-\lambda_1 + \frac{\left[(x_1 - \mu_1)^2(\lambda_1 - \sigma_{22}) + (x_2 - \mu_2)^2(\lambda_1 - \sigma_{11}) + 2\sigma_{12}(x_1 - \mu_1)(x_2 - \mu_2)\right]}{2\lambda_1 - (\sigma_{11} + \sigma_{22})} \quad .$$

However, from (3.3.6)

$$\frac{1}{\frac{2\lambda_1 - \sigma_{11} - \sigma_{22}}{\sigma_{12}}} = \frac{a_{12}}{1 + a_{12}^2}$$

so,

$$TIC_V(\underline{x}, \lambda_1) = -\lambda_1 + \frac{1}{1 + a_{12}^2}(x_1 - \mu_1)^2 + \frac{a_{12}^2}{1 + a_{12}^2}(x_2 - \mu_2)^2$$
$$+ \frac{2a_{12}}{1 + a_{12}^2}(x_1 - \mu_1)(x_2 - \mu_2) \quad .$$

and, using (3.3.7),

$$= -\lambda_1 + \alpha_{11}^2(x_1 - \mu_1)^2 + \alpha_{12}^2(x_2 - \mu_2)^2 + 2\alpha_{11}\alpha_{12}(x_1 - \mu_1)(x_2 - \mu_2)$$

$$= -\lambda_1 + Z_1^2 \quad , \tag{3.3.10}$$

where Z_1 is the score of $\underline{x}$ on the first principal component, and the subscript V denotes that it is for an eigenvalue from the covariance matrix. Similarly,

$$TIC_V(\underline{x}\,, \lambda_2) = -\lambda_2 + Z_2^2 \quad .$$

These results are very similar to the sample expressions in (3.2.1) or (3.2.2) which again shows that we can use the empirical for deletion or addition of points. The above indicates that the influence of a point on an eigenvalue depends only on its score for that principal component. We were only able to show a similar result in the sample case when an observation has a zero score on all the other principal components. In Chapter 4 we shall look at comparisons of the actual sample change and the estimated change based on the theoretical result, as well as the consideration of second order terms for the eigenvalues that do involve the other principal component scores for an observation.

We will now derive $TIC_V(\underline{x}\,, \underline{a}_1)$. Since a_{11} is defined to be unity, its influence function $TIC_V(\underline{x}\,, a_{11}) = 0$.

$$\tilde{a}_{12} = \frac{\tilde{\lambda}_1 - \tilde{\sigma}_{11}}{\tilde{\sigma}_{12}}$$

Using (2.3.3) and (3.3.10),

$$= \frac{a_{12} + \epsilon\left(-a_{12} + \dfrac{Z_1^2 - (x_1 - \mu_1)^2}{\sigma_{12}}\right)}{\left[1 + \dfrac{\epsilon}{\sigma_{12}}\left(-\sigma_{12} + (x_1 - \mu_1)(x_2 - \mu_2)\right)\right]} + o(\epsilon^2)$$

Expanding up the denominator and ignoring terms of $o(\epsilon^2)$, after simple algebra, results in

$$TIC_V(\underline{x}\,, a_{12}) = \frac{Z_1^2 - (x_1 - \mu_1)^2 - a_{12}(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_{12}} \quad .$$

Since we saw that the expression for λ_1 was simple when expressed in terms of

the principal components we will do the same here. If Γ contains the eigenvectors as columns then

$$(\underline{x} - \underline{\mu}) = \Gamma \underline{Z}$$

and so,

$$x_1 - \mu_1 = \alpha_{11}Z_1 + \alpha_{21}Z_2$$
$$x_2 - \mu_2 = \alpha_{12}Z_1 + \alpha_{22}Z_2 \quad .$$

Substituting these into $TIC_V(\underline{x}\ , a_{12})$ gives

$$TIC_V(\underline{x}\ , a_{12}) = \frac{1}{\sigma_{12}}\Big[(1 - \alpha_{11}^2 - a_{12}\alpha_{11}\alpha_{12})Z_1^2 - (\alpha_{21}^2 + a_{12}\alpha_{21}\alpha_{22})Z_2^2$$
$$- (2\alpha_{11}\alpha_{21} + \alpha_{11}\alpha_{22}a_{12} + \alpha_{21}\alpha_{12}a_{12})Z_1Z_2\Big] \quad .$$

From (3.3.7) $\alpha_{12} = a_{12}\alpha_{11}$. Using this and $\underline{\alpha}'_i\underline{\alpha}_i = 1$ and $\underline{\alpha}'_i\underline{\alpha}_j = 0$ then

$$1 - \alpha_{11}^2 - a_{12}\alpha_{11}\alpha_{12} = 1 - \alpha_{11}^2 - \alpha_{12}^2 = 0$$

$$\alpha_{21}^2 + a_{12}\alpha_{21}\alpha_{22} = \frac{\alpha_{21}}{\alpha_{11}}(\alpha_{21}\alpha_{11} + \alpha_{12}\alpha_{22}) = 0$$

$$2\alpha_{11}\alpha_{21} + \alpha_{11}\alpha_{22}a_{12} + \alpha_{21}\alpha_{12}a_{12} = \alpha_{11}(\alpha_{21} + \alpha_{22}a_{12}) + \frac{\alpha_{21}}{\alpha_{11}}(\alpha_{11}^2 + \alpha_{21}^2)$$
$$= \frac{\alpha_{21}}{\alpha_{11}} \quad .$$

Hence,

$$TIC_V(\underline{x}\ , a_{12}) = -\frac{Z_1Z_2}{\sigma_{12}} \times \frac{\alpha_{21}}{\alpha_{11}} \quad . \qquad (3.3.11)$$

We thus find that the influence function for a_{12} is a function of both principal component scores. If one eigenvector changes then the other must change also, since the eigenvectors must stay orthogonal,
there is no such restriction for the covariance eigenvalues which
may account for their differing form. The above influence gives a zero change in a_{12} if a point lies along either of the principal components, since Z_1 or Z_2 will be zero . This was noted for the sample results in § 3.2.

3.3.2. The Case for $p = 3$

The characteristic equation when $p = 3$ is

$$\lambda^3 - \lambda^2(\sigma_{11} + \sigma_{22} + \sigma_{33}) + \lambda(\sigma_{11}\sigma_{22} + \sigma_{11}\sigma_{33} + \sigma_{22}\sigma_{33} - \sigma_{12}^2 - \sigma_{13}^2 - \sigma_{23}^2)$$
$$- (\sigma_{11}\sigma_{22}\sigma_{33} - \sigma_{12}^2\sigma_{33} - \sigma_{13}^2\sigma_{22} - \sigma_{23}^2\sigma_{11} + 2\sigma_{12}\sigma_{13}\sigma_{23}) = 0 \quad .$$

If we re-express this as

$$\lambda^3 - \lambda^2 B + \lambda C + D = 0$$

then the roots of this equation are obtained by reducing to standard form

$$\gamma^3 + U\gamma + V = 0$$

where

$$\gamma = \lambda - \frac{B}{3}$$

$$U = -\frac{B}{3} + C$$

$$V = 2\left(\frac{B}{3}\right)^3 - \frac{BC}{3} + D \quad .$$

This gives the solutions,

$$\gamma_1 = E + F$$

$$\gamma_{23} = -\frac{E+F}{2} \pm i\frac{E-F}{2}\sqrt{3}$$

where,

$$E = \sqrt[3]{-\frac{U}{2} + \sqrt{Q}} \quad , \quad F = \sqrt[3]{-\frac{U}{2} - \sqrt{Q}}$$

$$Q = \left(\frac{U}{3}\right)^3 + \left(\frac{V}{2}\right)^2 \quad .$$

Using these formulae it is possible to work through, with similar algebra to the previous section, to obtain

$$TIC_V(\underline{x}, \lambda_k) = -\lambda_k + Z_k^2 , \qquad j = 1,2,3$$

and,

$$TIC_V(\underline{x}, a_{k2}) =$$

$$\frac{1}{\left[(\sigma_{22} - \lambda_k)(\sigma_{33} - \lambda_k) - \sigma_{23}^2\right]} \frac{Z_k}{\alpha_{k1}} \left\{ \sum_{\substack{t=1 \\ t \neq k}}^{p} \left[-(\sigma_{33} - \lambda_k)\alpha_{t2} + \sigma_{23}\alpha_{t3} \right] Z_t \right\}$$

$$TIC_V(\underline{x}\,, a_{k3}) =$$

$$\frac{1}{\left[(\sigma_{22}-\lambda_k)(\sigma_{33}-\lambda_k)-\sigma_{23}^2\right]}\frac{Z_k}{\alpha_{k1}}\left\{\sum_{\substack{t=1\\t\neq k}}^{p}\left[-(\sigma_{22}-\lambda_k)\alpha_{t2}+\sigma_{23}\alpha_{t3}\right]Z_t\right\} \quad .$$

These can be re-expressed as,

$$\begin{pmatrix} T_V(\underline{x}\,, a_{k2}) \\ T_V(\underline{x}\,, a_{k3}) \end{pmatrix} = -B^{-1}\frac{Z_k}{\alpha_{k1}}\begin{pmatrix} \sum_{t\neq k}\alpha_{t2}Z_t \\ \sum_{t\neq k}\alpha_{t3}Z_t \end{pmatrix} \quad . \qquad (3.3.12)$$

No proofs of the above expressions are given as the algebra is rather long and uninteresting. Again, a simple algebraic expression for the sample curve would not be possible due to the complicated nature of the expressions for the eigenvalues. The fact that one can obtain expressions using the theoretical shows what a useful tool it can be in providing some idea of what will be influential in the most complicated of situations.

3.4 Theoretical Influence Functions for the Eigenvalues and Eigenvectors from the Correlation Matrix for Small p

3.4.1. The Case for $p = 2$

The eigenvalues and eigenvectors from the 2×2 correlation matrix with $\rho_{12} \geq 0$ are,

$$\begin{aligned} \lambda_1 &= 1+\rho_{12} \qquad & \underline{\alpha}_1 &= (1/\sqrt{2} \quad 1/\sqrt{2}) \\ \lambda_2 &= 1-\rho_{12} \qquad & \underline{\alpha}_2 &= (1/\sqrt{2} \quad -1/\sqrt{2}) \quad . \end{aligned} \qquad (3.4.1)$$

If $\rho_{12} < 0$, then λ_1 and $\underline{\alpha}_1$ are interchanged with λ_2 and $\underline{\alpha}_2$. The perturbed λ_1, when $\rho_{12} > 0$, is

$$\begin{aligned} \tilde{\lambda}_1 &= 1+\tilde{\rho}_{12} \\ &= 1+\rho_{12}+\epsilon\, TIC(\underline{x}\,, \rho_{12}) + o(\epsilon^2) \quad . \end{aligned}$$

Thus,

$$TIC_R(\underline{x}\,, \lambda_1) = TIC(\underline{x}\,, \rho_{12}) \quad .$$

Similarly,

$$TIC_R(\underline{x}, \lambda_2) = -TIC(\underline{x}, \rho_{12}) \quad . \tag{3.4.2}$$

When $p = 2$ and $\rho_{12} > 0$, the influence function for the largest eigenvalue is the same as that for ρ_{12} and when $\rho_{12} < 0$ it is minus the influence function for the bivariate correlation. We have $TIC_R(\underline{x}, \lambda_1) = -TIC_R(\underline{x}, \lambda_2)$, and this reflects the fact that the sum of the two eigenvalues must equal 2. Thus, a change in one eigenvalue must be offset by an equal and opposite change in the other eigenvalue. This is different to the eigenvalues from a covariance matrix where the eigenvalues can change independently.

The eigenvectors from a 2×2 correlation matrix are given by (3.4.1) for all values of ρ_{12} . Hence the influence functions for the eigenvectors are zero, unless the eigenvectors swop in order. To do this the correlation must change sign when a point is added in. The influence functions for the correlation eigenvectors will not generally be zero for other values of p .

3.4.2. The Case for $p = 3$

The characteristic equation for the eigenvalues from a 3×3 correlation matrix is,

$$\lambda^3 - 3\lambda^2 + \lambda(3 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2) + (-1 + \rho_{12}^2 + \rho_{13}^2 + \rho_{23}^2 - 2\rho_{12}\rho_{13}\rho_{23}) = 0$$

Using the expressions for the roots of a cubic equations, as given in § 3.3.2, it is possible to arrive at the results which we will obtain for general p in § 3.7. However, the work required is very tedious and since we do not gain anything from looking at the algebra we will not discuss it further.

3.5 Theoretical Influence Functions for the Eigenvalues and Eigenvectors from a Symmetric Matrix W

Since we will need to find the influence functions for the eigenvalues and eigenvectors from a variety of symmetric matrices we will first derive the influence functions for a general symmetric matrix W.

Wilkinson (1965) quotes a result from Goursat (1933) which he uses to show that if λ_k is a simple eigenvalue of W and $\tilde{\lambda}_k$ is an eigenvalue of $W + \epsilon U$ then we can write

$$\tilde{\lambda}_k = \lambda_k + \epsilon c_1 + \epsilon^2 c_2 + \ldots\ldots$$

Similarly,

$$\tilde{\underline{\alpha}}_k = \underline{\alpha}_k + \epsilon \underline{d}_1 + \epsilon^2 \underline{d}_2 + \ldots\ldots$$

When $W + \epsilon U$ is $W + \epsilon TIC(\underline{x}, W)$ we can write

$$\tilde{\lambda}_k = \lambda_k + \epsilon TIC(\underline{x}, \lambda_k) + o(\epsilon^2)$$

$$\tilde{\underline{\alpha}}_k = \underline{\alpha}_k + \epsilon TIC(\underline{x}, \underline{\alpha}_k) + o(\epsilon^2) \quad .$$

We saw this was true in our 2×2 results.

3.5.1. The Influence Function for the Eigenvalues

We have the relationship

$$(\tilde{W} - \tilde{\lambda}_k I)\tilde{\underline{\alpha}}_k = 0 \quad . \tag{3.5.1}$$

We will use the eigenvectors $\underline{\alpha}_k$ which are normalised such that $\underline{\alpha}_k' \underline{\alpha}_k = 1$ and $\underline{\alpha}_k' \underline{\alpha}_j = 0$. Considering only terms up to $o(\epsilon^2)$ we have

$$\left[W - \lambda_k I + \epsilon(TIC(\underline{x}, W) - TIC(\underline{x}, \lambda_k) I)\right]\left[\underline{\alpha}_k + \epsilon TIC(\underline{x}, \underline{\alpha}_k)\right] = 0 \quad .$$

The term in ϵ is,

$$(W - \lambda_k I) TIC(\underline{x}, \underline{\alpha}_k) + (TIC(\underline{x}, W) - TIC(\underline{x}, \lambda_k) I)\underline{\alpha}_k = 0 \tag{3.5.2}$$

This expression involves both unknowns $TIC(\underline{x}, \lambda_k)$ and $TIC(\underline{x}, \underline{\alpha}_k)$ but pre-multiplying it through by $\underline{\alpha}_k'$ will remove the term in $TIC(\underline{x}, \underline{\alpha}_k)$ giving

$$\underline{\alpha}_k' TIC(\underline{x}, \lambda_k)\underline{\alpha}_k = \underline{\alpha}_k' TIC(\underline{x}, W)\underline{\alpha}_k$$

and since $TIC(\underline{x}, \lambda_k)$ is a constant the influence function for the eigenvalues is

$$TIC(\underline{x}, \lambda_k) = \underline{\alpha}_k' TIC(\underline{x}, W)\underline{\alpha}_k \quad . \tag{3.5.3}$$

We would arrive at the same expression (3.5.3) if we had used the eigenvectors $\underline{a}_k$ which have the alternative normalisation of the first coefficient set to one in (2.5.1). In the above proof we would have arrived at,

$$TIC(\underline{x}, \lambda_k)\underline{a}_k'\underline{a}_k = \underline{a}_k' TIC(\underline{x}, W)\underline{a}_k$$

and since $\underline{\alpha}_k = \dfrac{\underline{a}_k}{(\underline{a}_k'\underline{a}_k)^{1/2}}$ this also gives (3.5.3).

3.5.2 The Influence Function for the Eigenvectors

Having obtained $TIC(\underline{x}, \lambda_k)$ we can return to (3.5.2) to get $TIC(\underline{x}, \underline{\alpha}_k)$ which is now the only unknown. Re-arranging (3.5.2) gives,

$$(W - \lambda_k I)TIC(\underline{x}, \underline{\alpha}_k) = \left[TIC(\underline{x}, \lambda_k) I - TIC(\underline{x}, W)\right]\underline{\alpha}_k \quad . \tag{3.5.4}$$

$(W - \lambda_k I)$ is singular, as one of the conditions of eigenanalysis is that $|W - \lambda_k I| = 0$, so we cannot obtain $TIC(\underline{x}, \underline{\alpha}_k)$ on its own on the LHS by inverting $(W - \lambda_k I)$. Several options arise.

The results for $p = 2$ and 3 suggest that we can convert our problem to looking at $TIC(\underline{x}, \underline{a}_k)$, and we can obtain the influence function for the $(p - 1)$ non-zero coefficients (we will assume the first coefficient is set to zero). The influence function for the non-zero coefficients is the given by inverting $B = (W - \lambda_k I)^*$ which is $(W - \lambda_k I)$ with the first row and column removed. We can then obtain the influence function for the alternative normalised eigenvectors by noting

$$\tilde{\underline{\alpha}}_k = \frac{\tilde{\underline{a}}_k}{(\tilde{\underline{a}}_k'\tilde{\underline{a}}_k)^{1/2}}$$

Substituting $\tilde{\underline{a}}_k = \underline{a}_k + \epsilon TIC(\underline{x}, \underline{a}_k) + o(\epsilon^2)$ and ignoring terms of $o(\epsilon^2)$,

$$(\tilde{\underline{a}}_k \tilde{\underline{a}}_k) = \underline{a}_k'\underline{a}_k + 2\epsilon \underline{a}_k' TIC(\underline{x}, \underline{a}_k) \quad ,$$

so,

$$\underline{\tilde{\alpha}}_k = \frac{\left[\underline{a}_k + \epsilon TIC(\underline{x}, \underline{a}_k)\right]}{(\underline{a}_k' \underline{a}_k)^{1/2}} \left[1 + \frac{2\epsilon \underline{a}_k' TIC(\underline{x}, \underline{a}_k)}{\underline{a}_k' \underline{a}_k}\right]^{-1/2}$$

$$= \underline{\alpha}_k - \epsilon \underline{\alpha}_k \underline{\alpha}_k' \frac{TIC(\underline{x}, \underline{a}_k)}{(\underline{a}_k' \underline{a}_k)^{1/2}} + \epsilon \frac{TIC(\underline{x}, \underline{a}_k)}{(\underline{a}_k' \underline{a}_k)^{1/2}}$$

and so, noting $\alpha_{k1} = 1/(\underline{a}_k' \underline{a}_k)^{1/2}$

$$TIC(\underline{x}, \underline{\alpha}_k) = \alpha_{k1}\left[I - \underline{\alpha}_k \underline{\alpha}_k'\right] TIC(\underline{x}, \underline{a}_k) \qquad (3.5.5)$$

An alternative approach is given by Radhakrishnan and Kshirsagar (1981). They make $(W - \lambda_k I)$ non-singular by adding in $\underline{\alpha}_k \underline{\alpha}_k'$ and so,

$$TIC(\underline{x}, \underline{\alpha}_k) = (W - \lambda_k I + \underline{\alpha}_k \underline{\alpha}_k')^{-1}\left[TIC(\underline{x}, \lambda_k) I - TIC(\underline{x}, W)\right]\underline{\alpha}_k \qquad (3.5.6)$$

This can be done as,

$$\underline{\alpha}_k' TIC(\underline{x}, \underline{\alpha}_k) = 0 \quad , \qquad (3.5.7)$$

which can be seen from our normalisation constraints which are

$$\underline{\alpha}_k' \underline{\alpha}_k = 1, \qquad \underline{\tilde{\alpha}}_k' \underline{\tilde{\alpha}}_k = 1. \qquad (3.5.8)$$

To $o(\epsilon)$,

$$\underline{\tilde{\alpha}}_k' \underline{\tilde{\alpha}}_k = \underline{\alpha}_k' \underline{\alpha}_k + 2\epsilon \underline{\alpha}_k' TIC(\underline{x}, \underline{\alpha}_k)$$

but for both expressions in (3.5.8) to hold the term in ϵ above must be zero, which is (3.5.7).

The third and final alternative considered here uses generalised inverses and was first presented by Sibson (1979). We multiply each side of (3.5.4) by the generalised inverse of $(W - \lambda_k I)$, which we will denote by $(W - \lambda_k I)^+$, to give

$$(W - \lambda_k I)^+ (W - \lambda_k I) TIC(\underline{x}, \underline{\alpha}_k) = (W - \lambda_k I)^+ \left[TIC(\underline{x}, \lambda_k) I - TIC(\underline{x}, W)\right]\underline{\alpha}_k \quad .$$

Taking the generalised inverse as

$$(W - \lambda_k I)^+ = \sum_{\substack{j=1 \\ j \neq k}}^{p} \underline{\alpha}_j (\lambda_j - \lambda_k)^{-1} \underline{\alpha}_j'$$

and writing

$$(W-\lambda_k I)=\sum_{\substack{t=1\\ t\neq k}}^{p}\underline{\alpha}_t(\lambda_t-\lambda_k)\underline{\alpha}_t' \quad .$$

Since

$$\underline{\alpha}_j'\underline{\alpha}_t=\delta_{jt} \quad \text{where} \quad \delta_{jt}=\begin{cases}0 & j\neq t\\ 1 & j=t\end{cases}$$

we obtain,

$$(W-\lambda_k I)^{+}(W-\lambda_k I)TIC(\underline{x},\underline{\alpha}_k)=\sum_{\substack{j=1\\ j\neq k}}^{p}\underline{\alpha}_j\underline{\alpha}_j'TIC(\underline{x},\underline{\alpha}_k) \quad .$$

Noting that $\Gamma'\Gamma=\Gamma\Gamma'=I$,

$$=(I-\underline{\alpha}_k\underline{\alpha}_k')TIC(\underline{x},\underline{\alpha}_k) \quad .$$

Using (3.5.7),

$$=TIC(\underline{x},\underline{\alpha}_k) \quad .$$

Hence,

$$\begin{aligned}TIC(\underline{x},\underline{\alpha}_k)&=\sum_{\substack{j=1\\ j\neq k}}^{p}\underline{\alpha}_j(\lambda_j-\lambda_k)^{-1}\underline{\alpha}_j'\left[TIC(\underline{x},\lambda_k)I-TIC(\underline{x},W)\right]\underline{\alpha}_k\\ &=-\sum_{\substack{j=1\\ j\neq k}}^{p}\underline{\alpha}_j(\lambda_j-\lambda_k)^{-1}\underline{\alpha}_j'TIC(\underline{x},W)\underline{\alpha}_k \qquad (3.5.9)\end{aligned}$$

This approach is used by Critchley (1985) and is used throughout this thesis as generalised inverses lead to the simplest expressions for the theoretical influence functions of eigenvectors. This also means our corresponding empirical curves are quicker to calculate. Specifically, the advantage of generalised inverses is that we do not need to calculate and store the p inverses corresponding to each eigenvector that the two other approaches require. However, we will illustrate and compare the three approaches further by applying them all to the eigenvectors from PCA on the covariance matrix.

3.6 Theoretical Influence Functions for Principal Component Analysis Using the Covariance Matrix

3.6.1 Theoretical Influence Function for the Eigenvalues

Substituting (2.2.6) into (3.5.3) gives

$$\begin{aligned} TIC_V(\underline{x}, \lambda_k) &= \underline{\alpha}_k' \left[-\Sigma + (\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})' \right] \underline{\alpha}_k \\ &= -\lambda_k + Z_k^2 \end{aligned} \qquad (3.6.1)$$

where Z_k^2 is the score of the added in point on the kth axis. We can see that (3.6.1) is the same as our 2×2 results. We will not discuss the form of $TIC_V(\underline{x}, \lambda_k)$ here, see § 3.8. where it will be examined with the other influence functions including those from the correlation matrix.

3.6.2 Theoretical Influence Function for the Eigenvectors

Substituting (2.2.6) into (3.5.4) gives

$$(\Sigma - \lambda_k I) TIC_V(\underline{x}, \underline{\alpha}_k) = \left[\Sigma - \lambda_k I + Z_k^2 I - (\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})' \right] \underline{\alpha}_k$$

and using (3.1.1),

$$= \left[Z_k^2 I - (\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})' \right] \underline{\alpha}_k \quad .$$

Using the first approach in § 3.5.2 we have

$$TIC_V(\underline{x}, a_{k1}) = 0.$$

$$TIC_V(\underline{x}, \underline{a}_k^*) = B^{-1} \left[Z_k^2 I^* - (\underline{x} - \underline{\mu})^* (\underline{x} - \underline{\mu})' \right] \underline{a}_k$$

where * indicates that the first row has been omitted, so

$$\underline{a}_k^* = (a_{k2}, \ldots\ldots, a_{kp})' \quad .$$

and $B = (\Sigma - \lambda_k I)^*$, which is $(\Sigma - \lambda_k I)$ with the first row and column removed. Dividing through by $1/\alpha_{k1} = (\underline{a}_k' \underline{a}_k)^{1/2}$,

$$\alpha_{k1} TIC_V(\underline{x}, \underline{a}_k^*) = B^{-1} \left[Z_k^2 \underline{\alpha}^*{}_k - (\underline{x} - \underline{\mu})^* Z_k \right]$$

Re-expressing $(\underline{x} - \underline{\mu})^*$ in terms of the principal components, $\Gamma^* \underline{Z}$, the lth row

of $\left[Z_k^2 \underline{\alpha}^*{}_k - (\underline{x} - \underline{\mu})^* Z_k\right]$ is,

$$Z_k^2 \alpha_{kl} - \sum_{j=1}^{p} \alpha_{jl} Z_j Z_k = -Z_k \sum_{\substack{j=1 \\ j \neq k}}^{p} \alpha_{jl} Z_j \quad .$$

Hence,

$$TIC_V(\underline{x}, \underline{a}_k^*) = -\frac{B^{-1}}{\alpha_{k1}} Z_k \sum_{\substack{j=1 \\ j \neq k}}^{p} \underline{\alpha}_j^* Z_j \quad . \tag{3.6.2}$$

The result for the alternative normalisation is obtained using (3.5.5) and

$$TIC_V(\underline{x}, \underline{a}_k) = \begin{pmatrix} TIC_V(\underline{x}, a_{k1}) \\ TIC_V(\underline{x}, \underline{a}_k^*) \end{pmatrix} \quad .$$

Radhakrishnan and Kshirsagar (1981) leave their expression as

$$TIC_V(\underline{x}, \underline{\alpha}_k) = (\Sigma - \lambda_k I + \underline{\alpha}_k \underline{\alpha}_k')^{-1} \left[(\underline{\alpha}_k'(\underline{x} - \underline{\mu}))^2 I - (\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})'\right] \underline{\alpha}_k \quad .$$

By similar reasoning to the above this can be re-expressed as,

$$TIC_V(\underline{x}, \underline{\alpha}_k) = -(\Sigma - \lambda_k I + \underline{\alpha}_k \underline{\alpha}_k')^{-1} Z_k \sum_{\substack{j=1 \\ j \neq k}}^{p} \underline{\alpha}_j Z_j \tag{3.6.3}$$

The generalised inverse approach gives,

$$\begin{aligned} TIC_V(\underline{x}, \underline{\alpha}_k) &= \sum_{\substack{j=1 \\ j \neq k}}^{p} \underline{\alpha}_j (\lambda_j - \lambda_k)^{-1} \underline{\alpha}_j' \left[Z_k^2 \underline{\alpha}_k - (\underline{x} - \underline{\mu}) Z_k\right] \\ &= -\sum_{\substack{j=1 \\ j \neq k}}^{p} \underline{\alpha}_j (\lambda_j - \lambda_k)^{-1} \underline{\alpha}_j' (\underline{x} - \underline{\mu}) Z_k \\ &= -Z_k \sum_{\substack{j=1 \\ j \neq k}}^{p} \underline{\alpha}_j (\lambda_j - \lambda_k)^{-1} Z_j \quad . \end{aligned} \tag{3.6.4}$$

All three expressions (3.6.2), (3.6.3) and (3.6.4) are similar but the generalised inverse approach gives the simplest form. When the sample equivalents are substituted in we obtain the same numerical results from the three expressions. There were numerical problems with the matrix inversions when using (3.6.2) and (3.6.3) when $k = 1$ and $k = p$ due to lack of positive definiteness. However, this problem was avoided if we did not invert the matrices but left them on the LHS of the expressions and solved the resulting

simultaneous equations. There were also problems due to ' ill-conditioning ' for the Radhakrishnan and Kshirsagar approach using the simultaneous method when it was applied to the correlation eigenvectors.

3.6.3 Theoretical Influence Function for the Principal Component Scores

Let Z_{kc} be the value of the point $\underline{c}$ on the kth principal component. Then

$$Z_{kc} = \underline{\alpha}_k'(\underline{c} - \underline{\mu}) \quad .$$

Using the product rule for influence discussed in Chapter 1,

$$TIC_V(\underline{x}, Z_{kc}) = TIC_V(\underline{x}, \underline{\alpha}_k)'(\underline{c} - \underline{\mu}) + \underline{\alpha}_k' TIC_V(\underline{x}, \underline{c} - \underline{\mu}) \quad . \qquad (3.6.5)$$

As $\underline{c}$ is a fixed point $TIC_V(\underline{x}, \underline{c} - \underline{\mu}) = -TIC(\underline{x}, \underline{\mu})$ and from Campbell (1978)

$$-TIC(\underline{x}, \underline{\mu}) = -(\underline{x} - \underline{\mu}) \quad .$$

Substituting (3.6.4) into (3.6.5) gives,

$$\begin{aligned} TIC_V(\underline{x}, Z_{kc}) &= -Z_k \sum_{\substack{j=1 \\ j \neq k}}^{p} \underline{\alpha}_j'(\lambda_j - \lambda_k)^{-1} Z_j(\underline{c} - \underline{\mu}) - \underline{\alpha}_k'(\underline{x} - \underline{\mu}) \\ &= -Z_k \left[1 + \sum_{\substack{j=1 \\ j \neq k}}^{p} (\lambda_j - \lambda_k)^{-1} Z_j Z_{jc} \right] \end{aligned} \qquad (3.6.6)$$

where Z_k is the score of the added in point $\underline{x}$. This will be discussed further in § 3.8.

3.7 Theoretical Influence Functions for Principal Component Analysis Using the Correlation Matrix

We shall not discuss the form of these influence curves until § 3.8, except to note that the results are much more complicated than for the corresponding covariance results. We replace $TIC(\underline{x}, W)$ with the influence function for the correlation matrix, R, in the definitions of the influence functions for the eigenvalues and eigenvectors in § 3.5. The influence function for the correlation matrix is a matrix whose diagonal elements are zero and off diagonal elements are the influence functions for the bivariate

correlations.

$$TIC(\underline{x}, R) = \begin{pmatrix} 0 & TIC(\underline{x}, \rho_{12}) & .. & .. & TIC(\underline{x}, \rho_{1p}) \\ .. & 0 & .. & .. & .. \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ TIC(\underline{x}, \rho_{1p}) & TIC(\underline{x}, \rho_{2p}) & .. & .. & 0 \end{pmatrix} \qquad (3.7.1)$$

3.7.1 Theoretical Influence Function for the Eigenvalues

Substituting (3.7.1) into (3.5.3) gives,

$$TIC_R(\underline{x}, \lambda_k) = 2 \sum_{s=1}^{p} \sum_{\substack{t=1 \\ t>s}}^{p} \alpha_{ks}\alpha_{kt} TIC(\underline{x}, \rho_{st}) \quad . \qquad (3.7.2)$$

When $p = 2$ this specialises to our 2×2 results in § 3.4 as,

$$TIC_R(\underline{x}, \lambda_k) = 2\alpha_{k1}\alpha_{k2} TIC(\underline{x}, \rho_{12}) \quad .$$

Thus, when $k = 1$ and $\rho_{12} > 0$

$$TIC_R(\underline{x}, \lambda_k) = 2\frac{1}{\sqrt{2}}\frac{1}{\sqrt{2}} TIC(\underline{x}, \rho_{12}) = TIC(\underline{x}, \rho_{12}) \quad .$$

We shall now re-express $TIC_R(\underline{x}, \lambda_k)$ in terms of the principal components, which gave quite simple expressions in the covariance case. Substituting (2.3.5) into (3.7.2) we have,

$$\begin{aligned} TIC_R(\underline{x}, \lambda_k) &= -\sum_{s=1}^{p} \sum_{\substack{t=1 \\ t\neq s}}^{p} \alpha_{ks}\alpha_{kt}(y_s^2 + y_t^2)\frac{\rho_{st}}{2} + \sum_{s=1}^{p} \sum_{\substack{t=1 \\ t\neq s}}^{p} \alpha_{ks}\alpha_{kt} y_s y_t \\ &= -\sum_{s=1}^{p} \sum_{\substack{t=1 \\ t\neq s}}^{p} \alpha_{ks}\alpha_{kt} y_s^2 \rho_{st} + \sum_{s=1}^{p} \sum_{\substack{t=1 \\ t\neq s}}^{p} \alpha_{ks}\alpha_{kt} y_s y_t \end{aligned} \qquad (3.7.3)$$

The first term in (3.7.3) is,

$$-\sum_{s=1}^{p} \alpha_{ks} y_s^2 \sum_{\substack{t=1 \\ t\neq s}}^{p} \alpha_{kt}\rho_{st} = (1-\lambda_k)\left[\sum_{s=1}^{p} \alpha_{ks}^2 y_s^2\right]$$

as $(R - \lambda_k I)\underline{\alpha}_k = 0$, so that $-\sum_{\substack{t=1 \\ t\neq s}}^{p} \alpha_{kt}\rho_{st} = (1-\lambda_k)\alpha_{ks}$. The first term in (3.7.3) is thus,

$$(1-\lambda_k)\left[\sum_{s=1}^{p}\alpha_{ks}^2 y_s^2\right]=(1-\lambda_k)Z_k^2-(1-\lambda_k)\sum_{s=1}^{p}\sum_{\substack{t=1\\t\neq s}}^{p}\alpha_{ks}\alpha_{kt}y_s y_t \quad ,$$

since, $[\alpha_{k1}y_1+.....+\alpha_{kp}y_p]^2=Z_k^2$. So, (3.7.3) becomes,

$$TIC_R(\underline{x}\,,\lambda_k)=(1-\lambda_k)Z_k^2+\lambda_k\sum_{s=1}^{p}\sum_{\substack{t=1\\t\neq s}}^{p}\alpha_{ks}\alpha_{kt}y_s y_t \quad .$$

We now need to write y_s and y_t in terms of the principal components. This gives,

$$TIC_R(\underline{x}\,,\lambda_k)=(1-\lambda_k)Z_k^2+\lambda_k\sum_{s=1}^{p}\sum_{\substack{t=1\\t\neq s}}^{p}\alpha_{ks}\alpha_{kt}\Bigg[\sum_{u=1}^{p}\alpha_{us}\alpha_{ut}Z_u^2+\sum_{u=1}^{p}\sum_{\substack{v=1\\v\neq s}}^{p}(\alpha_{us}\alpha_{vt}+\alpha_{ut}\alpha_{vs})Z_u Z_v\Bigg]$$

$$TIC_R(\underline{x}\,,\lambda_k)=\left[1-\lambda_k+2\lambda_k\sum_{s=1}^{p}\sum_{\substack{t=1\\t>s}}^{p}\alpha_{ks}^2\alpha_{kt}^2\right]Z_k^2$$

$$+2\lambda_k\sum_{\substack{u=1\\u\neq k}}^{p}\left[\sum_{s=1}^{p}\sum_{\substack{t=1\\t>s}}^{p}\alpha_{ks}\alpha_{kt}\alpha_{us}\alpha_{ut}\right]Z_u^2$$

$$+2\lambda_k\sum_{u=1}^{p}\sum_{\substack{v=1\\v\neq s}}^{p}\left[\sum_{s=1}^{p}\sum_{\substack{t=1\\t\neq s}}^{p}\alpha_{ks}\alpha_{kt}(\alpha_{us}\alpha_{vt}+\alpha_{ut}\alpha_{vs})\right]Z_u Z_v$$

From the normalisations of our eigenvectors we have the relationships,

$$\sum_{s=1}^{p}\sum_{\substack{t=1\\t\neq s}}^{p}\alpha_{ks}^2\alpha_{kt}^2=\sum_{s=1}^{p}\alpha_{ks}^2(1-\alpha_{ks}^2)$$

$$=1-\sum_{s=1}^{p}\alpha^4{}_{ks}$$

$$\sum_{s=1}^{p}\sum_{\substack{t=1\\t\neq s}}^{p}\alpha_{ks}\alpha_{kt}\alpha_{us}\alpha_{ut}=\sum_{s=1}^{p}\alpha_{ks}\alpha_{us}(0-\alpha_{ks}\alpha_{us})$$

$$=-\sum_{s=1}^{p}\alpha_{ks}^2\alpha_{us}^2 \quad ,$$

substituting these and similar relationships into $TIC_R(\underline{x}\,,\lambda_k)$ gives after some algebra,

$$TIC_R(\underline{x}\,,\lambda_k)=\left[1-\lambda_k\sum_{s=1}^{p}\alpha^4{}_{ks}\right]Z_k^2+\sum_{\substack{u=1\\u\neq k}}^{p}-\lambda_k\sum_{s=1}^{p}\alpha_{ks}^2\alpha_{us}^2Z_u^2$$

$$+\sum_{u=1}^{p}\sum_{\substack{v=1\\v\neq s}}^{p}-2\lambda_k\sum_{s=1}^{p}\alpha_{ks}^2\alpha_{us}\alpha_{vs}Z_u Z_v \qquad (3.7.4)$$

3.7.2 Theoretical Influence Function for the Eigenvectors

We will only use the generalised inverse approach for the theoretical influence function of the correlation eigenvectors. Substituting (3.7.1) into (3.5.9) gives,

$$TIC_R(\underline{x}, \underline{\alpha}_k) = -\sum_{\substack{j=1 \\ j \neq k}}^{p} \underline{\alpha}_j (\lambda_j - \lambda_k)^{-1} \underline{\alpha}_j' TIC(\underline{x}, R) \underline{\alpha}_k$$

$$= -\sum_{\substack{j=1 \\ j \neq k}}^{p} \underline{\alpha}_j (\lambda_j - \lambda_k)^{-1} \sum_{s=1}^{p} \sum_{\substack{t=1 \\ t \neq s}}^{p} \alpha_{js} \alpha_{kt} TIC(\underline{x}, \rho_{st}) \qquad (3.7.5)$$

A similar method to that given for the eigenvalues gives $TIC_R(\underline{x}, \underline{\alpha}_k)$ in terms of the principal components as

$$TIC_R(\underline{x}, \underline{\alpha}_k) = -\sum_{\substack{j=1 \\ j \neq k}}^{p} \underline{\alpha}_j (\lambda_j - \lambda_k)^{-1} \left[-\frac{1}{2}(\lambda_j + \lambda_k) \sum_{u=1}^{p} \sum_{t=1}^{p} \alpha_{ju} \alpha_{kt} \alpha^2_{ut} Z_u^2 \right.$$

$$\left. + Z_j Z_k - (\lambda_j + \lambda_k) \sum_{u=1}^{p} \sum_{\substack{v=1 \\ v>u}}^{p} \sum_{t=1}^{p} \alpha_{kt} \alpha_{ju} \alpha_{ut} \alpha_{vt} Z_u Z_v \right] \qquad (3.7.6)$$

The theoretical influence functions for the eigenvalues and eigenvectors from the correlation matrix do not have a simple form when expressed in terms of the principal components. The accuracy of these expressions were checked by comparing numerically with the curves not expressed in terms of the principal components, i.e. (3.7.2) and (3.7.5). In practice one would probably use the expressions given by (3.7.2) and (3.7.5). However, the alternative forms do provide some insight into the form of the influence curves. For example, the second term in (3.7.6) is the same as the influence function for the covariance eigenvectors. The interpretation of these curves will be discussed further in § 3.8.

3.7.3 Theoretical Influence Function for the Principal Component Scores

$$TIC_R(\underline{x}, Z_{kc}) = TIC_R(\underline{x}, \underline{\alpha}_k)' \underline{y}_c + \underline{\alpha}_k' TIC(\underline{x}, \underline{y}_c)$$

$Z_{kc} = \underline{\alpha}_k' \underline{y}_c$ where $\underline{y}_c$ is the point $\underline{c}$ whose elements have been standardised The

perturbed tth element of $\underline{y}_c$ is,

$$\frac{c_t - \tilde{\mu}_t}{\sqrt{\tilde{\sigma}_{tt}}} = [c_t - \mu_t - \epsilon(x_t - \mu_t)]\frac{1}{\sqrt{\sigma_{tt}}}\left[1 + \frac{\epsilon}{\sigma_{tt}}\left(-\sigma_{tt} + (x_t - \mu_t)^2\right)\right]^{-1/2}$$

To $o(\epsilon)$

$$= \left[y_{ct} - \epsilon y_t\right]\left[1 - \frac{\epsilon}{2}(-1 + y^2{}_t)\right]$$

where y_{ct} is the tth standardised variable of the point c, and y_t is the tth standardised variable of our added in point. Hence,

$$TIC_R(\underline{x}, \underline{y}_c) = \frac{1}{2}\underline{y}_c - \frac{1}{2}\begin{pmatrix} y_{c1}y_1^2 \\ \vdots \\ y_{cp}y_p^2 \end{pmatrix} - \underline{y}$$

where $\underline{y}$ is the standardised added in point. This gives,

$$TIC_R(\underline{x}, Z_{kc}) = -\sum_{\substack{j=1 \\ j\neq k}}^{p}(\lambda_j - \lambda_k)^{-1}Z_{jc}\sum_{s=1}^{p}\sum_{\substack{t=1 \\ t\neq s}}^{p}\alpha_{js}\alpha_{ks}TIC(\underline{x}, \rho_{st})$$

$$+ \frac{1}{2}Z_{kc} - \frac{1}{2}\sum_{t=1}^{p}\alpha_{kt}y_{ct}y_t^2 - Z_k \qquad (3.7.7)$$

3.8. Plots and Comparisons of the Influence Functions

3.8.1. Summary of the Influence Functions

$$TIC_V(\underline{x}, \lambda_k) = -\lambda_k + Z_k^2 \tag{3.8.1}$$

$$TIC_V(\underline{x}, \underline{\alpha}_k) = -Z_k \sum_{\substack{j=1 \\ j \neq k}}^{p} \underline{\alpha}_j (\lambda_j - \lambda_k)^{-1} Z_j \tag{3.8.2}$$

$$TIC_V(\underline{x}, Z_{kc}) = -Z_k \left[1 + \sum_{\substack{j=1 \\ j \neq k}}^{p} (\lambda_j - \lambda_k)^{-1} Z_j Z_r \right] \tag{3.8.3}$$

$$TIC_R(\underline{x}, \lambda_k) = 2 \sum_{s=1}^{p} \sum_{\substack{t=1 \\ t>s}}^{p} \alpha_{ks} \alpha_{kt} TIC(\underline{x}, \rho_{st}) \tag{3.8.4}$$

$$TIC_R(\underline{x}, \lambda_k) = \left[1 - \lambda_k \sum_{s=1}^{p} \alpha^4_{ks} \right] Z_k^2 + \sum_{\substack{u=1 \\ u \neq k}}^{p} - \lambda_k \sum_{s=1}^{p} \alpha^2_{ks} \alpha^2_{us} Z_u^2$$

$$+ \sum_{u=1}^{p} \sum_{\substack{v=1 \\ v \neq s}}^{p} -2\lambda_k \sum_{s=1}^{p} \alpha^2_{ks} \alpha_{us} \alpha_{vs} Z_u Z_v \tag{3.8.5}$$

$$TIC_R(\underline{x}, \underline{\alpha}_k) = -\sum_{\substack{j=1 \\ j \neq k}}^{p} \underline{\alpha}_j (\lambda_j - \lambda_k)^{-1} \sum_{s=1}^{p} \sum_{\substack{t=1 \\ t \neq s}}^{p} \alpha_{js} \alpha_{kt} TIC(\underline{x}, \rho_{st}) \tag{3.8.6}$$

$$TIC_R(\underline{x}, \underline{\alpha}_k) = -\sum_{\substack{j=1 \\ j \neq k}}^{p} \underline{\alpha}_j (\lambda_j - \lambda_k)^{-1} \left[-\frac{1}{2}(\lambda_j + \lambda_k) \sum_{u=1}^{p} \sum_{t=1}^{p} \alpha_{js} \alpha_{kt} \alpha^2_{ut} Z_u^2 \right.$$

$$\left. + Z_j Z_k - (\lambda_j + \lambda_k) \sum_{u=1}^{p} \sum_{\substack{v=1 \\ v>u}}^{p} \sum_{t=1}^{p} \alpha_{kt} \alpha_{js} \alpha_{ut} \alpha_{vt} Z_u Z_v \right] \tag{3.8.7}$$

$$TIC_R(\underline{x}, Z_{kc}) = -\sum_{\substack{j=1 \\ j \neq k}}^{p} (\lambda_j - \lambda_k)^{-1} Z_{jc} \sum_{s=1}^{p} \sum_{\substack{t=1 \\ t \neq s}}^{p} \alpha_{js} \alpha_{kt} TIC(\underline{x}, \rho_{st})$$

$$+ \frac{1}{2} Z_{kc} - \frac{1}{2} \sum_{t=1}^{p} \alpha_{kt} y_{ct} y^2_t - Z_k \quad . \tag{3.8.8}$$

3.8.2. Examination and Comparisons of the Theoretical Influence Functions

As we go down the functions in § 3.8.1 we see that they gradually become more and more complicated. Algebra extracted from the correlation matrix is always more complicated and often results can only be represented for the covariance matrix. However, in practice we often prefer to use the correlation matrix so any results we can derive for it are worthwhile.

Expression (3.8.1) is the simplest and reflects that the only point that will be highly influential on a covariance eigenvalue is a point that is extreme on its corresponding principal component. If $|z_k| \leq \sqrt{\lambda_k}$ then an eigenvalue can decrease, and if $|z_k| \leq \sqrt{\lambda_k}$ for all k then all our eigenvalues could decrease when we add the extra point. This could happen for example if our additional point $\underline{x} = \underline{\mu}$. (Note, we subtract the original from the perturbed here so a negative influence means a decrease in the eigenvalue). The most influential points are those with a large component score and they lead to the eigenvalue increasing in value when they are added (or a decrease when they are omitted).

It is impossible for all the eigenvalues to decrease when another point is added for the correlation matrix, just as it would be impossible for them all to increase, since

$$\sum_{k=1}^{p} \tilde{\lambda}_k = \sum_{k=1}^{p} \lambda_k = p \quad .$$

There is no such restriction on the eigenvalues from the covariance matrix. We thus find that when we write $TIC_R(\underline{x}, \lambda_k)$ in terms of the principal components it involves all the square and cross product terms in the principal components. Unlike the covariance eigenvalues the most influential observations on a given eigenvalue can be a mixture of positive and negative influences. We seem to get a large positive influence i.e. an increase in the eigenvalue when a point is included, if it is extreme in the direction of the component. A large negative influence may occur when the change is a compensatory change for a large increase in one of the other eigenvalues. The complicated nature of the influence functions for the correlation matrix eigenvalues makes it more difficult to say what sort of points will generally be influential, particularly on the larger eigenvalues. If λ_k is very small in

(3.8.5) then,

$$TIC_R(\underline{x}, \lambda_k) \approx Z_k^2 \quad .$$

Thus, the expression for small λ_k from the correlation matrix is more like that for the eigenvalues from the covariance matrix since both depend on the square of the principal component score for that direction. We tend to get very small λ_ks when we have highly correlated data. We will examine the function further when we look at contour plots for small p in the next section.

One question of interest is whether, from expression (3.8.4) and a similar one for the covariance eigenvalues which is

$$TIC_V(\underline{x}, \lambda_k) = \underline{\alpha}_k' \, TIC(\underline{x}, \Sigma)\underline{\alpha}_k = \sum_{s=1}^{p} \sum_{t=1}^{p} \alpha_{ks}\alpha_{kt} TIC(\underline{x}, \sigma_{st}) \quad ,$$

we could decide from looking at influence on the bivariate correlations (or covariances and variances) what will be influential in our PCA from the correlation (covariance) matrix. If an observation has a large affect on some of the correlations then it is likely to be influential on some part of our analysis. However, affects can be cancelled out in the summation term of (3.8.4), and it would not be obvious what component the observation may come out as influential on. This is important as we usually only retain a few of the principal components so the observation although influential on the bivariate correlations may not come out in an 'important' direction. Investigating influence on the $p(p-1)/2$ possible correlations would also be much more complicated than considering the eigenvalues and eigenvectors from the desired principal component directions. In § 4.7 we shall compare the observations that are influential on the bivariate correlations, corresponding to variables with large coefficients in the latter eigenvectors, with those influential on these eigenvectors and the corresponding eigenvalues.

Expressions (3.8.1) and (3.8.2) reveal how different observations can be influential on the covariance eigenvalues and eigenvectors. An observation lying far out along the direction of the kth principal component will be very influential on λ_k but not on $\underline{\alpha}_k$. The two curves and differences between them will be discussed further in the next section. Expressions for the eigenvalues and eigenvectors from the correlation matrix provide little information on what types of observations may be influential on both or just one of them. Unfortunately, the plots in the next section do not help us as those for the correlation eigenvectors have no obvious pattern.

If we add a point along the direction of an existing principal component we find that the correlation eigenvectors may change although those from the covariance matrix do not. The correlation eigenvectors will change due to the square terms in (3.8.7), whereas (3.8.2) only has cross product terms. The second order terms for the eigenvectors from the covariance matrix, see § 4.4, also show that the covariance eigenvectors will not change when only one principal component score is non-zero. We did note in § 3.7.2 that one of the terms in $TIC_R(\underline{x}, \underline{\alpha}_k)$ was the same as $TIC_V(\underline{x}, \underline{\alpha}_k)$, see (3.8.2) and (3.8.7). The other terms in (3.8.7) are very complicated involving all the principal component scores. Both eigenvector expressions (3.8.2) and (3.8.7) involve terms in $(\lambda_j - \lambda_k)^{-1}$ which comes from the generalised inverse, and it means that when we have close eigenvalues λ_k and λ_u the changes in the corresponding eigenvectors can be very large. However, the changes in both eigenvectors $\underline{\alpha}_k$ and $\underline{\alpha}_u$ will be nearly equal as $(\lambda_u - \lambda_k)^{-1}$ will dominate both influence functions. Since the eigenvalues are very close the large changes in the eigenvectors may just represent them rotating within a relatively unchanged subspace. This problem may arise more for eigenvalues from the correlation matrix since the eigenvalues sum to p so there are often smaller

and closer eigenvalues. However, we will see in § 4.3 that it occurs in the covariance and correlation matrices alike.

We will not discuss the influence function for the scores here. An example of influence on the scores is given in § 4.8

3.8.3. Contour Plots of the Theoretical Influence Functions for small p

In this section we will examine contour plots of some of the expressions in § 3.8.1 for $p = 2$ and $p = 3$. The heights of the contour plots are the values of the theoretical influence functions for the values of the principal component scores, or variables, given along the axes. We will see in Chapter 4 that when we substitute the sample equivalents into our theoretical influence curves we obtain good estimates of the sample changes. Thus, the type of point we observe as being influential here will have some relevance to what observations are influential in samples.

Fig 3.8.1 is a plot of (3.8.1), for the largest eigenvalue from the covariance matrix,

$$\Sigma = \begin{pmatrix} 6 & -2 \\ -2 & 3 \end{pmatrix} \quad . \tag{3.8.9}$$

Expression (3.8.1) is the simplest of all the influence functions and we do not really need a plot to understand its behaviour. We include one here for completeness and to make comparisons. The contours in Fig. 3.8.1 are vertical lines cutting the first principal component axis at right angles. This shows that the value of the added point on the second principal component has no importance in determining its affect on λ_1. The contour plot for λ_2 is similar, with the contours cutting the second principal component at right angles. The contour plots for the covariance eigenvalues would have the same form for all covariance matrices.

Figure 3.8.1 Contour plot of $TIC_V(x, \lambda_1)$ for the 2×2 covariance matrix of (3.8.9) (axes $\times 10^{-1}$)

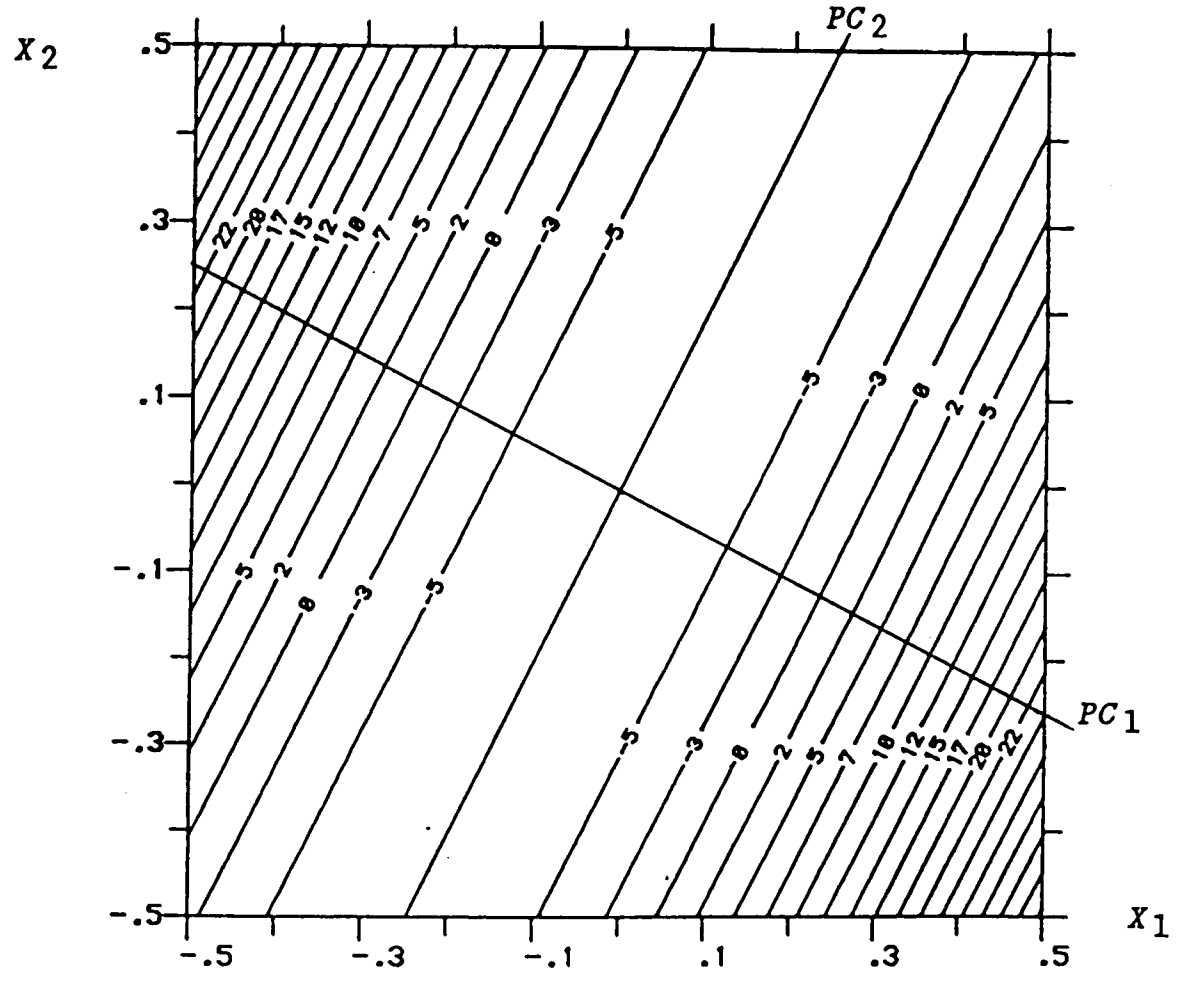


Figure 3.8.2 Plot of $I_R(x, \lambda_1)$ for a 2×2 correlation matrix with $\bar{\rho} = 0.2$ (contours $\times 10$)

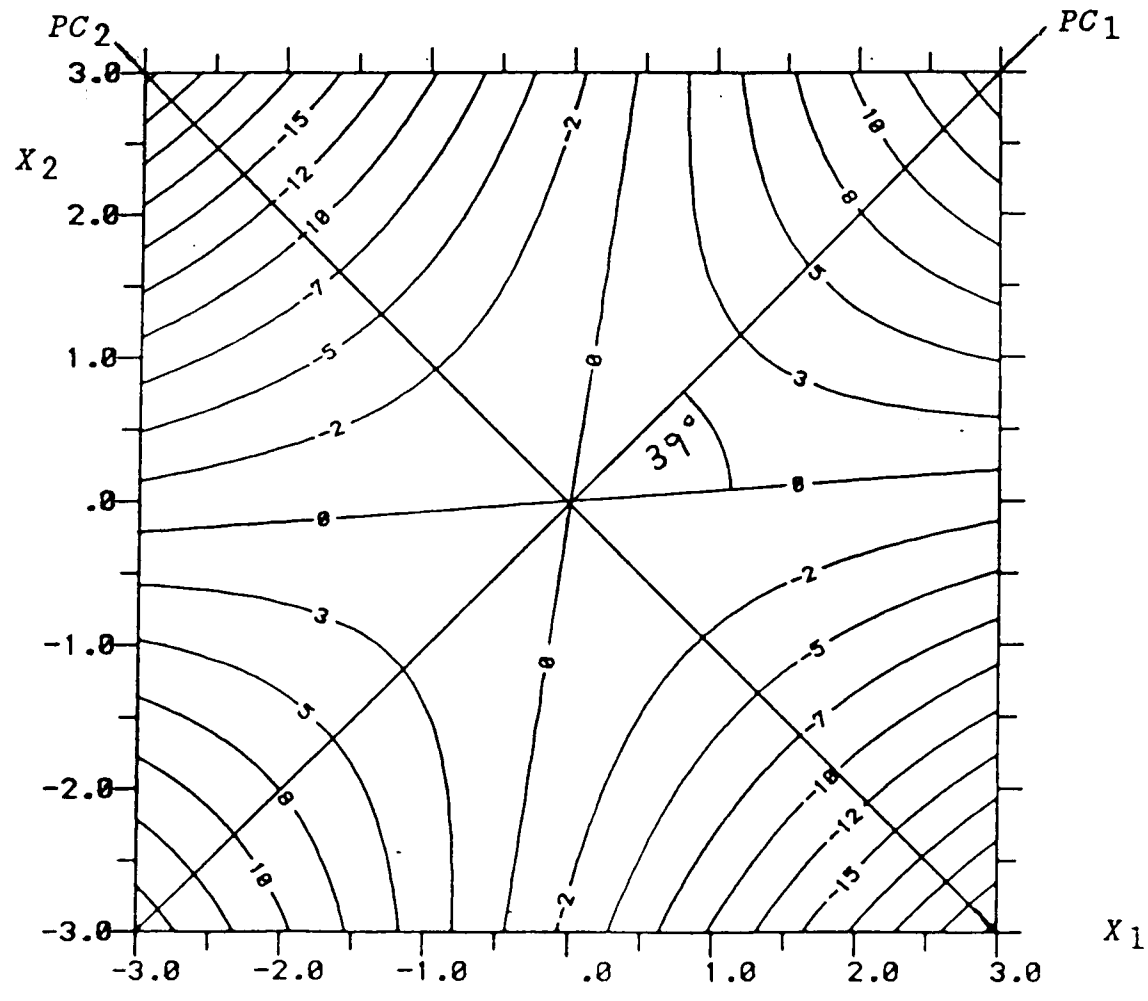


Fig. 3.8.2 is a plot of the theoretical influence function for λ_1 from the 2×2 correlation matrix with $\rho_{12} = 0.2$. We observed in § 2.3.2 that this is the same as the influence function for ρ_{12}. The plot is different to the plots for the covariance eigenvalue and reflects the fact that we cannot have independent changes in the λs due to the fixed sum of the eigenvalues. This means the contour plot for λ_2 is the same as Fig. 3.8.2 but the contours have the opposite sign. The asymptotes, which correspond to no change in the eigenvalues, make an angle θ with the first principal component axis where,

$$\theta = \tan^{-1} \pm \left(\frac{2 - \lambda_1}{\lambda_1} \right)^{1/2} \qquad (3.8.10)$$

and Table 3.8.1 gives some values of θ for some choices of ρ_{12}. We will prove result (3.8.10) for general p later in this section. As ρ_{12} increases the angle θ decreases so that the value on the second principal component becomes more important in determining the influence of a point on λ_1. This could reflect the fact that as λ_1 becomes more distinct, as ρ_{12} increases, a point needs to be further out along the first axis to be 'unusual'.

Table 3.8.1

θ (degrees) for Given Values of ρ

ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
θ	45	42.1	39.2	36.3	33.2	30.0	26.7	22.8	18.4	12.9

Since the influence functions for the correlation eigenvalues are much more complicated than those for the covariance eigenvalues, and so provide less information on what points are influential, we will look in detail at plots for $p = 3$. Three types of correlation matrices will be considered and we shall plot the influence function for each eigenvalue in the first two dimensions for $z_3 = 0$. The correlation matrices are:

(i) A matrix with high correlations

$$\begin{pmatrix} 1 & 0.9 & 0.8 \\ & 1 & 0.85 \\ & & 1 \end{pmatrix}$$

$\lambda_1 = 2.7$, $\lambda_2 = 0.21$ and $\lambda_3 = 0.09$.

(ii) A matrix with low correlations

$$\begin{pmatrix} 1 & 0.3 & 0.2 \\ & 1 & 0.25 \\ & & 1 \end{pmatrix}$$

$\lambda_1 = 1.5$, $\lambda_2 = 0.81$ and $\lambda_3 = 0.69$.

(iii) A matrix whose correlations are wider apart

$$\begin{pmatrix} 1 & 0.2 & 0.4 \\ & 1 & 0.8 \\ & & 1 \end{pmatrix}$$

$\lambda_1 = 2.0$, $\lambda_2 = 0.84$ and $\lambda_3 = 0.17$.

First consider Figs 3.8.3, 3.8.4 and 3.8.5, which are the contour plots of $TIC_R(\underline{x}, \lambda_1)$ from the above three matrices respectively in the first two dimensions with z_3 fixed at zero. As for the 2×2 results the angles that the asymptotes make with the first axis decrease as the correlations increase. Thus, the value of an observation on the first principal component becomes less important in determining its influence on λ_1. The plots show that points with a large principal component score on the second axis decrease λ_1 when they are included. (Conversely they would increase λ_1 if they were deleted). This must occur since if the variance along one direction increases some of the other variances (eigenvalues) must decrease to maintain the constant sum. The angles that the two asymptotes make with the first axis differ most for the third matrix. We can show when all correlations are equal that the angles that the two asymptotes make with the axes for the plot of $TIC_R(\underline{x}, \lambda_1)$ in the first two dimensions, are the same. The angle is given by,

Figure 3.8.3 Plot of $TIC_R(x\ ,\ \lambda_1)$ for
the matrix with high correlations $(\overline{Z}_3 = 0)$(contours × 10)

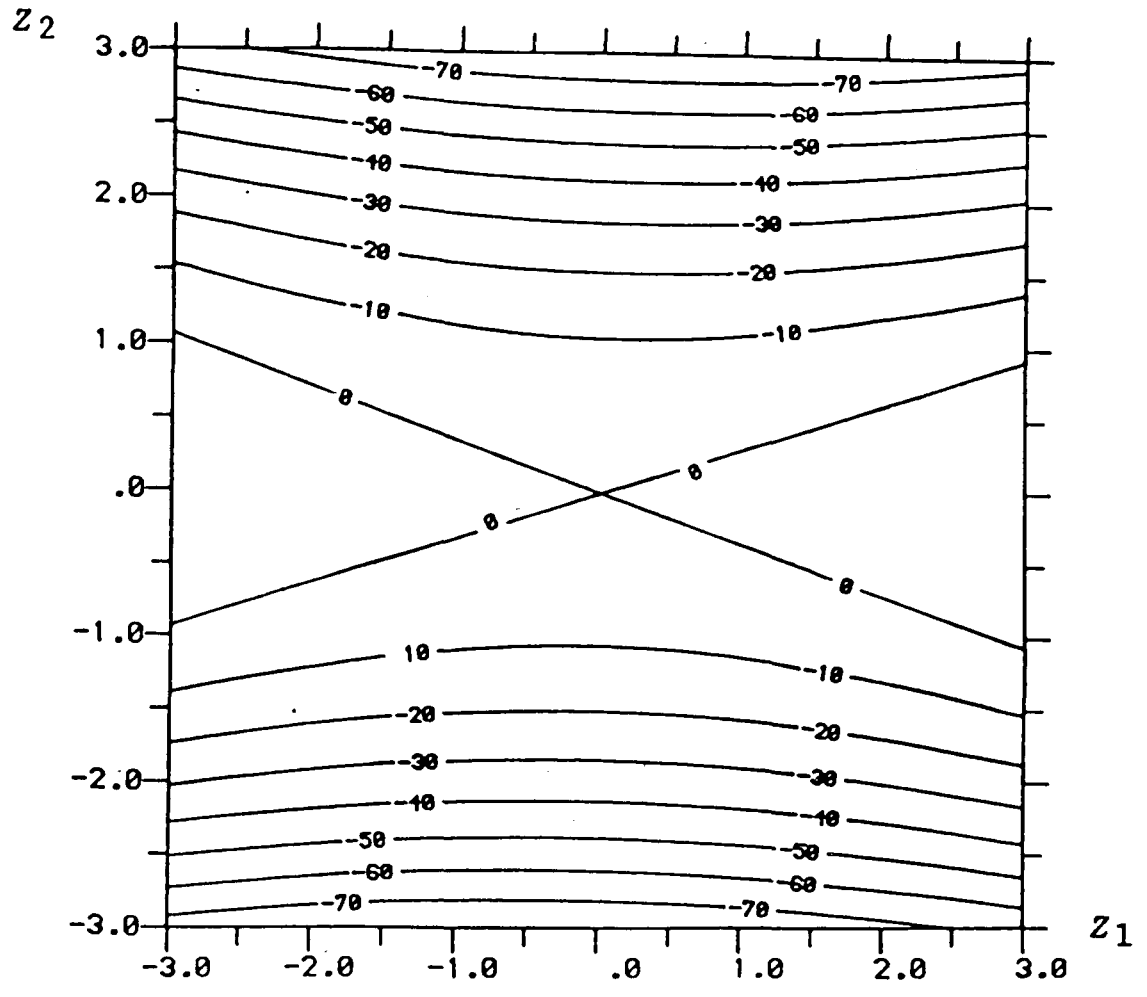


Figure 3.8.4 Plot of $TIC_R(x\ ,\ \lambda_1)$ for
the matrix with low correlations $(\overline{Z}_3 = 0)$ (contours × 10)

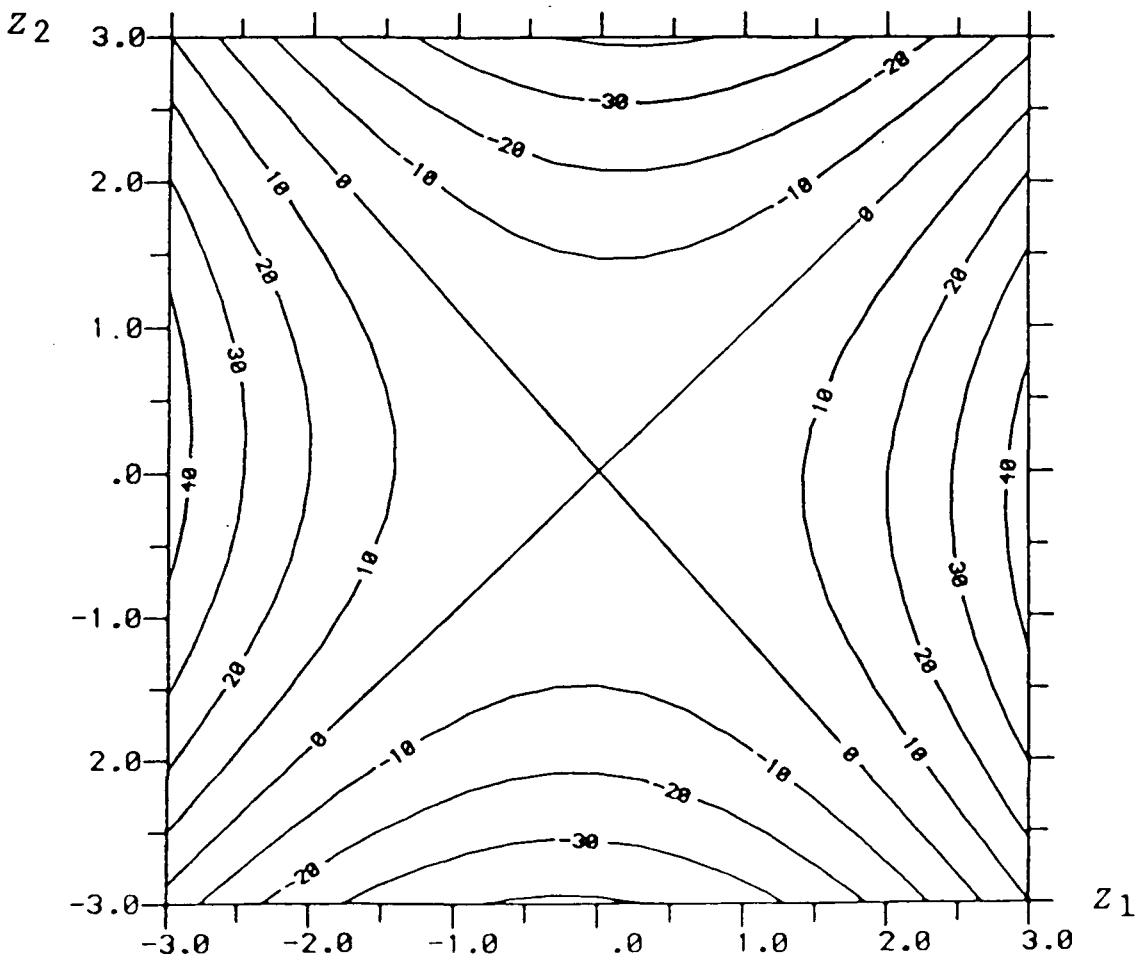
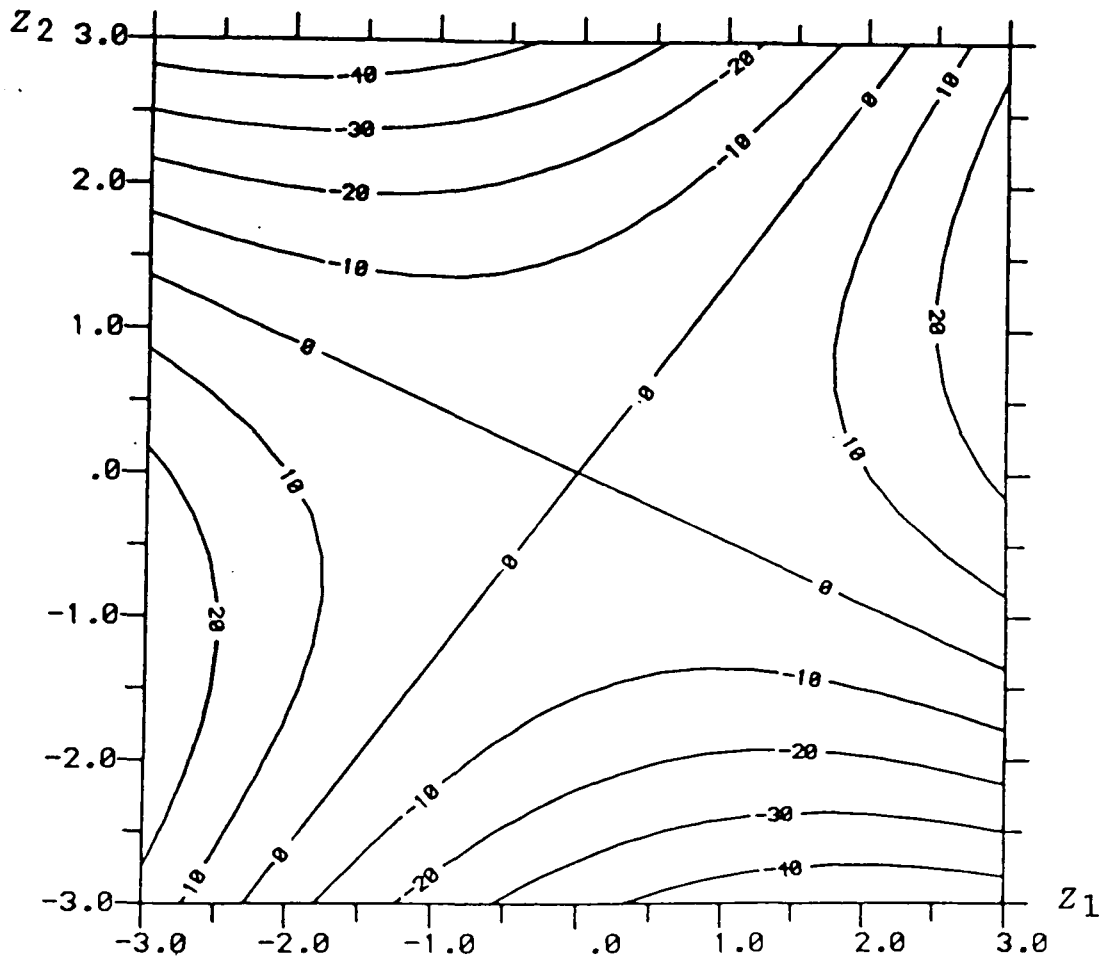


Figure 3.8.5 Plot of $TIC_R(\underline{x}, \lambda_1)$ for the matrix whose correlations are wider apart ($Z_3 = 0$) (contours × 10)



$$\theta = \tan^{-1} \pm \left[\frac{p - \lambda_1}{\lambda_1} \right]^{1/2} \quad . \tag{3.8.11}$$

Proof

When all off diagonal elements are equal to ρ we have

$$\lambda_1 = 1 + (p-1)\rho \quad , \quad \lambda_j = 1 - \rho \quad j \neq 1$$

and, $\underline{\alpha}_1 = (1/\sqrt{p} \ \ldots\ldots 1/\sqrt{p})$. So expression (3.8.5) becomes,

$$\begin{aligned} TIC_R(\underline{x}, \lambda_1) = & \left[1 - \lambda_1 \sum_{j=1}^{p} 1/p^2 \right] Z_1^2 \\ & + \sum_{k=2}^{p} \left[-\lambda_1 \sum_{j=1}^{p} 1/p\, \alpha^2{}_{kj} \right] Z_k^2 \\ & - \sum_{k=1}^{p} \left[\sum_{\substack{s=1 \\ s \neq k}}^{p} 2\lambda_1 \sum_{j=1}^{p} 1/p\, \alpha_{kj} \alpha_{sj} \right] Z_k Z_s \\ = & \left[1 - \frac{\lambda_1}{p} \right] Z_1^2 - \frac{\lambda_1}{p} \sum_{k=2}^{p} Z_k^2 \quad . \end{aligned}$$

Setting all but one of the Z_k to zero and fixing $TIC_R(\underline{x}, \lambda_1) = 0$ (which represents a point on the asymptotes) we have,

$$\begin{aligned} 0 &= \left[1 - \frac{\lambda_1}{p} \right] Z_1^2 - \frac{\lambda_1}{p} Z^2{}_j \\ \Rightarrow & \pm \left[\frac{\lambda_1}{p - \lambda_1} \right]^{1/2} Z_j = Z_1 \quad . \end{aligned}$$

Letting $Z_j = 1$ say, then

$$\tan\theta = \pm \left[\frac{p - \lambda_1}{\lambda_1} \right]^{1/2}$$

which leads to (3.8.11), and (3.8.10) is a special case of this. The correctness of the angles given by (3.8.11) has been checked by the examination of many contour plots. Plots for $TIC_R(\underline{x}, \lambda_1)$ when Z_2 rather than Z_3 is zero are not presented here but they are very similar to these plots, differing most for the third matrix perhaps due to the two smallest eigenvalues being wider apart than for the other two matrices.

Figs 3.8.6, 3.8.7 and 3.8.8 are plots of $TIC_R(\underline{x}, \lambda_2)$, when $Z_3 = 0$, for the above three correlation matrices respectively. We find the value a point has on the second principal axis is very important in determining its influence on λ_2, unlike the case for λ_1 were the value on its own axis was not very important. We will see later that the influence contours for $TIC_R(\underline{x}, \lambda_3)$ in the first two dimensions are not very large so that λ_1 and λ_2 in these dimensions are mostly compensating each others changes to maintain the constant sum of 3. The contours, particularly in Fig. 3.8.6 are almost straight and perpendicular to the second axis. This is even more so in the plots (not presented here) of $TIV_R(\underline{x}, \lambda_2)$ when Z_1 rather than Z_3 is set to zero. This pattern is also true for $TIC_R(\underline{x}, \lambda_3)$ when Z_1 or Z_2 is set to zero but these plots are not presented here. As noted in the previous section expression (3.8.5) tends to Z_k^2 as the eigenvalue tends to zero.

Finally, Figs. 3.8.9, 3.8.10 and 3.8.11 represent $TIC_R(\underline{x}, \lambda_3)$ for the three correlation matrices when $Z_3 = 0$. The set for $TIC_R(\underline{x}, \lambda_2)$ when $Z_2 = 0$ are almost identical to these, all of which are ellipses. Since the plots are only unique up to a change in sign of the initial eigenvectors the different direction of the ellipse in Fig. 3.8.11 is not important since this is just a reflection about the axis. The value of the first principal component plays a more important role than the second component on $TIC_R(\underline{x}, \lambda_3)$, even though a fixed value of three, say, would be more extreme on the second than first axis. A possible explanation for this can be seen from the corresponding plots of $TIC_R(\underline{x}, \lambda_1)$ and $TIC_R(\underline{x}, \lambda_2)$ when $Z_3 = 0$. For all the three types of correlation matrix we can see that the value of Z_1 is more important in determining $TIC_R(\underline{x}, \lambda_1)$ than $TIC_R(\underline{x}, \lambda_2)$ (although it is the value of Z_2 that plays the greatest role for both). Since we need to keep the sum $\sum_{k=1}^{3} \lambda_k = 3$ the last eigenvalue has changed accordingly. The only plots not mentioned are those

Figure 3.8.6 Plot of $TIC_R(x, \lambda_2)$ for the matrix with high correlations $(\overline{Z}_3 = 0)$ (contours × 10)

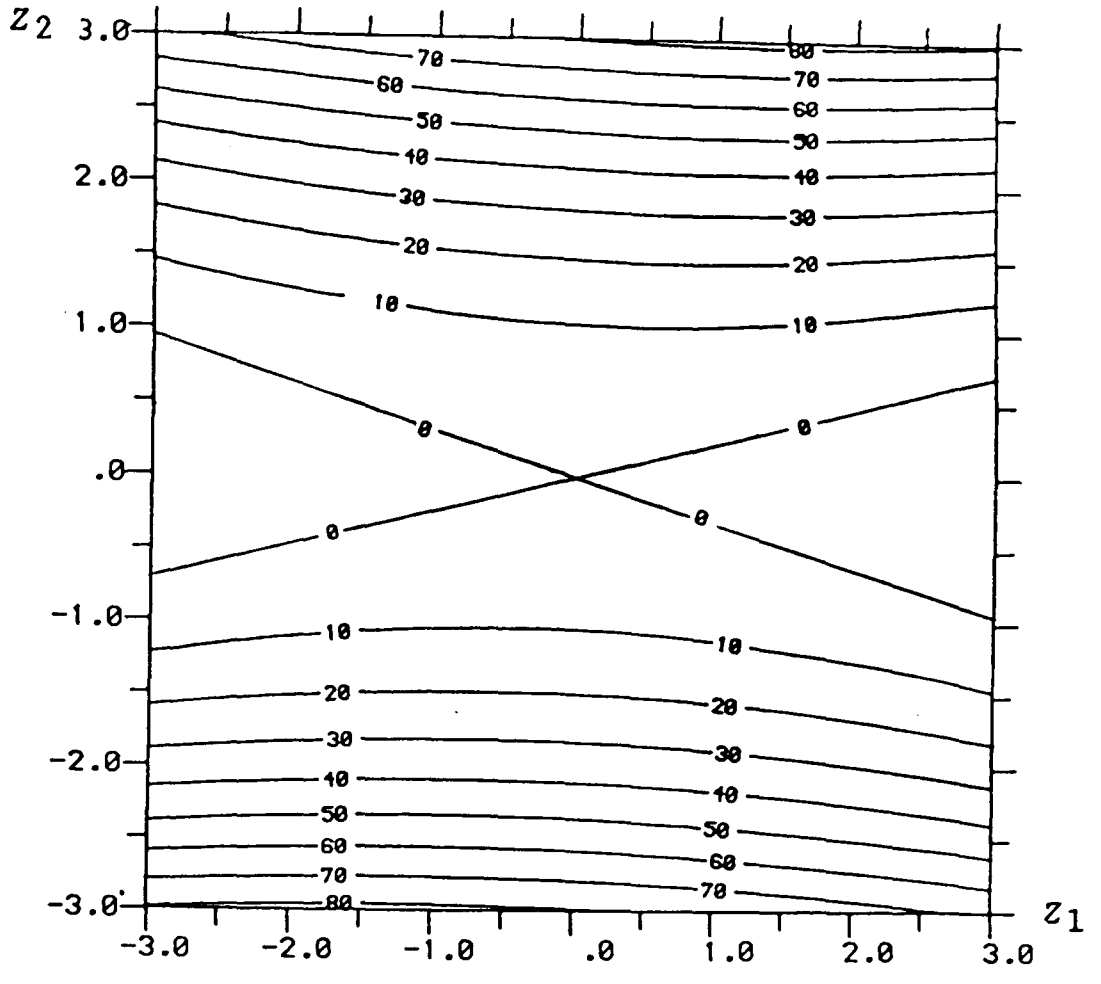


Figure 3.8.7 Plot of $TIC_R(x, \lambda_2)$ for the matrix with low correlations $(\overline{Z}_3 = 0)$ (contours × 10)

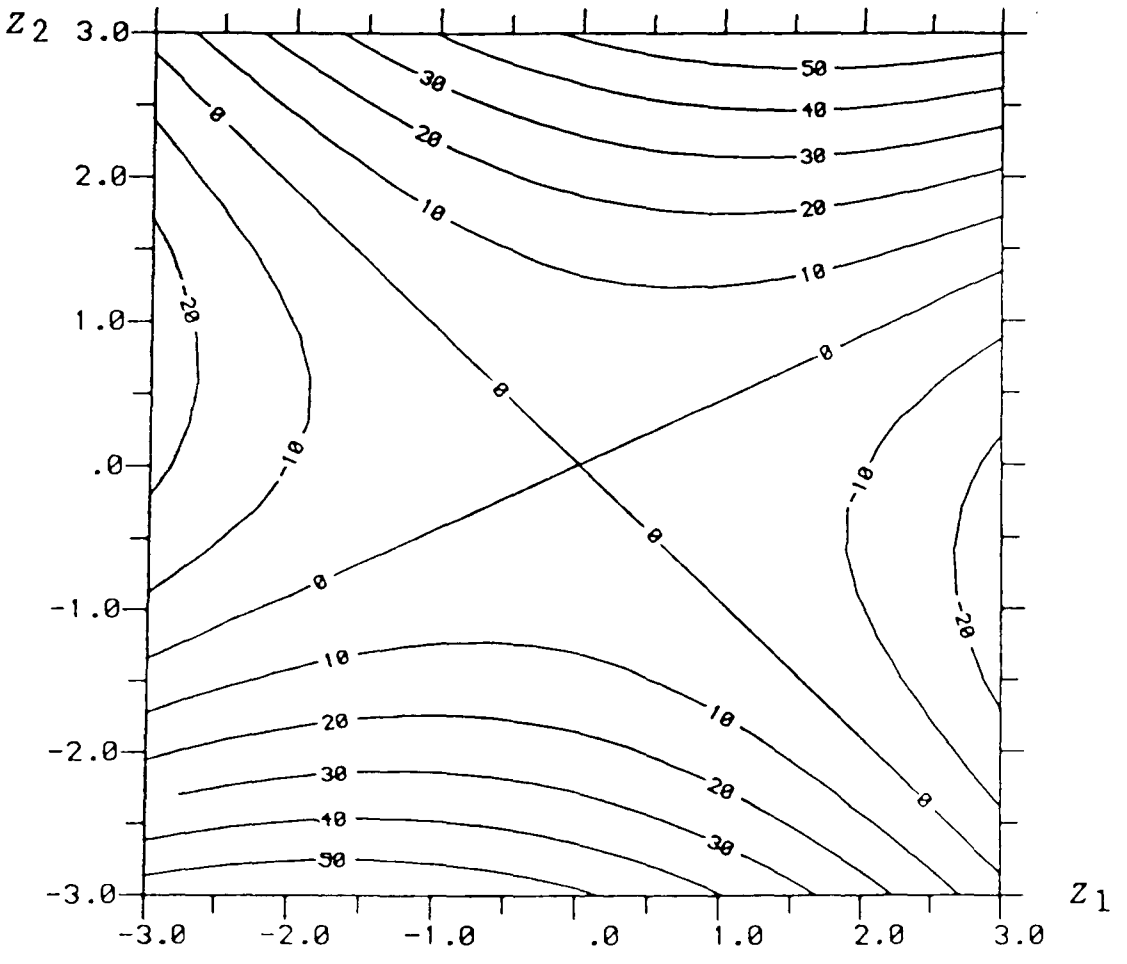


Figure 3.8.8 Plot of $TIC_R(\underline{x}, \lambda_2)$ for the matrix whose correlations are wider apart ($Z_3=0$) (contours × 10)

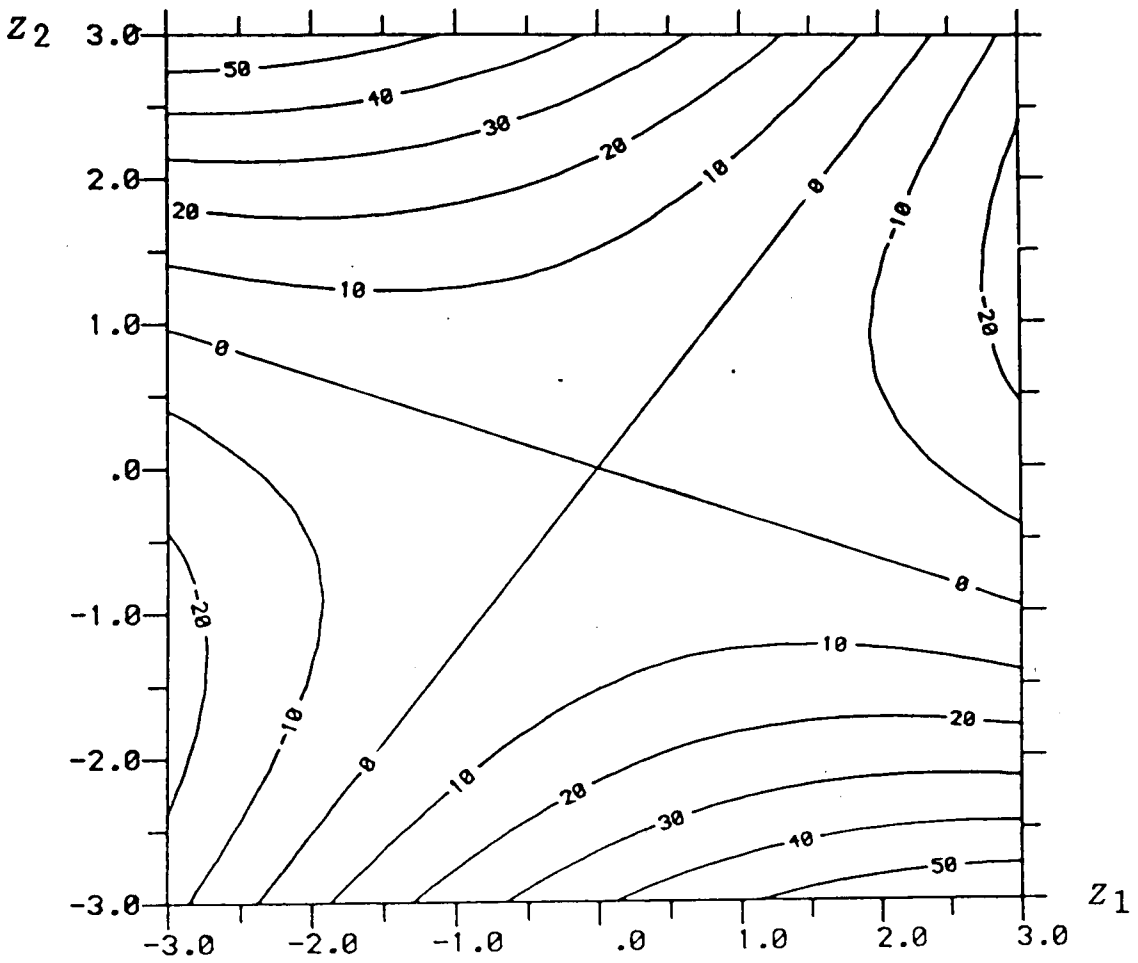


Figure 3.8.9 Plot of $TIC_R(x\ ,\lambda_3)$ for
the matrix with high correlations ($\overline{Z}_3=0$) (contours × 10)

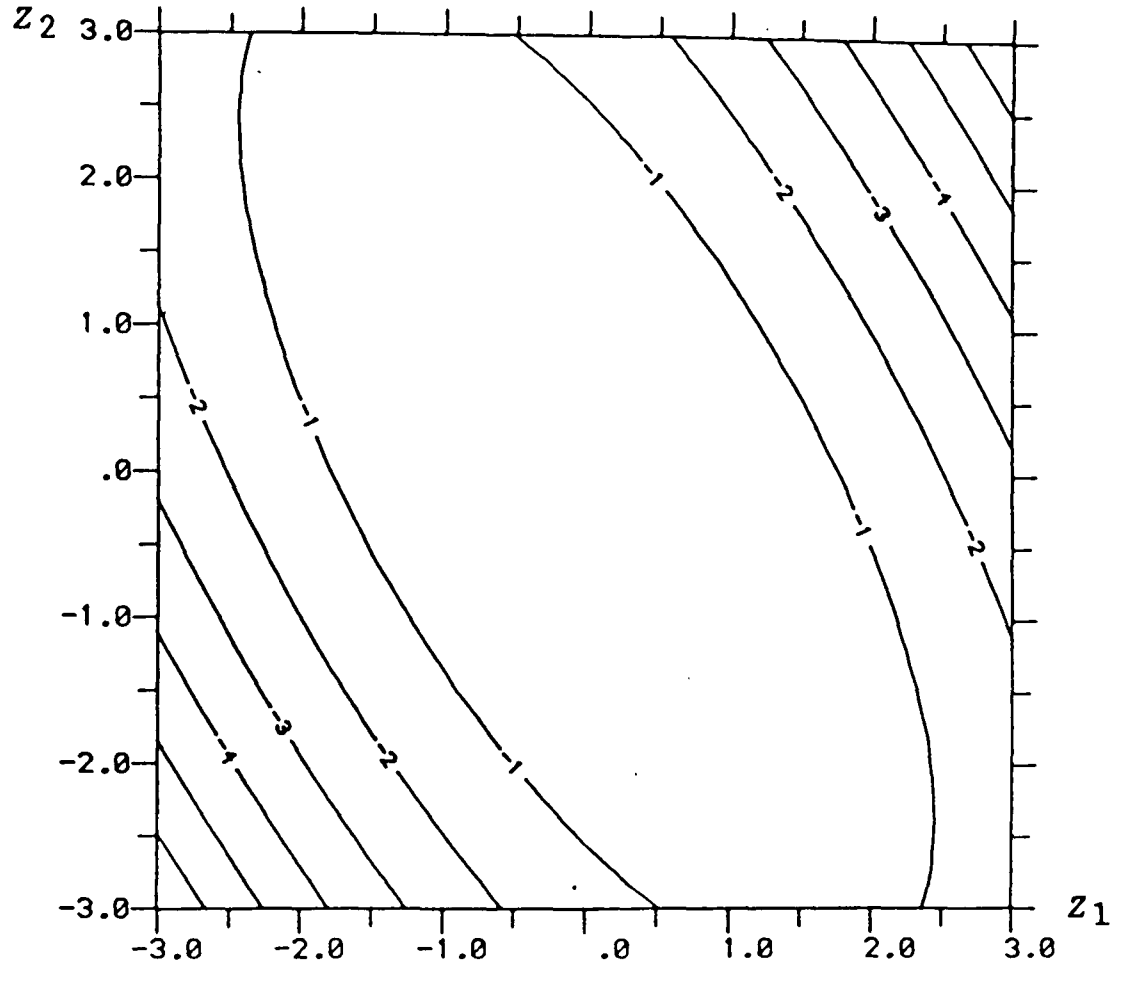


Figure 3.8.10 Plot of $TIC_R(x\ ,\lambda_3)$ for
the matrix with low correlations ($\overline{Z}_3=0$)(contours × 10)

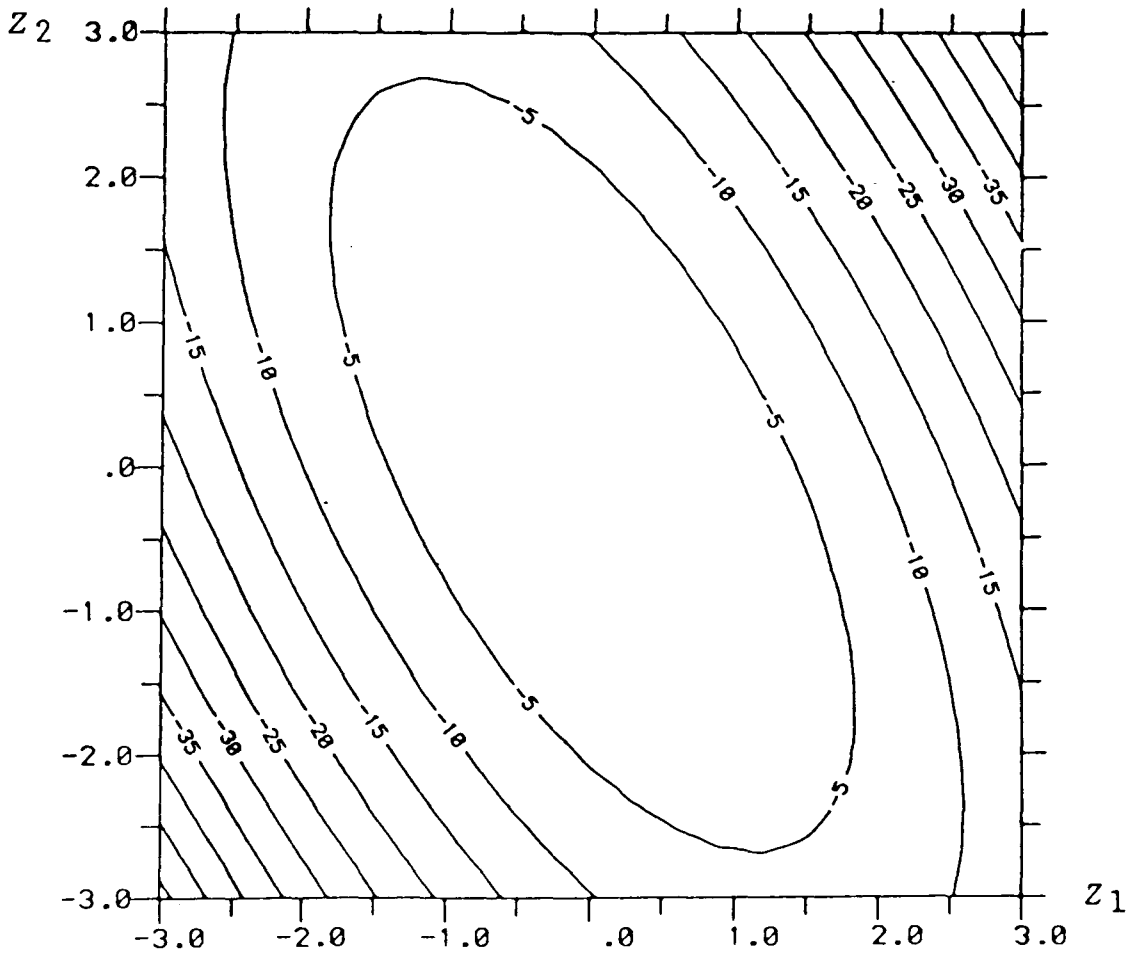
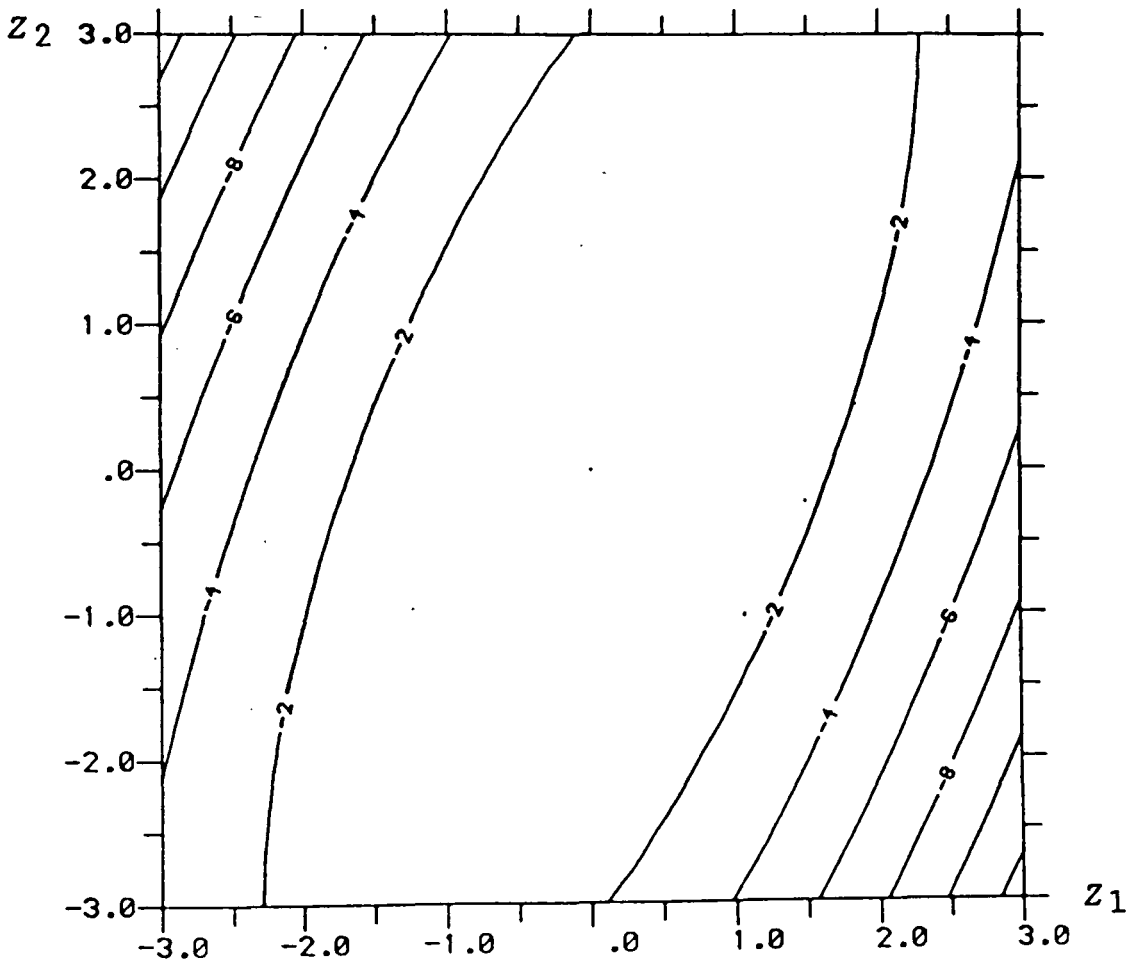


Figure 3.8.11 Plot of $TIC_R(\underline{x}, \lambda_3)$ for the matrix whose correlations are wider apart ($Z_3 = 0$) (contours × 10)



for $TIC_R(\underline{x}, \lambda_1)$ for $Z_1 = 0$. These plots are almost circular for the first two correlation matrices but more ellipsoidal for the last matrix. This may be due to the closeness of the last two eigenvalues in the first two matrices compared to the last matrix.

Fig. 3.8.12 is a contour plot for the coefficient α_{11} from the covariance matrix in (3.8.9). From (3.8.2),

$$TIC_V(\underline{x}, \alpha_{11}) = -Z_1 Z_2 (\lambda_2 - \lambda_1)^{-1} \alpha_{21}$$
$$TIC_V(\underline{x}, \alpha_{12}) = -Z_1 Z_2 (\lambda_2 - \lambda_1)^{-1} \alpha_{22}$$
$$TIC_V(\underline{x}, \alpha_{21}) = Z_1 Z_2 (\lambda_2 - \lambda_1)^{-1} \alpha_{11}$$
$$TIC_V(\underline{x}, \alpha_{22}) = Z_1 Z_2 (\lambda_2 - \lambda_1)^{-1} \alpha_{12}$$

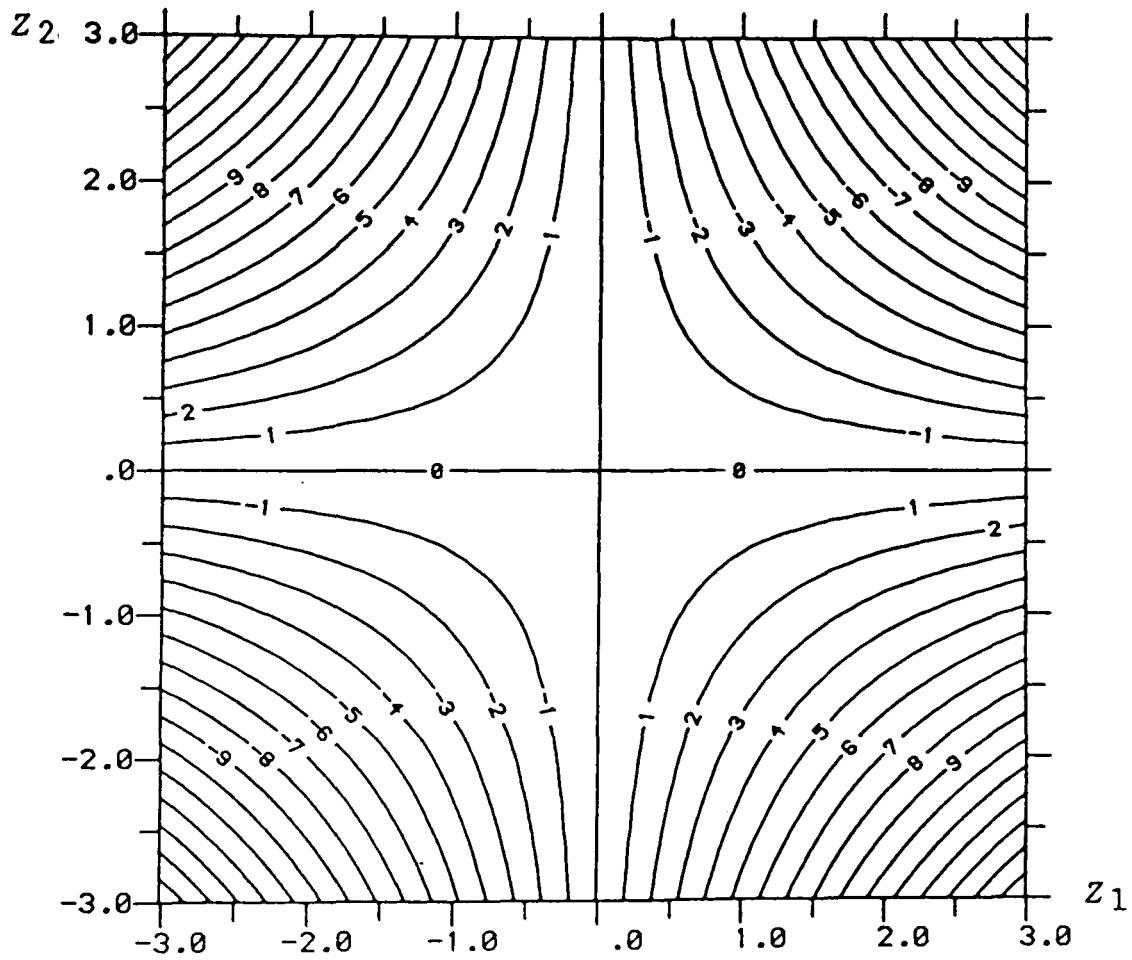
and from (3.3.11)

$$TIC_V(\underline{x}, a_{12}) = -\frac{Z_1 Z_2}{\sigma_{12}} \frac{\alpha_{21}}{\alpha_{11}}$$

so we find that the contour plots for all these coefficients are similar but with different constant terms multiplying the contours. The contour plots for the coefficients of an eigenvector from a 3×3 covariance matrix, where one of the component scores not involving the eigenvector number is set to zero, is also similar to Fig. 3.8.12. This occurs since we will only have one term in (3.8.2) which will be similar to the above expressions.

As discussed in the previous section, points placed out along an existing principal component will have no influence on the eigenvector coefficients. From the plot we see that the component score involving the eigenvector number is not more important than the other score in determining the influence on the eigenvector coefficient. Influence increases most rapidly as the two component scores increase together. This is partly due to the special nature of (3.8.2) when we have only two non-zero components as Z_1 and Z_2 play an equal role in the expression. If we look at contour plots of the eigenvectors from a $p \times p$ matrix in the first two dimensions, then as we

Figure 3.8.12 Plot of $TIC_V(\underline{x}, \alpha_{11})$ for the 2×2 covariance matrix (3.8.9) (contours $\times$ 10)



increase the values of the other component scores, rather than fix them at zero, the contours may become much straighter (although this is not always the case). We can write

$$TIC_V(\underline{x}, \alpha_{1j}) = -Z_1Z_2(\lambda_2 - \lambda_1)^{-1}\alpha_{2j} - Z_1\sum_{t=3}^{p} Z_t(\lambda_t - \lambda_1)^{-1}\alpha_{tj}$$

$$= -Z_1Z_2(\lambda_2 - \lambda_1)^{-1}\alpha_{2j} - Z_1C \quad ,$$

where C is some constant. C can be large or small depending on whether the terms in the sum add together or cancel each other out. The larger C is the more weight Z_1 will have compared to Z_2 in determining the influence of a point on the first eigenvector. Figs 3.8.13 and 3.8.14 are contour plots, in the first two dimensions, for α_{11} from the 3×3 covariance matrix,

$$\Sigma = \begin{pmatrix} 6 & 3 & 4 \\ 3 & 5 & 2 \\ 4 & 2 & 5 \end{pmatrix} \qquad (3.8.12)$$

for $Z_3 = 1$ and $Z_3 = 3$ respectively. Thus, if we were looking for influential points on the covariance eigenvectors by examining 2 way plots of the principal components it may not be obvious what would be influential due to the changing nature of these contour plots. The numerical evaluation of the influence functions for certain observations is the best way to detect influential points, and this will be discussed in detail in the next chapter.

No contour plots for the correlation eigenvectors are presented here. Little pattern was found to the contours. Sometimes they were found to be similar to plot for the corresponding eigenvalue and sometimes not. However, as noted in the last section the eigenvectors may change if we add along an existing component, unlike the covariance eigenvectors.

3.8.4. Comparisons with Other Influence Techniques

Greenacre (1984,§ 8.1) gives measures, based on the work of Escofier

Figure 3.8.13 Plot of $TIC_V(x, \alpha_{11})$ for the 3×3 covariance matrix (3.8.12) $\overline{(Z_3 = 1)}$ (contours × 10)

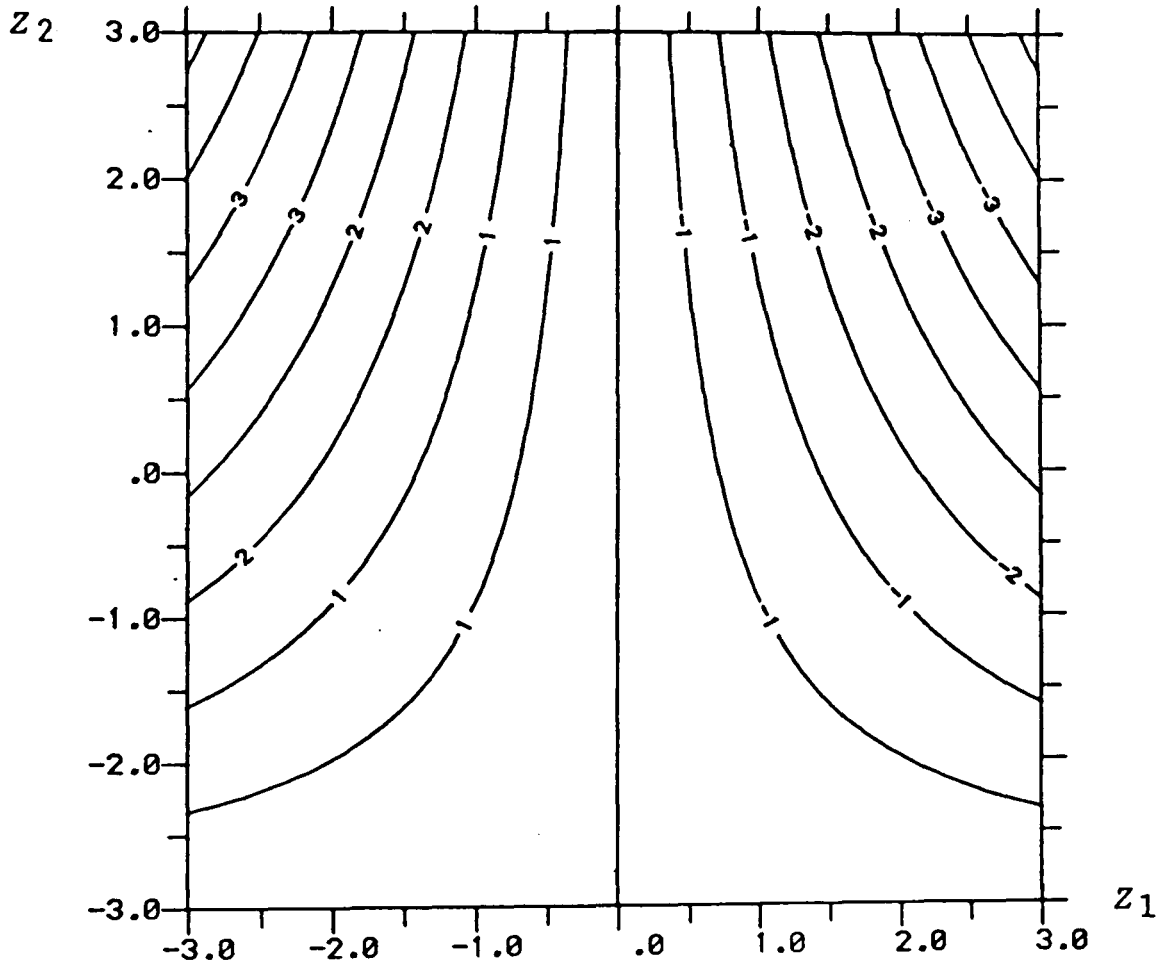
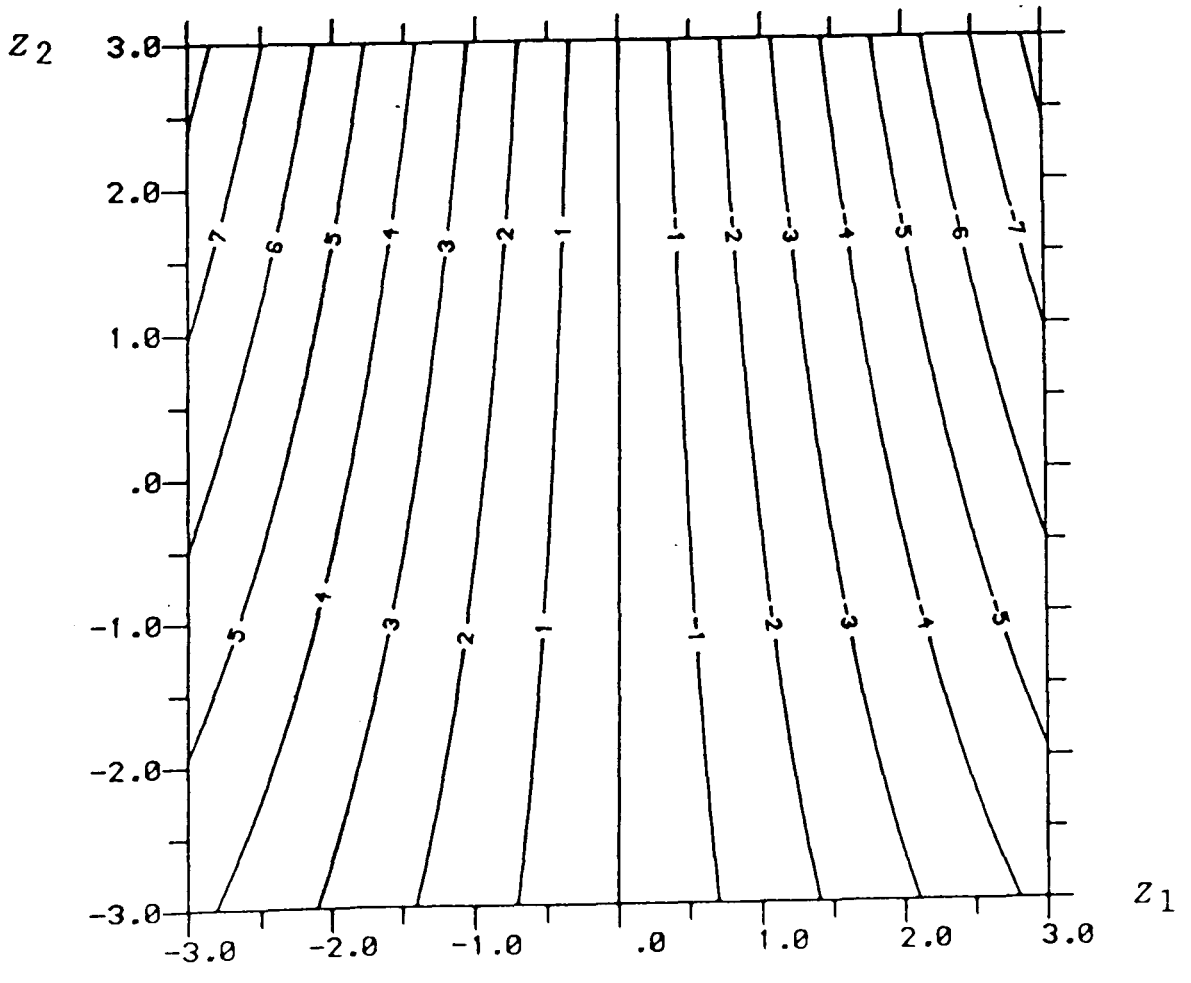


Figure 3.8.14 Plot of $TIC_V(x, \alpha_{11})$ for the 3×3 covariance matrix (3.8.12) $\overline{(Z_3 = 3)}$ (contours × 10)



and Le Roux (1976), for the affect of observations on eigenvalues and eigenvectors in general. For the eigenvalues from a covariance or correlation principal component analysis (using the divisor n so that $\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} 1/n = 1$) this specialises to,

$$I(\underline{x}, \lambda_k) = \frac{1}{n-1}\lambda_k - \frac{n}{(n-1)^2}\hat{Z}_{ki}^2 \quad , \tag{3.8.13}$$

where $\hat{Z}_{ki}$ is the kth sample principal component score of the ith observation, and I represents that it is an influence measure not based on the influence function. Expression (3.8.13) is the same as the empirical influence function, divided through by a function of n, for the covariance eigenvalue (obtained by substituting the sample equivalents into the theoretical influence function). However, it is not like the empirical expression for the correlation eigenvalues. Expression (3.8.13) has a different sign to the empirical as we have subtracted the original from the perturbed. Expression (3.8.13) comes about as follows. The sample variance of the n principal component scores, $\hat{Z}_{k1},.....\hat{Z}_{kn}$ is $\hat{\lambda}_k$, the kth eigenvalue, and when we omit the ith observation the perturbed variance of the fixed set of principal component scores using (2.2.4), but on divisor n is,

$$\tilde{V} = \frac{n}{n-1}\hat{\lambda}_k - \frac{n}{(n-1)^2}\hat{Z}_{ki}^2 \quad .$$

Subtracting the original eigenvalue (variance) from this gives expression (3.8.13). The theoretical influence function for the correlation eigenvalues is more complicated than the influence measure (3.8.13) which allows all the different variances (eigenvalues) to change independently. However, we need to maintain the constant sum of p. Greenacre (1984) notes that the above measure does not account for any changes in the metrics, which for PCA from the correlation matrix would be the standard deviations. The specialisation of the expression in Greenacre (1984) to correspondence analysis will be

discussed in § 6.3.1 and § 6.3.4.

The influence of points on the eigenvectors is given as upper bounds for the rotation of the axis when an observation is removed. For PCA on the covariance or correlation matrix this results in,

$$h = \frac{\frac{1}{n-1}\sum_{s=k}^{p} \hat{Z}^2_{si}}{(\lambda_k - \lambda_{k+1})} \tag{3.8.14}$$

with the simplest upper bound as,

$$\sin 2\phi \leq h \quad .$$

Often $h \geq 1$, so putting $h = 1$ results in a bound of $\phi = 45^o$. A more refined bound is given by,

$$h \geq 1: \quad \tan 2\phi \leq h \sin 2\theta_{sk} / (1 - h\cos^2\theta_{sk})$$

$$h < 1: \quad \tan 2\phi \leq h \sin 2\theta_{sk} / (1 - h\cos 2\theta_{sk}) \tag{3.8.15}$$

where,

$$\cos^2\theta_{sk} = \frac{\hat{Z}^2_{ki}}{n\sum_{s=1}^{p} \hat{Z}^2_{si}} \quad . \tag{3.8.16}$$

Greenacre notes that these bounds are only approximate and they assume that the principal components previous to the kth are negligibly rotated by omitting the point $\underline{x}_i$. In practice this will not usually hold particularly for the later components. In particular, the term $(\hat{\lambda}_k - \hat{\lambda}_{k+1})^{-1}$ in (3.8.14) shows the angle for the kth axis will be large when the kth and $(k+1)$th eigenvalues are close, but the above measure, unlike the theoretical influence function, would not record large changes in the $k+1$th eigenvector as well. One disadvantage of the bounds above is we cannot find them for the last eigenvector, which we may be interested in as it often defines near constant relationships between the variables. The bounds will be examined further in § 4.9, where they are

applied to a dataset for a covariance PCA. We will see that the bounds are good for the first two dimensions but less so for the latter.

Krzanowski (1984) considers an alternative form of influence, called sensitivity analysis, which is not based on case deletion. Expressions are obtained for the maximum change in the coefficients of the eigenvector $\hat{\underline{\alpha}}_k$ given a small increase (or decrease), ϵ , in the corresponding eigenvalue $\hat{\lambda}_k$. The angle between the original eigenvector and the vector which differs maximally from it, but with a variance at most ϵ less than it is given by,

$$\cos\theta = \left[1 + \frac{\epsilon}{\lambda_k - \lambda_{k+1}}\right]^{-1/2} \quad . \tag{3.8.17}$$

Similarly, the angle between the original eigenvector and that which differs maximally from it , but with a variance at most ϵ greater than it is,

$$\cos\theta = \left[1 + \frac{\epsilon}{\lambda_{k-1} - \lambda_k}\right]^{-1/2} \quad . \tag{3.8.18}$$

We again see that the closeness of the eigenvalues is important in determining the sensitivity of the eigenvectors, but as above only the closeness of the eigenvalue in one direction is involved.

Benasseni (1985) derives upper and lower bounds for the change in the eigenvalues from a covariance and correlation matrix when the frequency of the ith observation is increased or decreased. Adding an observation, when all the observations have the same frequency, is a special case of an increase in the frequency of the $(n+1)$th observation, with the original frequency, $f_{n+1} = 0$, and the perturbed frequency, $f^*_{n+1} = 1/(n+1)$. Similarly, for omitting an observation the original frequency, $f_i = 1/n$, and the perturbed frequency is $f^*_i = 0$. The bounds for the eigenvalues from a covariance matrix when we add an observation specialise to,

$$\frac{n}{n+1}\hat{\lambda}_k \leq \hat{\lambda}^*_k \leq \frac{n}{n+1}\hat{\lambda}_k + \frac{n}{(n+1)^2}(\underline{x}_{n+1} - \underline{\bar{x}})'(\underline{x}_{n+1} - \underline{\bar{x}}) \tag{3.8.19}$$

where $\hat{\lambda}_k^*$ is the perturbed eigenvalue.

From our derivation of the theoretical influence function we can write our perturbed eigenvalue as,

$$\lambda_k^* = \lambda_k + \epsilon(-\lambda_k + Z_k^2) + o(\epsilon^2) \quad .$$

Letting $\epsilon = 1/(n+1)$ and putting in the sample equivalents gives,

$$\hat{\lambda}_k^* = \hat{\lambda}_k + \frac{1}{n+1}(-\hat{\lambda}_k + \hat{Z}_k^2) + o(\epsilon^2) \quad .$$

To $o(\epsilon)$

$$= \frac{n}{n+1}\hat{\lambda}_k + \frac{1}{n+1}\hat{Z}_k^2 \quad .$$

Taking $\hat{Z}_k = 0$, the smallest value for the perturbed eigenvalue, $\hat{\lambda}_k$, is $\frac{n}{n+1}\hat{\lambda}_k$ which coincides with the above lower bound. The upper bound in (3.8.19) can be re-expressed as,

$$\frac{n}{n+1}\hat{\lambda}_k + \frac{n}{(n+1)^2}\sum_{j=1}^{p}\hat{Z}_j^2$$

since

$$\Gamma'(\underline{x}_{n+1} - \underline{\bar{x}}) = \underline{\hat{Z}}_{n+1} \quad \text{and} \quad \Gamma'\Gamma = \Gamma\Gamma' = I \quad .$$

If $\hat{Z}_j = 0$ for $j \neq k$ then the bound is similar to the theoretical expression and if only $\hat{Z}_k = 0$ then the two differ, as they do for any value of $\hat{Z}_k$, as the bound considers the other dimensions. The bounds do seem generous as the lower bound was found to be independent of the observation and the upper bound accounts for the decomposition of $\underline{x}_{n+1}$ on all the axes. Another upper bound is also derived that involves the Mahalanobis distance of the observation.

The bounds for adding an observation for the correlation matrix eigenvalues specialise to,

$$\hat{\lambda}_k \min_j(1 + \frac{1}{n+1}y_{ij}^2)^{-1} \leq \hat{\lambda}_k^* \leq (\hat{\lambda}_k + \frac{1}{n+1}\underline{y}'\underline{y}) \max_j(1 + \frac{1}{n+1}y_{ik}^2)^{-1}$$

where y is the standardised variable. Again, another upper bound is given.

These bounds to not compare easily with the theoretical expressions as both are quite complicated. The bounds are also used by Benasseni to give conditions on the variables so that the change in an eigenvalue does not exceed some specified level when the frequency is increased from $f_i = 1/n$.

3.9. Influence Functions for Adding m Observations

We usually have two options open to us when we wish to look for multiple outliers or subsets of influential points. The first is to carry out the techniques for finding individual outliers (influential points) sequentially, but due to 'masking' we may prefer to use block procedures. These involve the deletion of m observations together but it can lead to heavy computational problems due to the number of possible subsets one can consider for each value of m. Also it can be difficult to decide on the value of m. In PCA we find that a good estimate of the sample change when we add (or omit) m observations, provided m is reasonably small compared to n, is the sum of the m individual changes when the m observations are added to (or omitted from) the analysis singularly. In this section we will present the algebra that gives this result and show to what extent it is likely to be supported in samples. Numerical comparisons of the sum of the individual sample changes with the sample change when all m observations are omitted together are given in § 4.5

If we perturb the underlying distribution by adding in the distribution functions $\delta_{\underline{x}_1}, \ldots\ldots \delta_{\underline{x}_m}$ (where the subscript on $\underline{x}$ is for convenience) which have mass 1 at $\underline{x}_i$ then as a simple extension of the result (2.3.3)

$$\tilde{\Sigma} = (1 - \epsilon m)\Sigma + \epsilon \sum_{i=1}^{m} (\underline{x}_i - \underline{\mu})(\underline{x}_i - \underline{\mu})'$$

which gives the influence function

$$TIC(\underline{x}_m, \Sigma) = -m\Sigma + \sum_{i=1}^{m} (\underline{x}_i - \underline{\mu})(\underline{x}_i - \underline{\mu})' \quad . \tag{3.9.1}$$

We obtain the theoretical influence function for λ_k from the covariance matrix, as by (3.5.3), giving

$$TIC(\underset{\sim}{x}_m, \lambda_k) = -m\lambda_k + \sum_{i=1}^{m} Z^2_{ki} \quad . \tag{3.9.2}$$

Thus, the effect of omitting m observations is the same as the sum of the individual effects. When $m = 1$ we noted that the empirical influence function for the covariance matrix and sample function were very similar, see § 2.2.2, and it is partly for this reason that we will see in the next chapter how well the empirical can be used to estimate the sample changes in eigenvalues and eigenvectors from the covariance matrix. We will examine here how well (3.9.1) approximates the sample influence function when we add (or omit) m observations. This will provide some insight into how well the additive property for the eigenvalues will hold in practice. The sample covariance matrix with m extra observations included is,

$$S_{(+m)} = \frac{n-1}{n+m-1} S + \frac{1}{n+m} \sum_{i=n+1}^{n+m} (\underset{\sim}{x}_i - \underline{\bar{x}})(\underset{\sim}{x}_i - \underline{\bar{x}})'$$
$$- \frac{2}{(n+m)(n+m-1)} \sum_{i=n+1}^{n+m} \sum_{\substack{j=n+1 \\ j \neq i}}^{n+m} (\underset{\sim}{x}_i - \underline{\bar{x}})(\underset{\sim}{x}_j - \underline{\bar{x}})' \tag{3.9.3}$$

where $\underset{\sim}{x}_{n+1}, \ldots\ldots \underset{\sim}{x}_{n+m}$ are the added observations. Similarly,

$$S_{(-m)} = \frac{n-1}{n-m-1} S - \frac{n-m+1}{(n-m)(n-m-1)} \sum_{i \in M} (\underset{\sim}{x}_i - \underline{\bar{x}})(\underset{\sim}{x}_i - \underline{\bar{x}})'$$
$$- \frac{2}{(n-m)(n-m-1)} \sum_{i \in M} \sum_{\substack{j \in M \\ j \neq i}} (\underset{\sim}{x}_i - \underline{\bar{x}})(\underset{\sim}{x}_j - \underline{\bar{x}})' \tag{3.9.4}$$

where $\underset{\sim}{x}_i \; i \in M$ are the m observations omitted from the dataset. These expressions specialise to (2.2.1) and (2.2.4) respectively when $m = 1$ as the final terms in each expression will not exist. We shall prove result (3.9.3), the proof for (3.9.4) follows in a similar way.

$$(n+m-1)S_{+m} = \sum_{i=1}^{n+m} (\underset{\sim}{x}_i - \underline{\bar{x}}^*)(\underset{\sim}{x}_i - \underline{\bar{x}}^*)' \quad .$$

Adding in and taking out $\underline{\bar{x}}$ from each bracket and multiplying out gives,

$$= \sum_{i=1}^{n} (\underline{x}_i - \underline{\bar{x}})(\underline{x}_i - \underline{\bar{x}})' + \sum_{i=n+1}^{n+m} (\underline{x}_i - \underline{\bar{x}})(\underline{x}_i - \underline{\bar{x}})'$$
$$- (n+m)(\underline{\bar{x}} - \underline{\bar{x}}^*)(\underline{\bar{x}} - \underline{\bar{x}}^*)' \quad .$$

Substituting in

$$\underline{\bar{x}}^* = \frac{n\underline{\bar{x}} + m\underline{\bar{x}}_m}{n+m}$$

where $\underline{\bar{x}}_m$ is the mean of the m added observations so that

$$\underline{\bar{x}} - \underline{\bar{x}}^* = \frac{m(\underline{\bar{x}} - \underline{\bar{x}}_m)}{n+m} = -\sum_{i=n+1}^{n+m} \frac{(\underline{x}_i - \underline{\bar{x}})}{n+m}$$

gives,

$$S_{+m} = \frac{n-1}{n+m-1} S + \frac{1}{n+m-1} \sum_{i=n+1}^{n+m} (\underline{x}_i - \underline{\bar{x}})(\underline{x}_i - \underline{\bar{x}})'$$
$$- \frac{1}{(n+m)(n+m-1)} \sum_{i=n+1}^{n+m} (\underline{x}_i - \underline{\bar{x}}) \sum_{i=n+1}^{n+m} (\underline{x}_i - \underline{\bar{x}})' \quad .$$

This then gives result (3.9.3).

The first term in (3.9.3) can be expressed as $(1 - \frac{m}{n+m-1})S$ and in (3.9.4) as $(1 + \frac{m}{n-m-1})S$, and taking the sample influence functions as $(n+m)(S_{(+m)} - S)$ and $(n-m)(S - S_{(-m)})$ respectively gives

$$SIC_{(+m)}(\underline{x}_m, S) = -\frac{m(n+m)}{n+m-1} S + \sum_{i=n+1}^{n+m} (\underline{x}_i - \underline{\bar{x}})(\underline{x}_i - \underline{\bar{x}})'$$
$$- \frac{2}{(n+m-1)} \sum_{i=n+1}^{n+m} \sum_{\substack{j=n+1 \\ j \neq i}}^{n+m} (\underline{x}_i - \underline{\bar{x}})(\underline{x}_j - \underline{\bar{x}})' \qquad (3.9.5)$$

$$SIC_{(-m)}(\underline{x}_m, S) = -\frac{m(n-m)}{n-m-1} S + \frac{n-m+1}{(n-m-1)} \sum_{i \in M} (\underline{x}_i - \underline{\bar{x}})(\underline{x}_i - \underline{\bar{x}})'$$
$$+ \frac{2}{(n-m-1)} \sum_{i \in M} \sum_{\substack{j \in M \\ j \neq i}} (\underline{x}_i - \underline{\bar{x}})(\underline{x}_j - \underline{\bar{x}})' \quad . \qquad (3.9.6)$$

Comparing (3.9.5) and (3.9.6) with (2.2.3) and (2.2.5) respectively, we see that to the first order, the sample curves (apart from slight changes in the functions of n) exhibit the same additive structure as the theoretical influence

function. However, there is also an extra term of $o(1/n)$ in the above sample influence functions which will affect the accuracy of the estimated change based on (3.9.2) when we delete m observations. Comparisons of how well this additive approximation holds are given in § 4.5. There seems to be a greater restriction on the size of m in the deleted sample function for this additive property to hold as the second order term in (3.9.6) has the coefficient $1/(n-m-1)$ multiplying it, and this increases as m increases.

We can look at this additive nature from an alternative angle, but we will restrict our attention to $m=2$ for illustration purposes. In Section 3.6.3 we derived the influence function for the principal component scores from the covariance matrix when we add the point $\underline{x}$. We would expect this to affect the additive nature of the theoretical influence for the covariance eigenvalues whose influence functions involve the component scores since, if we add one point and then the other, the (perturbed) score of the point added second will not be the same as if it was added first. If $\tilde{\lambda}_k$ is the perturbed eigenvalue when we add the first point $\underline{x}_1$ then from (3.6.1)

$$\tilde{\lambda}_k = \lambda_k + \epsilon(-\lambda_k + Z_{k1}^2) + o(\epsilon^2) \tag{3.9.7}$$

where Z_{k1} is the principal component score. If we now let $\tilde{\lambda}^*_k$ denote the perturbed $\tilde{\lambda}_k$ when we add the second point $\underline{x}_2$ then,

$$\tilde{\lambda}^*_k = -\tilde{\lambda}_k + \epsilon(-\tilde{\lambda}_k + \tilde{Z}_{k2}^2) + o(\epsilon^2) \tag{3.9.8}$$

where $\tilde{Z}_{k2}$ is the score of the second point on the new axis after $\underline{x}_1$ has been included. Putting (3.9.7) into (3.9.8) we obtain by writing

$$\tilde{Z}_{k2}^2 = Z_{k2}^2 + 2\epsilon Z_{k2} TIC(\underline{x}_1, Z_{k2})$$
$$\tilde{\lambda}^*_k = \lambda_k + \epsilon(-2\lambda_k + Z_{k1}^2 + Z_{k2}^2) + o(\epsilon^2) \quad .$$

This means the change in the scores, if we were using the theoretical for estimating the changes when adding or deleting m observations, only affect the perturbed λ_k to $o(\epsilon^2)$ and so will not come into the theoretical influence

function. We thus get that the theoretical influence function for the eigenvalues is additive.

We similarly find that the theoretical influence function for the eigenvectors is additive, and from (3.9.1) we can write

$$TIC(\underset{\sim}{x}_m, \rho_{kj}) = -\frac{\rho_{kj}}{2} \sum_{i=1}^{m} y^2_{ik} y^2_{ij} - \sum_{i=1}^{m} y_{ik} y_{ij} \quad .$$

The theoretical influence function for the correlation matrix is also additive and so we find the theoretical influence function for its eigenvalues and eigenvectors are also additive.

The additivity of the theoretical influence functions has interesting properties. We noted in the previous sections that the correlation eigenvalues, particularly on the larger eigenvalues, often have as large positive influences as negative. Thus, when we delete two or more influential points they can cancel out each others effects. This, is less true for eigenvalues from the covariance matrix were the largest influences all tend to be of the same sign and correspond to large principal component score. Some examples of this cancelling out of influence will be seen in our practical examples in § 4.5. The additive property means we should not need to carry out multiple procedures in practice, particularly where we have large datasets. This is extremely advantageous in such datasets where there would be heavy computational problems. The above cancelling out property shows it could in fact be confusing to look at multiple case deletion. In small datasets where the asymptotic result may not hold so well there may be some need to consider multiple case deletion using the sample influence curve. This additivity is not likely to hold when two observations are extreme in the direction of a principal component so that this component disappears when both are omitted. We will see in § 4.8 that the empirical does not reflect the sample

changes well when one observation causes a dimension to disappear. However, it still gives this observation as being 'highly influential'.

Chapter 4: Practical Applications of the Influence Functions in Principal Component Analysis

4.1. Introduction

In this chapter we concentrate on the practical applications of the influence functions in PCA. We first consider, in § 4.2, an appropriate scalar measure of influence for the changes in the coefficients of the k th eigenvector. This is chosen to be the angle between the original and perturbed eigenvectors, which can be written in terms of the the sample influence function for the eigenvectors. We shall see that the empirical influence function for the eigenvectors provides a second order approximation to this angle. In § 4.3 we compare the empirical and sample curves for the 'actual' change in the eigenvalues and the angle discussed above. Comparisons for the eigenvalues and eigenvectors from both the covariance and correlation matrices are considered. We shall see that the influences compare well, but they often differ when we have close eigenvalues or if a dimension should completely disappear when an observation is deleted. In § 4.4 we shall consider the use of second order terms to improve the accuracy of the empirical, with emphasis on the covariance matrix eigenvalues and eigenvectors. In § 3.9 we gave the theoretical justification for influence being additive. This was supported for the sample covariance matrix provided m is not too large compared to n. We compare the sample influence when m observation are omitted with the sum of the individual sample changes, for both the covariance and correlation eigenvalues, in § 4.5.

In § 4.6 we discuss simulated critical values for the covariance and correlation eigenvalues. For the eigenvalues from the covariance matrix these are seen to take a fairly simple form. The influence values for the eigenvectors are largely determined on how close the eigenvalues are, due to the terms $(\lambda_j - \lambda_k)^{-1}$. It is possible if there are a number of close eigenvalues that the influences could be small if the terms cancelled each other out.

Alternatively, and as discussed in § 4.3, if there are only two close eigenvalues we can obtain large changes in the two corresponding eigenvectors. Hence, determining critical values for the changes in the eigenvectors could be difficult.

This chapter is concluded with an examination of influence in three datasets. In the first dataset emphasis is placed on the differences and similarities of the outliers and influential points that have been detected in the dataset (this is continued to a certain extent for the other datasets as well). Only the principal components for the correlation matrix are considered in this dataset as the data are standardised. Also considered in this dataset is how well changes in the bivariate correlations can be used to indicate what will be influential on the principal component analysis. In the second dataset the influence of observations in both the covariance and correlation principal component analyses are considered. We shall consider the changes of one particular observation, which when omitted leads to a dimension disappearing. Influence for the covariance and correlation principal component scores is also looked at. In the final dataset we consider the covariance principal component analysis only, and consider its application to the covariance biplot.

4.2. Measures of Influence

Influence procedures for one principal component analysis, if one looks at both the eigenvalues and eigenvectors, can result in many measures to examine. The problem is enhanced by the vector valued influence functions for the eigenvectors. However, one usually only retains a fraction of the p dimensions so we would restrict our influence procedures to those dimensions of interest. It is not really worthwhile to compute an overall measure of influence for an observation on the principal component analysis since important features of change could be lost when combining the dimensions.

This is particularly important in principal component analysis where we usually interpret the dimensions individually and so influence measures applied to each dimension singularly provide invaluable information on our interpretations and conclusions. However, it is worthwhile to convert the p-vector influence functions for the eigenvectors into scalar measures. One possible scalar measure is the sums of squares of the individual changes in the eigenvector coefficients. The sums of squares for the covariance eigenvectors using the theoretical influence function (3.8.2) is,

$$TIC_V(\underline{x}\,,\underline{\alpha}_k)'TIC_V(\underline{x}\,,\underline{\alpha}_k) = Z_k^2 \sum_{\substack{j=1 \\ j\neq k}}^{p} (\lambda_j - \lambda_k)^{-2} Z_j^2 \quad .$$

Similarly, for the correlation eigenvectors we get,

$$TIC_R(\underline{x}\,,\underline{\alpha}_k)'TIC_R(\underline{x}\,,\underline{\alpha}_k) = \sum_{\substack{j=1 \\ j\neq k}}^{p} (\lambda_j - \lambda_k)^{-2} \left[\sum_{s=1}^{p} \sum_{\substack{t=1 \\ t\neq s}}^{p} \alpha_{ks}\alpha_{kt} TIC(\underline{x}\,,\rho_{st}) \right]^2 \quad .$$

For both the covariance and correlation eigenvectors we find

$$1 - \frac{\epsilon^2}{2} TIC(\underline{x}\,,\underline{\alpha}_k)'TIC(\underline{x}\,,\underline{\alpha}_k)$$

is the cosine of the angle between the original and perturbed eigenvector up to terms in $o(\epsilon^2)$.

First we will consider the sample results. If $\hat{\underline{\alpha}}_k^*$ is the perturbed sample eigenvector and

$$\frac{1}{n-1} SIC_V(\underline{x}\,,\hat{\underline{\alpha}}_k) = \hat{\underline{\alpha}}_k - \hat{\underline{\alpha}}_k^*$$

then

$$\frac{1}{(n-1)^2} SIC_V(\underline{x}\,,\hat{\underline{\alpha}}_k)'SIC_V(\underline{x}\,,\hat{\underline{\alpha}}_k) = \hat{\underline{\alpha}}_k'\hat{\underline{\alpha}}_k + \hat{\underline{\alpha}}_k^{*\prime}\hat{\underline{\alpha}}_k^* - 2\hat{\underline{\alpha}}_k'\hat{\underline{\alpha}}_k^*$$

and since $\hat{\underline{\alpha}}_k'\hat{\underline{\alpha}}_k = 1$ and $\hat{\underline{\alpha}}_k^{*\prime}\hat{\underline{\alpha}}_k^* = 1$ we get,

$$\frac{1}{(n-1)^2} SIC_V(\underline{x}\,,\hat{\underline{\alpha}}_k)'SIC_V(\underline{x}\,,\hat{\underline{\alpha}}_k) = 2 - 2\cos\theta_k$$

$$\Rightarrow \quad \cos\theta_k = 1 - \frac{1}{2(n-1)^2} SIC_V(\underline{x}\,,\hat{\underline{\alpha}}_k)'SIC_V(\underline{x}\,,\hat{\underline{\alpha}}_k) \quad .$$

For the theoretical we only consider terms up to $o(\epsilon^2)$. Let the perturbed eigenvector be,

$$\tilde{\underline{\alpha}}_k = \underline{\alpha}_k + \epsilon \underline{c}_1 + \frac{\epsilon^2}{2}\underline{c}_2 + o(\epsilon^3)$$

then,

$$\tilde{\underline{\alpha}}_k' \tilde{\underline{\alpha}}_k = \underline{\alpha}_k' \underline{\alpha}_k + 2\epsilon \underline{\alpha}_k' \underline{c}_1 + \epsilon^2 \left[\underline{c}_1' \underline{c}_1 + \underline{\alpha}_k' \underline{c}_2 \right] \quad .$$

Since $\underline{\alpha}_k' \underline{\alpha}_k = 1$ to have $\tilde{\underline{\alpha}}_k' \tilde{\underline{\alpha}}_k = 1$ all terms in $\epsilon^r = 0 \;\; r = 1,2....$, this implies that,

$$\underline{\alpha}_k' \underline{c}_1 = 0 \;\text{ and }\; \underline{\alpha}_k' \underline{c}_2 = -\underline{c}_1' \underline{c}_1 \quad .$$

The angle is given by,

$$\begin{aligned} \cos\theta_k &= \underline{\alpha}_k' \tilde{\underline{\alpha}}_k \\ &= \underline{\alpha}_k' \underline{\alpha}_k + \epsilon \underline{\alpha}_k' \underline{c}_1 + \frac{\epsilon^2}{2} \underline{\alpha}_k' \underline{c}_2 + o(\epsilon^3) \quad . \end{aligned}$$

Using the relationships above

$$\begin{aligned} &= 1 - \frac{\epsilon^2}{2} \underline{c}_1' \underline{c}_1 + o(\epsilon^3) \\ &= 1 - \frac{\epsilon^2}{2} TIC(\underline{x}, \underline{\alpha}_k)' TIC(\underline{x}, \underline{\alpha}_k) + o(\epsilon^3) \quad . \end{aligned} \qquad (4.2.1)$$

Looking at the angle rather than just the sums of squares of individual coefficient changes gives a more meaningful interpretation to the changes in the eigenvectors. It is also better at highlighting rotations and swops in the eigenvectors, particularly when using the sample influence curve. One also needs to be careful when using the sums of squares of the sample changes in the eigenvector coefficients that the perturbed eigenvector is returned with the same sign as the original eigenvector. We can usually guard against this by changing the sign of the perturbed eigenvector if $p/2$ or $(p+1)/2$ of its signs are different from the original eigenvector but the odd one may slip through. This may occur when the eigenvector is dominated by one or two coefficients, and it is whether the signs of these coefficients have changed that is the most important. If a change of sign is not picked up the influences will appear

unusually large since one has in affect added the original and perturbed eigenvectors. This problem does not occur when using the theoretical influence function since the perturbed eigenvectors are not calculated. The problem of change in sign of the sample eigenvectors is avoided when we use the angular measure of influence as this just leads to an angle greater than 90° and subtracting this from 180° gives us the same angle as would be obtained if the sign had not changed.

Due to the similarity of the sample influence functions for the covariance matrix when we add or delete an observation, see § 2.2, we find that the theoretical influence functions can be used for the deletion of observations in principal component analysis as well as addition. We also noted in § 2.2.2 that if we considered the sample curve for the covariance matrix as $n \to \infty$ we would obtain the same expression as the empirical. The same holds here for the eigenvalues from the covariance matrix whose influence function from (3.5.3) only depends on the influence function for the covariance matrix. Substituting the sample influence curve for the covariance matrix for V in (3.5.3) and letting $n \to \infty$ would give the same as the empirical curves based on (3.8.1) to (3.8.3). We thus find that rather than just take $\epsilon = 1/(n-1)$ to form our estimated change from the theoretical, if we let our estimate for the sample change in the covariance eigenvalues be

$$E_V(\underline{x}_i, \hat{\lambda}_k) = -\frac{\hat{\lambda}_k}{n-2} + \frac{n}{(n-1)(n-2)}\hat{Z}_{ki}^2 \qquad (4.2.2)$$

then the estimated change in the covariance eigenvalues gives closer values to the actual sample change. The functions of n in (4.2.2) come from those in (2.2.5) when it is divided through by $(n-1)$ to give the actual sample change in the sample covariance matrix. The actual sample change for the eigenvalues and eigenvectors is found by using the expression for the sample change in the covariance matrix and then using the eigenvalue and

eigenvector routines. No adjustments to the other empirical influence functions lead to a noticeable improvement in the estimates.

Our measures of influence are thus taken as the change in the individual eigenvalues when observations are omitted from the sample and the angle between the original and perturbed eigenvectors which can be calculated from the influence functions. A possible alternative to the change in the eigenvalues is the percentage change. We shall see in § 4.6 that for all the eigenvalues in a covariance principal component analysis we can expect the same percentage change irrespective of their initial values. This is not true for the correlation eigenvalues.

4.3 Numerical Comparisons of the Sample and Empirical Functions and Problems Arising from Close Eigenvalues

We will compare the empirical and sample curves by calculating the estimated and actual sample changes in the eigenvalues, and the actual and estimated angle between the original and perturbed eigenvectors, when an observation is removed, for two datasets. The estimated change in the covariance eigenvalues is given by (4.2.2) and for the correlation eigenvalues by the empirical based on (3.8.4) divided by $(n-1)$. The estimated angle is taken from (4.2.1), with the sample equivalents substituted in and $\epsilon = 1/(n-1)$. Both the covariance and correlation principal component analyses will be considered. The first dataset has 160 observations and five variables which are measurements on turtles. The other dataset was introduced in § 2.5.3 and its variables consist of seven anatomical measurements on Students at the University of Kent. There are 33 observations, so we may expect our comparisons with the asymptotic result to be poorer than for the first dataset. Tables 4.3.1 to 4.3.8 bring to light some interesting aspects of the similarities and differences between the influence

curves in principal component analysis.

4.3.1. Influence for the Eigenvalues from the Covariance and Correlation Matrices

Tables 4.3.1 and 4.3.3 compare the actual and estimated change for the three most influential and three least influential observations on each covariance eigenvalue from the two datasets. The order is determined by the ranked sample influences and the corresponding estimated change is given underneath. The rankings for the two curves can differ when we have observations with close influences or when we have problems such as eigenvectors swopping. We have considered all eigenvalues here, in practice one may only be interested in a subset of these. We have given the three smallest influences, which we also considered for the partial correlation coefficient, to show that the curves agree well for large and small influences. Tables 4.3.2 and 4.3.4 give similar comparisons for the correlation eigenvalues from the two datasets. We have only considered the most influential here, the smallest influences are very small.

From the tables we see that the estimates for the covariance and correlation matrix eigenvalues provide us with reliable information on what observations are most influential in datasets and on what eigenvalues. Comparisons for the first dataset, with 160 observations, are very good as we would hope since this is a reasonably sized dataset. However, the comparisons for the smaller dataset are also good, except for those involving observation 30, particularly in the principal component analysis from the covariance matrix.

The large differences in the empirical and sample either occur on observation 30, as noted above, or when the eigenvectors have swopped in order of importance. This means the variance (eigenvalue) in an earlier

Comparisons of the Sample and Empirical Functions for Eigenvalues from the Turtle Dataset

Table 4.3.1

Comparisons of the Actual Sample and Estimated Change for the Most and Least Influential Observations on the Covariance Eigenvalues

$\hat{\lambda}_k$ =	101845.62	177.34	51.93	29.24	13.27
Obs.	(143)	(56)	(10)	(103)	(151)
Actual	6360.46	13.69	1.55	1.25	2.15
Estimated	6361.06	13.74	1.57	1.20	2.11
Obs.	(121)	(119)	(104)	(142)	(143)
Actual	5714.93	9.62	1.49	0.71	1.20
Estimated	5715.25	9.67	1.50	0.72	1.05
Obs.	(61)	(141)	(38)	(157)	(81)
Actual	2936.10	7.83	1.33	0.71	1.13
Estimated	2936.16	7.85	1.33	0.67	1.04
" "	" "	" "	" "	" "	" "
" "	" "	" "	" "	" "	" "
" "	" "	" "	" "	" "	" "
Obs.	(83)	(58)	(62)	(74)	(78)
Actual	-53.25	0.05	0.01	-0.00	-0.00
Estimated	-53.24	0.06	0.01	-0.00	-0.00
Obs.	(151)	(46)	(73)	(93)	(159)
Actual	37.92	-0.05	0.00	-0.00	0.00
Estimated	37.94	-0.05	0.01	-0.00	-0.00
Obs.	(108)	(111)	(97)	(158)	(88)
Actual	18.25	0.03	0.00	0.00	-0.00
Estimated	18.26	0.02	0.00	-0.00	-0.00

Table 4.3.2

Comparisons of Actual Sample and Estimated Change for the Most Influential Observations on the Correlation Matrix

$\hat{\lambda}_k$ =	4.115	0.449	0.187	0.175	0.074
Obs.	(151)	(151)	(56)	(104)	(121)
Actual	-0.050	0.043	0.011	0.007	-0.003
Estimated	-0.047	0.042	0.016	0.006	-0.003
Obs.	(135)	(72)	(151)	(76)	(157)
Actual	0.022	0.018	0.008	0.007	0.002
Estimated	0.021	0.018	0.007	0.007	0.002
Obs.	(155)	(61)	(69)	(19)	(61)
Actual	0.020	0.015	0.006	0.005	-0.002
Estimated	0.020	0.015	0.006	0.003	-0.002

Comparison of the Sample and Empirical Functions for the Eigenvalues from the Student Dataset

Table 4.3.3

Comparisons of the Actual Sample and Estimated Change for the Most and Least Influential Observations on the Covariance Eigenvalues

$\hat{\lambda}_k$ =	35.31	2.78	1.41	0.86	0.68	0.38	0.25
Obs.	(19)	(25)	(31)	(30)	(24)	(27)	(30)
Actual	6.40	0.42	0.25	0.14	0.27	0.06	0.16
Estimated	6.42	0.42	0.22	0.46	0.25	0.05	0.01
Obs.	(10)	(8)	(8)	(29)	(30)	(30)	(18)
Actual	2.60	0.40	0.06	0.05	0.27	0.06	0.09
Estimated	2.62	0.44	0.03	0.09	0.03	-0.01	0.07
Obs.	(26)	(30)	(24)	(2)	(17)	(5)	(19)
Actual	1.57	0.29	0.06	-0.03	0.07	0.05	0.05
Estimated	1.58	0.43	0.10	-0.03	0.06	0.04	0.04
" "	" "	" "	" "	" "	" "	" "	" "
" "	" "	" "	" "	" "	" "	" "	" "
" "	" "	" "	" "	" "	" "	" "	" "
Obs.	(9)	(3)	(13)	(15)	(20)	(33)	(27)
Actual	-0.18	-0.02	0.00	-0.00	-0.01	-0.00	0.00
Estimated	-0.18	-0.02	0.00	-0.00	-0.01	-0.00	-0.00
Obs.	(28)	(10)	(22)	(6)	(7)	(3)	(22)
Actual	0.05	-0.01	-0.00	-0.00	0.01	0.00	0.00
Estimated	0.05	-0.01	-0.00	-0.00	0.01	-0.00	-0.00
Obs.	(1)	(32)	(23)	(5)	(18)	(16)	(29)
Actual	0.03	-0.00	0.00	-0.00	0.00	0.00	-0.00
Estimated	0.04	-0.00	0.01	-0.00	0.00	0.00	-0.00

Table 4.3.4

Comparisons of Actual Sample and Estimated Change for the Most Influential Observations on the Correlation Eigenvalues

$\hat{\lambda}_k$ =	4.80	0.93	0.40	0.37	0.28	0.13	0.10
Obs.	(30)	(30)	(24)	(27)	(30)	(19)	(18)
Actual	-0.58	0.46	-0.07	0.06	0.10	-0.02	0.03
Estimated	-0.33	0.39	-0.05	0.05	-0.04	-0.02	0.02
Obs.	(10)	(17)	(19)	(19)	(24)	(5)	(24)
Actual	0.12	-0.05	-0.03	-0.04	0.09	0.02	0.02
Estimated	0.11	-0.05	-0.02	-0.03	0.05	0.01	0.01
Obs.	(19)	(10)	(29)	(29)	(25)	(31)	(31)
Actual	0.12	-0.05	0.02	0.03	0.03	0.01	0.02
Estimated	0.10	-0.05	0.05	-0.00	0.03	0.01	0.01

direction has fallen down below that in the next direction (it would be possible for it to fall below two or more of the other eigenvalues) when observations are omitted. The estimate records a swop in the eigenvalues by some larger changes in the eigenvalues and the sample reflects this swop when we look at the eigenvectors by angles close to 90°. The theoretical influence function (and so the estimate) gives the change in the eigenvalue and eigenvector in the kth direction irrespective of whether it changes its ranked position when it is perturbed. The sample curve will compare the kth original eigenvalue and eigenvector with the perturbed $(k+1)$th eigenvalue and eigenvector when a swop occurs. This is one of the circumstances where the empirical will have a larger value than the deleted sample curve (although, using the adjustment in the functions of n for the estimated change in the covariance eigenvalues, discussed in § 4.2, we find the usual underestimation is not so obvious).

We shall look at the values of the perturbed eigenvalues given by the sample and empirical functions when swopping has occured. Note that a positive influence refers to a decrease in $\hat{\lambda}_k$, as we subtract the perturbed from the original eigenvalue. We will consider this for two examples. In Table 4.3.4 the original third and fourth eigenvalues are,

$$\hat{\lambda}_3 = 0.399 \qquad \hat{\lambda}_4 = 0.366 \quad .$$

The sample influences give the perturbed eigenvalues when observation 29 is omitted as

$$\hat{\lambda}_3^{*s} = 0.379 \qquad \hat{\lambda}_4^{*s} = 0.336 \quad .$$

The perturbed eigenvalues using the empirical curve are,

$$\hat{\lambda}_3^{*e} = 0.349 \qquad \hat{\lambda}_4^{*e} = 0.362 \quad .$$

There was another swop in these two dimensions when observation 5 was omitted but this is not seen from Table 4.3.4 as it was not in the top three ranked sample changes which the observations are presented by.

We shall consider the example from Table 4.3.2 which shows that swops can occur even in large datasets provided the original eigenvalues are close enough. The original third and fourth eigenvalues are,

$$\hat{\lambda}_3 = 0.187 \quad \hat{\lambda}_4 = 0.175 \quad .$$

The sample influences give the perturbed eigenvalues when observation 56 is omitted as

$$\hat{\lambda}_3^{*s} = 0.176 \quad \hat{\lambda}_4^{*s} = 0.171 \quad .$$

The perturbed eigenvalues using the empirical curve are,

$$\hat{\lambda}_3^{*e} = 0.171 \quad \hat{\lambda}_4^{*e} = 0.175 \quad .$$

The theoretical influence function records a swop in the eigenvalues/eigenvectors when observation 30 is omitted for the 4th and 5th dimensions in Table 4.3.3. However, this is not a good reflection of the sample change as we find the original fourth dimension is attributed to observation 30 only, and it disappears in the perturbed problem. This, and the deletion of observation 30 in the correlation PCA analysis, is discussed in detail in § 4.8.

4.3.2. Influence for the Eigenvectors from the Covariance and Correlation Matrices

Tables 4.3.5 to 4.3.8 compare the actual and estimated angle. These are given by (3.8.2) and (3.8.6). The angles for the least influential observations are not given as these are very small.

The angles using the sample and empirical curves are very close in Table 4.3.5 and Table 4.3.6 except for those under θ_3 and θ_4 for the latter table. Table 4.3.9 contains the most and least influential observations, and their angular changes, on the third and fourth eigenvectors from the correlation matrix of the Turtle dataset when using the two types of curves. The original

Comparisons of the Sample and Empirical Functions for the Eigenvectors from the Turtle Dataset

Table 4.3.5

Comparisons of the Actual and Estimated Angle Between the Original and Perturbed Eigenvectors from the Covariance Matrix for the most Influential Observations

Angle	θ_1	θ_2	θ_3	θ_4	θ_5
Obs.	(143)	(119)	(19)	(81)	(81)
Actual	0.14°	1.59°	3.04°	3.82°	3.70°
Estimated	0.13°	1.45°	2.88°	3.80°	3.65°
Obs.	(121)	(56)	(76)	(19)	(142)
Actual	0.10°	1.58°	2.32°	3.01°	2.12°
Estimated	0.10°	1.40°	2.14°	2.85°	2.00°
Obs.	(34)	(76)	(103)	(142)	(143)
Actual	0.05°	1.39°	2.19°	2.29°	1.57°
Estimated	0.05°	1.33°	2.22°	2.20°	1.43°

Table 4.3.6

Comparisons of the Actual and Estimated Between the Original and Perturbed Eigenvectors from the Correlation Matrix for the most Influential Observations

Angle	θ_1	θ_2	θ_3	θ_4	θ_5
Obs.	(151)	(151)	(56)	(56)	(56)
Actual	0.72°	5.11°	82.32°	82.32°	2.54°
Estimated	0.70°	4.23°	3.74°	3.99°	2.15°
Obs.	(61)	(143)	(151)	(151)	(119)
Actual	0.30°	5.11°	23.41°	22.94°	2.53°
Estimated	0.29°	4.48°	6.25°	4.69°	2.28°
Obs.	(72)	(81)	(19)	(19)	(76)
Actual	0.29°	2.69°	22.10°	22.11°	2.04°
Estimated	0.28°	2.49°	22.30°	22.30°	1.87°

Comparisons of the Sample and Empirical Functions for the Eigenvectors from the Student Dataset

Table 4.3.7

Comparisons of Actual and Estimated Angle Between the Original and Perturbed Eigenvectors from the Covariance Matrix for the Most Influential Observations

Angle	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7
Obs.	(19)	(30)	(24)	(30)	(30)	(30)	(30)
Actual	1.74°	13.96°	13.81°	65.88°	80.56°	69.36°	52.58°
Estimated	1.29°	14.65°	16.57°	49.96°	46.79°	6.90°	9.48°
Obs.	(24)	(8)	(31)	(29)	(24)	(24)	(24)
Actual	1.51°	11.55°	13.11°	26.95°	65.92°	64.43°	26.96°
Estimated	1.37°	7.81°	10.47°	23.26°	35.35°	4.49°	5.03°
Obs.	(10)	(31)	(8)	(24)	(29)	(27)	(27)
Actual	1.49°	9.09°	12.06°	16.09°	27.80°	16.98°	16.10°
Estimated	1.25°	9.26°	7.91°	32.49°	22.11°	8.36°	6.43°

Table 4.3.8

Comparisons of Actual and Estimated Angle Between the Original and Perturbed Eigenvectors from the Correlation Matrix for the Most Influential Observations

Angle	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7
Obs.	(30)	(30)	(5)	(5)	(30)	(31)	(30)
Actual	6.60°	70.25°	83.65°	82.54°	70.13°	26.86°	28.28°
Estimated	4.23°	8.94°	10.26°	4.92°	17.23°	21.97°	10.35°
Obs.	(19)	(29)	(29)	(30)	(27)	(30)	(31)
Actual	1.38°	5.31°	68.93°	69.64°	47.21°	26.06°	26.93°
Estimated	1.09°	5.12°	19.37°	11.44°	8.75°	9.09°	21.14°
Obs.	(27)	(19)	(30)	(29)	(17)	(12)	(24)
Actual	1.20°	4.69°	63.64°	69.01°	29.67°	20.22°	23.73°
Estimated	1.08°	3.91°	15.82°	18.63°	18.47°	26.82°	11.10°

Table 4.3.9

Angles Given by the Two Influence Functions for Observations in the Third and Fourth Dimensions from the Correlation Matrix Analysis of the Turtle Dataset

Sample Curve			Empirical Curve		
Obsn.	θ_3	θ_4	Obsn.	θ_3	θ_4
56	82°	82°	119	27°	27°
151	23°	23°	141	24°	24°
19	22°	22°	19	22°	22°
141	22°	22°	121	17°	17°
119	21°	21°	104	14°	14°
" "	" "	" "	" "	" "	" "
" "	" "	" "	" "	" "	" "
" "	" "	" "	" "	" "	" "
131	0.05°	0.05°	131	0.05°	0.05°
156	0.01°	0.01°	156	0.01°	0.01°

eigenvalues for these dimensions are $\hat{\lambda}_3 = 0.187$ and $\hat{\lambda}_4 = 0.175$. This shows that when we have two close eigenvalues the change in the corresponding eigenvectors when a point is omitted are nearly equal. This means that while the eigenvectors may be changing a lot in appearance they are just rotating within a relatively unchanged subspace. This occurs because the 'ellipse of variation' described by the two principal components with close eigenvalues is nearly circular. Rotation within this subspace will usually cause little change in the eigenvalues but large changes in the eigenvectors. Why large angles occur can be seen from the terms $(\lambda_j - \lambda_k)^{-1}$ in the influence functions for the eigenvectors from both the covariance and correlation matrices.

Figs. 4.3.1 and 4.3.2 provide an illustration of rotation. Fig 4.3.1 is a plot of the original principal component scores in the third and fourth dimensions for the problem above. Fig 4.3.2 is a plot of the principal component scores for the two perturbed dimensions when observation 19 is removed. On each plot we have labelled the same observations and also given are the angles between the two closest lines joining these points to the origin. Examination of these plots shows that the observations have stayed in the same relative positions to each other and hence the angles of 22° in Table 4.3.9 for observation 19 are mostly caused by rotation. A statistic is proposed in § 4.3.3 for measuring to what extent such eigenvectors have rotated out of the subspace.

We will now consider the differences in using the sample and empirical curves in Table 4.3.9. Observations 56 and 151 have the largest angles using the sample curve but do not appear in the top group on the empirical. The angles for observation 56 are close to 90° and this corresponds to the swop in the eigenvectors that was discussed in the previous section. The sample curve compares the original third eigenvector with the perturbed fourth eigenvector

Figure 4.3.1 Plot of the Original Principal Component Scores from the Third and Fourth Dimensions of the Covariance PCA of the Dataset on Turtles.

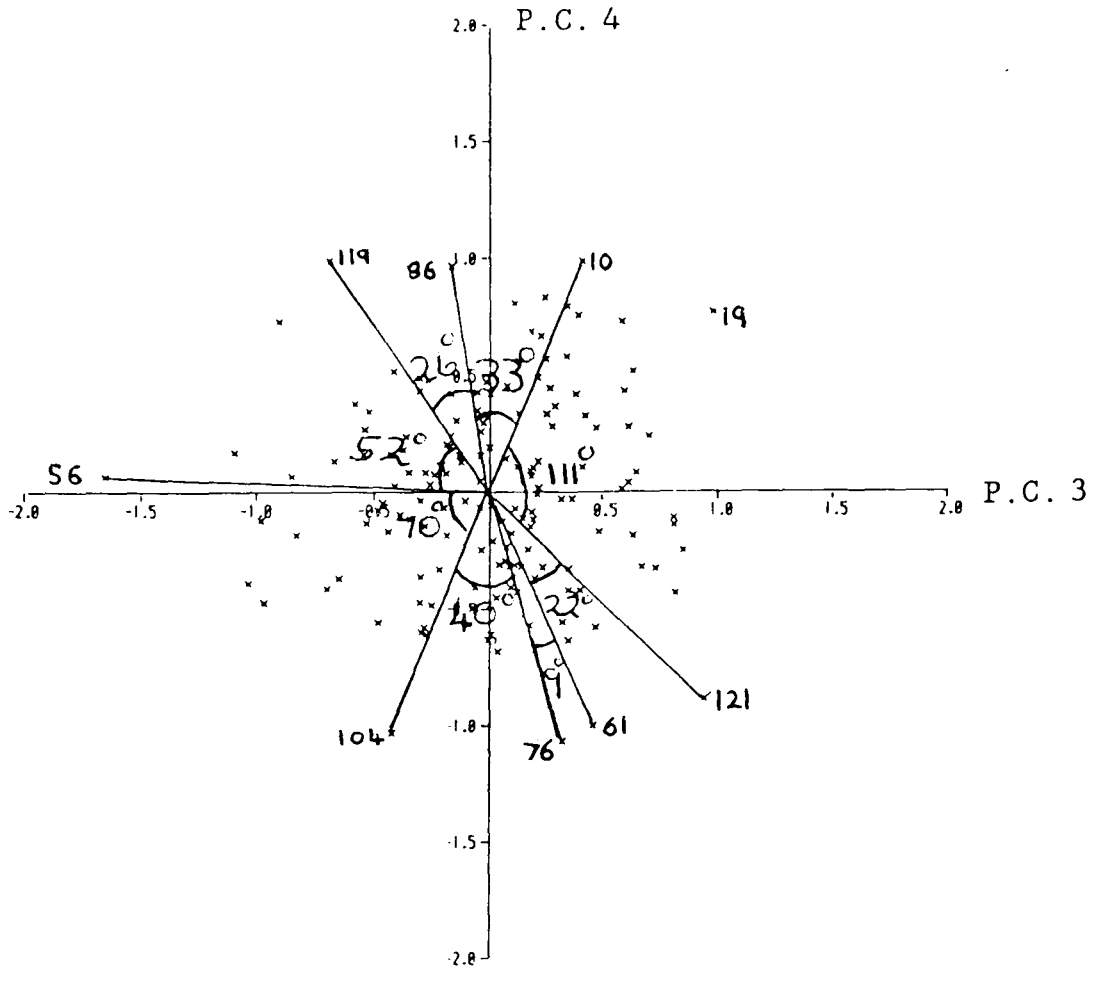
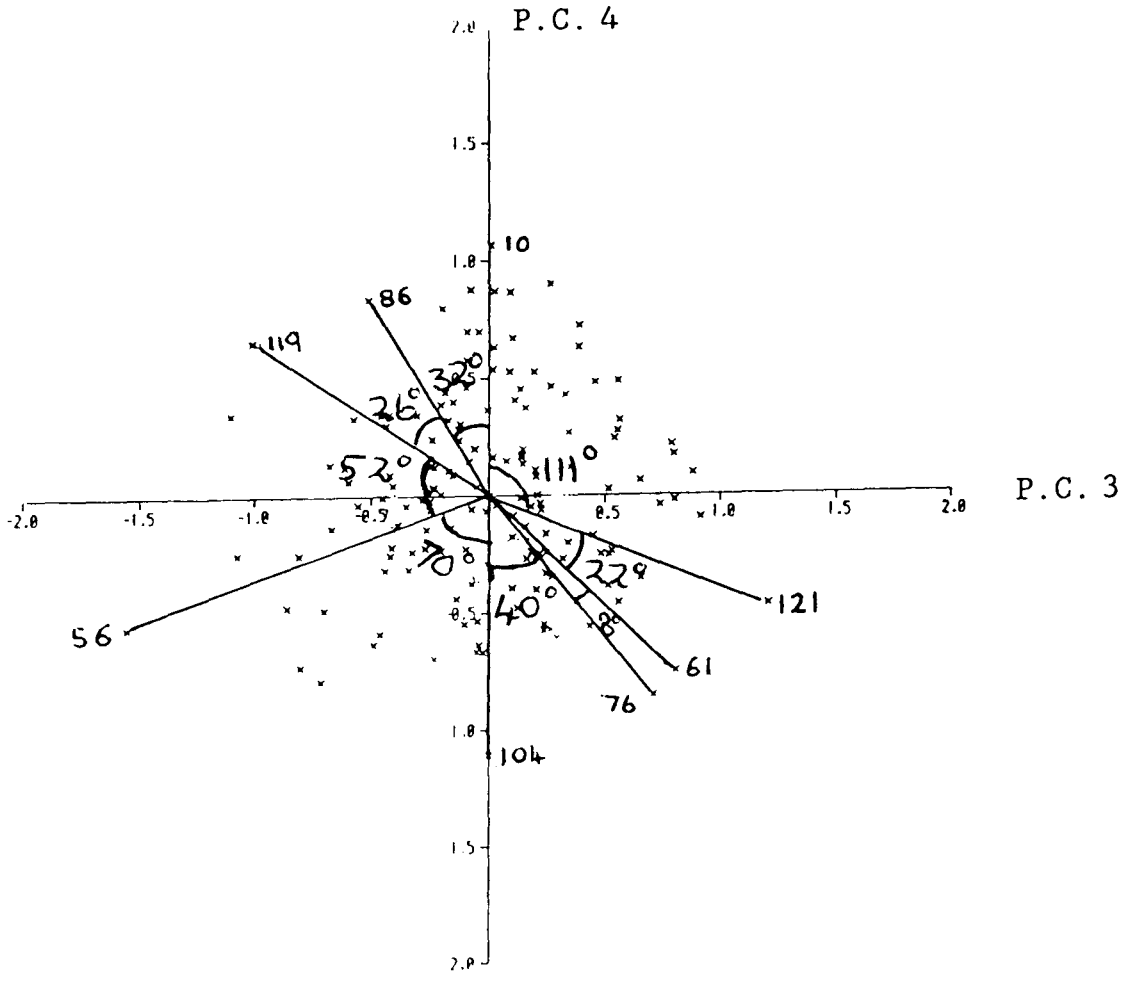


Figure 4.3.2 Plot of the Principal Component Scores in the Third and Fourth dimensions when Observation 19 is Omitted.



so we obtain large angles close to 90°. The empirical only looks at the change in the eigenvector defining that direction so we do not obtain large angles. The angles for observation 151 are not close to 90° and it does not appear in the top group of the empirical. This tends to occur when the perturbed eigenvalues are even closer than the original, here $\hat{\lambda}_3^* = 0.179$ and $\hat{\lambda}_4^* = 0.175$, and so the sample curve is affected by the greater indeterminacy of the two perturbed eigenvectors.

We will now consider the second dataset. In Table 4.3.7 the comparisons of the two curves are good for the first three dimensions but after that the two deviate. However, the deviations tend to be occurring for specific observations, namely observations 30 and 24. We will not discuss observation 30 until § 3.8 where we will look at the form of the original and perturbed eigenvectors in detail. Omitting observation 24 causes the fifth eigenvalue, which was originally close to the fourth, to decrease and become very close to the sixth eigenvalue instead. We find that the fifth and sixth eigenvectors rotate, see § 4.8, due to close perturbed eigenvalues. Again, the theoretical does not reflect the large changes caused by the perturbed rather than the original eigenvalues being close. In Table 4.3.8 the largest angles using the sample curve tend to occur for observations 5, 29 and 30. In the previous section we noted that observations 5 and 29 caused a swop in the third and fourth eigenvalues when they were omitted so this accounts for their large angles. The original second dimension, like the fourth from the covariance matrix analysis, is a contrast of variables 3 and 7. When observation 30 is omitted the variance along that direction falls, and in the sample case the perturbed fifth eigenvector is similar to the original second with leading coefficients in $\underline{x}_3$ and $\underline{x}_7$. However, the theoretical does not record a swop, although it gives a large change in the second eigenvalue. See § 4.8 for

further details.

In summary we find that the angles when using the sample or empirical curves are close in the first few dimensions where no swopping or rotation of the eigenvectors tends to happen. If one is only extracting the first few dimensions then one may not need to worry about such problems. However, in the correlation analysis for the Student dataset the eigenvalue from the second dimension decreases so much that by the 0.7 rule for retention of eigenvectors (Jolliffe, 1972) it would not be retained for inspection in the perturbed problem. We also find, see Section 4.8, that the second eigenvector changes its form so we can obtain problems in the early dimensions. The angles for the two curves tended to differ in two circumstances. The first is a swop in the eigenvectors. The empirical shows this has occurred by the changes in its eigenvalues but the sample reveals it in large angles, as it compares the original kth eigenvector with the perturbed $(k+1)$th eigenvector, when the kth and $(k+1)$th eigenvectors switch. Secondly, the two seem to differ when the perturbed eigenvalues are closer than the originals. Further comparisons of the two types of curves can be found in Jolliffe(1986, § 10.2) and Calder et. al. (1986).

4.3.3. A Measure of Influence for the Eigenvectors When There is Rotation

It is important to know if our eigenvectors are unsteady, due to close eigenvalues, since we usually interpret the sizes of their coefficients. However, it may be desirable to have some measure of influence that tells us to what extent the original subspace has changed, since it is possible that small angular changes may represent a larger change in the subspace than the big angular changes caused by rotation of the eigenvectors in a relatively unchanged subspace. The theoretical influence function can be used to give a measure of influence that shows to what extent an eigenvector has rotated out

of the original subspace.

We will use the theoretical influence function for the covariance matrix eigenvectors in the illustration, but exactly the same reasoning holds for eigenvectors from the correlation matrix. We will first consider the case of two close eigenvalues. The perturbed eigenvector from the covariance matrix, to $o(\epsilon)$, from (3.8.2) can be written as,

$$\tilde{\underline{\alpha}}_k = \underline{\alpha}_k - \epsilon Z_k \sum_{\substack{j=1 \\ j \neq k}}^{p} \underline{\alpha}_j (\lambda_j - \lambda_k)^{-1} Z_j \quad . \tag{4.3.1}$$

If λ_k and λ_{k+1} are close then from (4.3.1) we can see that we obtain large and nearly equal changes (and so angles) in the corresponding eigenvectors as the term in $(\lambda_k - \lambda_{k+1})^{-1}$ will dominate both their influence functions. We can write, to $o(\epsilon)$,

$$\tilde{\underline{\alpha}}_k = \underline{\alpha}_k + \epsilon(\lambda_k - \lambda_{k+1})^{-1} Z_k Z_{k+1} \underline{\alpha}_{k+1} + \epsilon \Psi_k$$
$$\tilde{\underline{\alpha}}_{k+1} = \underline{\alpha}_{k+1} - \epsilon(\lambda_k - \lambda_{k+1})^{-1} Z_k Z_{k+1} \underline{\alpha}_k + \epsilon \Psi_{k+1}$$

where,

$$\Psi_k = -Z_k \sum_{\substack{j=1 \\ j \neq k,k+1}}^{p} Z_j (\lambda_j - \lambda_k)^{-1} \underline{\alpha}_j$$

$$\Psi_{k+1} = -Z_{k+1} \sum_{\substack{j=1 \\ j \neq k,k+1}}^{p} Z_j (\lambda_j - \lambda_{k+1})^{-1} \underline{\alpha}_j \quad .$$

The terms $\epsilon\Psi_k$ and $\epsilon\Psi_{k+1}$ represent the extent to which the kth and $(k+1)$th eigenvectors respectively have rotated out of the original subspace generated by their close eigenvalues. The first two terms represent the rotation within the original subspace as $\underline{\alpha}_k$ and $\underline{\alpha}_{k+1}$ form a basis for the original subspace and a linear combination of these eigenvectors is still in the space. We can use the sums of squares of the elements in Ψ_k and Ψ_{k+1} as our scalar measures of influence for the two eigenvectors when we substitute in the sample equivalents and divide by $(n-1)^2$. We will refer to these measures as ER_k where,

$$ER_k = \frac{1}{(n-1)^2}\hat{\Psi}_k'\hat{\Psi}_k \quad . \tag{4.3.2}$$

This idea extends easily to higher dimensional subspaces, we merely need to omit terms from (4.3.1) when the eigenvalues differ by less than some preassigned value, for example 0.05. This can lead to one eigenvector being included in two different subspaces because of fairly close eigenvalues either side of it, but these eigenvalues on either side differ by more than the preassigned value. It is preferable for the value chosen to cover all eigenvalues in a set so we do not have overlapping subspaces. It is possible for eigenvectors to be steady when there are a number of close eigenvalues, as if there are number of large terms in $(\lambda_j - \lambda_k)^{-1}$ they can cancel each other out. This could result, say, in a stable last dimension that reflects a high correlation. Noting that the influence function for the correlation eigenvectors from (3.8.6) can be written as,

$$\tilde{\underline{\alpha}}_k = \underline{\alpha}_k - \epsilon \sum_{\substack{j=1 \\ j \neq k}}^{p} \underline{\alpha}_j (\lambda_j - \lambda_k)^{-1} L_{jk}$$

we see that exactly the same argument applies to the correlation eigenvectors. Another possible way of seeing whether the eigenvectors have rotated out of the subspace may be to subtract the two original angles, since rotation should lead to equal angular changes in the two eigenvectors. However, this assumes that it is only the eigenvector with the larger angular change that has rotated out, which may not be true. We will look at an example using measure (4.3.2) for the Turtle dataset on the third and fourth dimensions from the correlation matrix. Table (4.3.10) gives the values of ER_3 and ER_4 compared to the original sums of squares of the influences given by $\frac{1}{(n-1)^2}EIC(\underline{x}, \underline{\alpha}_3)'EIC(\underline{x}, \underline{\alpha}_3)$ and $\frac{1}{(n-1)^2}EIC(\underline{x}, \underline{\alpha}_4)'EIC(\underline{x}, \underline{\alpha}_4)$. We refer to these measures of influence in the table as SSE. This table shows how much

Table 4.3.10

Comparison of Influence Measures When We Have Rotation

Obsn.	Measure	$\underline{\alpha}_3$	$\underline{\alpha}_4$
119	SSE	0.1278	0.1286
	ER	0.0008	0.0008
141	SSE	0.1462	0.1463
	ER	0.0001	0.0002
19	SSE	0.1469	0.1471
	ER	0.0003	0.0002
121	SSE	0.0736	0.0734
	ER	0.0006	0.0003
104	SSE	0.0274	0.0257
	ER	0.0001	0.0009

the large angles in the third and fourth dimensions are caused by the one term in $(\hat{\lambda}_3 - \hat{\lambda}_4)^{-1}$ from the empirical influence functions. In this example all the influences have become small indicating there is probably little change in the eigenvectors other than from their rotation in the subspace. However, we can see that observation 104 had the smallest influences for SSE_3 and SSE_4 but has the largest influence in the Table for ER_4.

4.4. Using Second Order Terms

Since the estimated change is based on the asymptotic theoretical influence function, as n becomes smaller so our estimate becomes less accurate. To improve it one could use the second order terms from

$$\tilde{\lambda}_k = \lambda_k + \epsilon c_1 + \frac{\epsilon^2}{2} c_2 + \ldots\ldots \tag{4.4.1}$$

and similarly for the eigenvectors. The second order terms for the covariance eigenvalues and eigenvectors are

$$\Lambda = -2Z_k^2 \left\{ 1 + \sum_{\substack{j=1 \\ j \neq k}}^{p} Z_j^2 (\lambda_j - \lambda_k)^{-1} \right\} \tag{4.4.2}$$

$$\begin{aligned} \underline{\Pi} = &-Z_k^2 \sum_{\substack{j=1 \\ j \neq k}}^{p} Z_j^2 (\lambda_j - \lambda_k)^{-2} \underline{\alpha}_j - 2 \sum_{\substack{j=1 \\ j \neq k}}^{p} Z_j^2 (\lambda_j - \lambda_k)^{-1} TIC_V(\underline{x}, \underline{\alpha}_k) \\ &-2Z_k^3 \sum_{\substack{j=1 \\ j \neq k}}^{p} Z_j (\lambda_j - \lambda_k)^{-2} \underline{\alpha}_j \quad , \end{aligned} \tag{4.4.3}$$

see Critchley (1985). We could thus take as an estimate of the change in the k th eigenvalue when an observation is removed

$$E_V^*(\underline{x}, \lambda_k) = \frac{1}{n-1}(-\lambda_k + Z_k^2) + \frac{1}{(n-1)^2} \left(Z_k^2 + Z_k^2 \sum_{\substack{j=1 \\ j \neq k}}^{p} Z_j^2 (\lambda_j - \lambda_k)^{-1} \right) . \tag{4.4.4}$$

Note that the second order term is included with reversed sign. When we form the deleted sample curve we take $\epsilon = -1/(n-1)$ and subtract the

perturbed parameter from the original which maintains the sign of the empirical curve when used for addition or deletion of points. Substituting $\epsilon = -1/(n-1)$ into (4.4.1) and subtracting from the original eigenvalue we obtain as our estimate

$$E^* = \frac{1}{n-1}\hat{c}_1 - \frac{1}{2(n-1)^2}\hat{c}_2 \quad , \tag{4.4.5}$$

where $\hat{c}_1$ is (3.8.1) and $\hat{c}_2$ is (4.4.2) with the sample equivalents substituted in. If we were considering the addition of points the sign of the second order term would be maintained. It was found that (4.4.4) did not give better estimates of the sample change in the eigenvalues when we omit an observation than our previous estimate (4.2.2), which was the first order term with adjusted functions of n. However, some contour plots of the sample influence function (not given here) showed that while the contours tended to the straight lines of the theoretical contours quite quickly as n increased, for small n they were slightly bent indicating that some quadratic terms would be appropriate. We find that these second order terms do help to increase the accuracy of our estimated change in the eigenvalues when we again adjust our functions on n.

Note that the eigenvalues in (4.4.3) are the eigenvalues from the matrix $\Omega = (n/(n-1))S$, so we need to substitute in $(n/(n-1))\lambda_j$ to get the eigenvalues from S. This will affect some of the arguments in this Section. In particular, substituting $\Omega = (n/(n-1))S$ for Σ in (4.4.6) and $\bar{x}$ for μ and taking $\epsilon = 1/(n+1)$ will give expression (2.2.1) since (4.4.6) is exact to $o(\epsilon^2)$, as higher terms in ϵ are zero. This would lead to the appropriate functions of n for the addition of an observation. In deletion of an observation one can use expression (6) of Critchley (1985) to obtain more accurate results when using our second order terms, without considering adjustments to the functions of n. However, the conclusions from the original text below are similar to what would have been obtained had the above been used.

First we will reconsider our justification of the functions of n in (4.2.2). We have to $o(\epsilon)$,

$$\tilde{\Sigma} = (1-\epsilon)\Sigma + \epsilon(\underline{x}-\underline{\mu})(\underline{x}-\underline{\mu})'$$

and from (2.2.4)

$$S_{(-i)} = \left(1 + \frac{1}{n-2}\right)S - \frac{n}{(n-1)(n-2)}(\underline{x}_i - \underline{\bar{x}})(\underline{x}_i - \underline{\bar{x}})' \quad .$$

In the proof for $TIC_V(\underline{x}, \lambda_k)$, in § 3.6.1, we see that the $-\lambda_k$ term in the influence function comes from $-\Sigma$ in $TIC(\underline{x}, \Sigma)$, and the Z_k^2 term from the $(\underline{x}-\underline{\mu})(\underline{x}-\underline{\mu})'$. Since we would obtain the same empirical expression if we considered the asymptotic sample result from using $S_{(-i)}$ rather than Σ we

have taken

$$\tilde{\Sigma} = (1 - \epsilon_1)\Sigma + \epsilon_2(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})'$$

where $\epsilon_1 = -1/(n-2)$ and $\epsilon_2 = \frac{n}{(n-1)(n-2)}$. Our usual theoretical proof follows through in the same way as before since ϵ_1 and ϵ_2 are of the same order, and results in expression (4.2.2). This modification was found to improve our estimate. When we include the second order term we find

$$\tilde{\Sigma} = (1 - \epsilon)\Sigma + \epsilon(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})' - \epsilon^2(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})' \quad , \qquad (4.4.6)$$

see Critchley (1985). Taking $\epsilon = -1/(n-1)$ and substituting in the sample equivalents into (4.4.4) we obtain as an estimate of the perturbed sample covariance matrix

$$S^* = (1 + \frac{1}{n+1})S - \frac{n}{(n-1)^2}(\underline{x}_j - \underline{\bar{x}})(\underline{x}_j - \underline{\bar{x}})'$$

which is a less accurate estimate of the change in the sample covariance matrix than when we just used the first order term with modifications. We can write the perturbed sample covariance matrix

$$S_{(-i)} = \left(1 + \frac{1}{n-2}\right)S - \frac{1}{n-2}(\underline{x}_j - \underline{\bar{x}})(\underline{x}_j - \underline{\bar{x}})' - \frac{1}{(n-1)(n-2)}(\underline{x}_j - \underline{\bar{x}})(\underline{x}_j - \underline{\bar{x}})' \quad . \qquad (4.4.7)$$

So re-expressing (4.4.6) as

$$\tilde{\Sigma} = (1 - \epsilon_1)\Sigma + \epsilon_2(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})' + \epsilon_3(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})'$$

and following through the derivation of the second order terms with these different epsilon terms and substituting $\epsilon_1 = \epsilon_2 = 1/(n-2)$ and $\epsilon_3 = 1/(n-1)(n-2)$ which are taken from the functions of n in (4.4.7) we arrive at the estimate,

$$E^* = -\frac{\hat{\lambda}_k}{n-2} + \frac{n}{(n-1)(n-2)}\hat{Z}_k^2 + \frac{\hat{Z}_k^2}{(n-2)^2}\sum_{\substack{j=1 \\ j \neq k}}^{p} \hat{Z}_j^2(\hat{\lambda}_j - \hat{\lambda}_k)^{-1} \quad .$$

The justification for the above approach is the greater accuracy of the

estimate when compared against the actual sample changes in the eigenvalues.

No such modifications seemed necessary for the covariance eigenvectors. We thus take as our improved estimate the first and second order terms with sample equivalents substituted in. Again the sign of the second order term for deletion of observations from the dataset needs to be reversed. The use of the second order terms for the eigenvectors did improve the accuracy of the estimated angle. However, for the Student dataset, discussed in § 4.3.1 and § 4.3.2, we find that the theoretical seems to break down when using second order terms when observation 30 was omitted. The theoretical gives an argument out of the range of the cosine, this was the only example where this was seen to occur. We noted in § 4.3.1 that the empirical had problems when this observation was removed even when just the first order term was used. Tables 4.4.1 and 4.4.2 give comparisons for the Student dataset where $n = 33$, of the actual sample change and the estimated change in the eigenvalues and eigenvectors when the second order terms are used. These should be compared with Tables 4.3.3 and 4.3.7 respectively which are comparisons for the first order term. We see the second order terms have improved the estimate although not for observation 30 on the eigenvectors as discussed above. We still have the same disagreements due to swops and rotations though. In other datasets where we had less of these problems the comparisons were better.

The use of second order terms does not really seem to be worth the extra computational time, particularly for the eigenvectors where there are several parts to the second order term, as it is possible in some circumstances that the actual sample change may not take much longer to calculate. This would be particularly true for the correlation eigenvalues and eigenvectors whose first order terms were quite complicated. It is for this reason that the second order

Comparisons of Actual Sample Change and Estimated Change using Second Order Terms for the Student Dataset.

Table 4.4.1
Comparisons for the Covariance Eigenvalues

λ_k =	35.31	2.78	1.41	0.86	0.68	0.38	0.25
Obs.	(19)	(25)	(31)	(30)	(24)	(27)	(30)
Actual	6.40	0.42	0.25	0.14	0.27	0.06	0.16
Estimated	6.40	0.42	0.26	0.44	0.37	0.06	0.03
Obs.	(10)	(8)	(8)	(29)	(30)	(30)	(18)
Actual	2.60	0.40	0.06	0.05	0.27	0.06	0.09
Estimated	2.60	0.42	0.05	0.07	0.17	-0.00	0.09
Obs.	(26)	(30)	(24)	(2)	(17)	(5)	(19)
Actual	1.57	0.29	0.06	-0.03	0.07	0.05	0.05
Estimated	1.57	0.28	0.05	-0.03	0.07	0.05	0.05

Table 4.4.2
Comparisons for the Angles Between Perturbed and Original Eigenvectors

Angle	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7
Obs.	(19)	(30)	(24)	(30)	(30)	(30)	(30)
Actual	1.74°	13.96°	13.81°	65.88°	80.56°	69.36°	52.58°
Estimated	1.67°	15.37°	14.82°	-	-	16.54°	21.19°
Obs.	(24)	(8)	(31)	(29)	(24)	(24)	(24)
Actual	1.51°	11.55°	13.11°	26.95°	65.92°	64.43°	26.96°
Estimated	1.51°	10.99°	12.25°	33.67°	25.69°	10.19°	10.00°
Obs.	(10)	(31)	(8)	(24)	(29)	(27)	(27)
Actual	1.49°	9.09°	12.06°	16.09°	27.80°	16.98°	16.10°
Estimated	1.48°	9.45°	11.29°	11.07°	33.68°	12.54°	11.12°

terms for the correlation matrix have not been derived. In Chapter 7 we will discuss what type of curve may be preferable to use in certain circumstances which, takes into account the computation times.

4.5. Multiple Case Deletion

In § 3.9 we found from the theoretical influence function that the affect of deleting m observations was to sum the influences when each observation was removed individually. We will now give examples using the Turtle and Student datasets for the covariance and correlation matrix eigenvalues. We will consider the deletion of five observations for the Turtle dataset and two for the Student dataset which has a smaller sample size. We noted in § 3.9 that additivity of influence in samples is only likely to hold if m is small compared to the total sample size. The observations that we delete are those that were most influential on the largest eigenvalue when omitted individually. In Tables 4.5.1 to 4.5.4 we have written the individual sample changes in the eigenvalues when each point is omitted, these are then summed, and the sample change when all are omitted together is given underneath. The sums are actually calculated to greater accuracy than the two decimal points given in the tables for the individual changes. We can see the additive property is well supported in practice (however, we are not saying it is an exact relationship as it is only based on a first order approximation). In Table 4.5.4. we have examples, on the first two eigenvalues, showing the cancelling out of influence due to the different signs of the individual influences.

Comparisons of Summed Individual Sample Influences with Influence when Five Observations are Omitted Together from the Turtle Dataset

Table 4.5.1
Comparisons for the Covariance Eigenvalues

Obs	λ_1	λ_2	λ_3	λ_4	λ_5
143	6360.46	6.19	-0.31	-0.10	1.20
121	5714.93	3.23	0.30	-0.09	-0.08
61	2936.10	-0.89	0.45	-0.16	0.60
34	2926.98	1.18	-0.31	-0.18	-0.05
155	1687.17	-0.41	-0.31	-0.15	-0.05
Sum of Individual	19625.64	9.30	-0.18	-0.68	1.62
Block Deletion	20349.56	11.74	-0.13	-0.71	1.53

Table 4.5.2
Comparisons for the Correlation Eigenvalues

Obs	λ_1	λ_2	λ_3	λ_4	λ_5
151	-0.05	0.04	0.01	-0.00	-0.00
135	0.02	-0.01	-0.00	-0.00	-0.00
155	0.02	-0.01	-0.00	-0.00	-0.00
72	-0.02	0.02	-0.00	0.00	-0.00
101	0.02	-0.01	-0.00	-0.00	-0.00
Sum of Individual	-0.01	0.03	-0.01	-0.01	-0.01
Block Deletion	-0.01	0.03	-0.01	-0.01	-0.01

Comparisons of Summed Individual Sample Influences with Influence when Two Observations are Omitted Together from the Student Dataset

Table 4.5.3
Comparisons for the Covariance Eigenvalues

Obs	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
19	6.40	-0.08	-0.00	-0.03	-0.01	-0.01	0.05
10	2.60	-0.01	0.03	0.01	-0.02	-0.01	-0.01
Sum of Individual	9.00	-0.09	0.03	-0.02	-0.03	-0.02	0.04
Block Deletion	8.96	-0.10	0.05	-0.03	-0.03	-0.03	0.05

Table 4.5.4
Comparisons for the Correlation Eigenvalues

Obs	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
30	-0.58	0.46	-0.00	0.02	0.10	-0.00	0.01
10	0.12	-0.05	-0.02	-0.02	-0.02	-0.00	-0.00
Sum of Individual	-0.46	0.41	-0.02	0.00	0.08	-0.00	0.01
Block Deletion	-0.48	0.42	-0.03	0.01	0.08	-0.01	0.01

4.6. Simulated Critical Values for the Percentage Change in an Eigenvalue

To decide whether an observation is 'highly influential' we can rank the influences on a particular parameter and look at the gaps separating successive values. This is called the gap test, it is a very useful but subjective test for assessing the influence of observations. In this section we shall investigate the 5% simulated critical values for the largest percentage change in an eigenvalue for multivariate normal data. Both the sample and empirical influences will be considered and for both the covariance and correlation eigenvalues. Two types of simulation studies have been done. The first involved the generation of 1000 multivariate normal datasets from the same covariance (or correlation matrix) as an existing real dataset and for the same sample size. The largest absolute influence (which is the same as the largest non-absolute influences for the covariance matrix as the largest influences always occur for those observations which decrease the eigenvalue when they are omitted) from each simulated dataset is stored and the critical values formed from the resulting 1000 values. Some of these simulations are used in the final sections of this chapter. For the second simulation study, which we will concentrate on in this section, we generate a covariance matrix with a given set of eigenvalues and then simulate from this matrix 1,500 datasets for a given value of n. The critical values are then formed as above. This was also done for the correlation eigenvalues using the routine by Lin and Bendel (1985) to generate a correlation matrix from a given set of eigenvalues. Since we can combine any p eigenvalues together for the covariance matrix simulations it is easy to investigate the critical values for a variety of situations. It is more restrictive for the correlation eigenvalues since we need to satisfy $\sum_{k=1}^{p} \lambda_k = p$.

We will consider the simulated critical values for the covariance eigenvalues first. Tables 4.6.1 and 4.6.2 give the simulated critical values for the sample and empirical influences respectively for $n=150$ and $p=4,5$ and 6. The eigenvalues of each of the covariance matrices simulated from are specified by the columns that the critical values are entered in; and so the number of dimensions p is given by the number of entries in a given row. For examples involving the same p the same N(0,1)s were used to generate the datasets which will aid our comparisons of the critical values for different matrices. Two rows for $p=5$ in Tables 4.6.1 and 4.6.2 give the critical values, which are identical, for the same set of eigenvalues but from different covariance matrices. This shows the individual entries of the covariance matrices are not important in determining the critical values. The critical values for all the eigenvalues from a given covariance matrix are about equal, this means that we can expect the same percentage change in all the eigenvalues irrespective of their original sizes. Another example of this is given by the simulations from the same covariance matrix as that for the Turtle dataset. This matrix has a dominant eigenvalue $\hat{\lambda}_1 = 101845.62$ compared to $\hat{\lambda}_5 = 13.27$ and the sample 5% critical values for these two eigenvalues are 7.28% and 7.39% respectively. There is a tendency for the sample critical values on the smaller eigenvalues to be slightly larger. This is not so for the empirical (although generally the critical values are similar to the sample ones). This may occur due to the terms in the second order expression for the eigenvalues that involve $(\lambda_j - \lambda_k)^{-1}$, see (4.4.2), which will be large for the small eigenvalues which are usually closer together. Also from Tables 4.6.1 and 4.6.2 we see that the critical values change little over the different combinations of eigenvalue used to form the covariance matrices, and for the different values of p. There does seem to be a tendency for a

Table 4.6.1

Sample 5% Critical Values for Covariance Eigenvalues

λ =	1000	200	35	20	10	9	5	3	2
CRITICAL VALUES					7.72		7.78	7.53	7.96
				7.75	7.76			7.51	7.88
			7.74	7.80				7.52	7.87
		7.73	7.82					7.49	7.87
				7.66		7.54	7.84	7.83	8.19
				7.66		7.54	7.84	7.83	8.19
				7.66	7.59		7.83	7.83	8.19
			7.73		7.56		7.80	7.83	8.18
			7.65	7.67		7.84		7.75	8.16
			7.66	7.66	7.82			7.76	8.14
		7.82	7.50	7.93				7.73	8.08
	7.80	7.71	7.88					7.73	8.07
			7.71	7.70	7.64		7.91	7.67	8.31

Table 4.6.2

Empirical 5% Critical Values for Covariance Eigenvalues

λ =	1000	200	35	20	10	9	5	3	2
CRITICAL VALUES					7.80		7.83	7.55	7.62
				7.80	7.73			7.53	7.60
			7.81	7.79				7.53	7.63
		7.74	7.77					7.55	7.63
				7.74		7.58	7.79	7.80	7.72
				7.74		7.58	7.79	7.80	7.72
				7.71	7.61		7.78	7.79	7.71
			7.80		7.58		7.80	7.80	7.72
			7.77	7.59		7.73		7.82	7.76
			7.77	7.66	7.72			7.80	7.77
		7.83	7.60	7.76				7.77	7.78
	7.80	7.65	7.79					7.72	7.76
			7.87	7.73	7.63		7.78	7.46	7.87

Table 4.6.3
Sample 5% Critical Values for a Fixed Set of Eigenvalues and Varying n

$n \backslash \lambda$	1000	200	35	3	2
50	18.32	19.65	19.52	18.51	20.84
100	10.58	10.70	11.14	10.93	11.33
150	7.80	7.71	7.88	7.73	8.07
200	5.90	6.07	6.12	6.06	6.28

Table 4.6.4
Empirical 5% Critical Values for a Fixed Set of Eigenvalues and Varying n

$n \backslash \lambda$	1000	200	35	3	2
50	18.48	19.16	18.69	18.94	18.23
100	10.61	10.54	10.95	10.92	10.73
150	7.81	7.65	7.79	7.72	7.76
200	5.91	6.03	6.04	6.07	6.03

critical value to increase as the eigenvalue moves down in its ranked position, but there are exceptions. The steadiness of the critical values over the eigenvalues in a given dataset is very useful in assessing influences in the dataset. It also provides a very useful way of forming one influence statistic for a number of dimensions if one did not want to look at all the dimensions separately. Perhaps the best measure would be the maximum percentage change in any eigenvalue (or for those eigenvalues one is interested in) when we omit an observation. If one forms an average, say, one can overlook an observation that is highly influential on one eigenvalue only.

All the critical values in the above Tables have been based on the same value of n and we have seen that the critical values are steady over variations of other factors. In Tables 4.6.3 and 4.6.4 we have for the sample and empirical curves respectively the changes in the critical levels as we vary n for a covariance matrix with eigenvalues 1000,200,35,3 and 2. The critical values for another example examined, but not presented here, were very similar. When we go from $n=50$ to $n=100$ the critical values decrease by a little under $1/2$. We would expect this to be roughly so, as we form an estimate from the empirical by dividing by (n-1), although we did consider modifications for the covariance eigenvalues but these were still to $o(1/n)$. A rough guide to the critical values for the percentage change in the eigenvalues would seem to be $1000/(n-1)\%$, i.e. a proportional change of $10/(n-1)$ in any eigenvalue.

Tables 4.6.5 and 4.6.6 give the simulated 5% critical values for the percentage (absolute) change in the correlation eigenvalues for the sample and empirical respectively. For a given set of eigenvalues from the same correlation matrix the critical values are not steady over the different eigenvalues as they were for the covariance eigenvalues. The critical values increase as the eigenvalues decrease. We also find that the critical values for

Table 4.6.5

Sample 5% Critical Values for Correlation Eigenvalues

$\hat{\lambda}$ =	4.39	3.0	2.5	2.0	1.5	1.0	0.8	0.5	0.3	0.2	0.1	0.01
C V			2.04						10.73	11.89		
R A					3.81	4.85		9.79				
I L				4.34		6.40	8.49			10.46		
T U				4.32		6.31	8.61			10.55		
I E			2.96			5.69			10.42	11.04		
C S				5.42	5.97		8.07	10.60		10.75		
A			3.47		4.79			9.70	10.84	11.67		
L				4.61	6.35	7.17			10.68	11.27		
	1.50								10.84	11.57	12.52	12.92
		4.07				7.42		9.69	11.18	11.19		
			4.50			6.25	8.84	10.57		11.20		
			6.11		7.26	7.97		9.65	10.35	12.06		
	2.56					7.64			10.31	11.67	12.10	12.33

Table 4.6.6

Empirical 5% Critical Values for Correlation Eigenvalues

$\hat{\lambda}$ =	4.39	3.0	2.5	2.0	1.5	1.0	0.8	0.5	0.3	0.2	0.1	0.01
C V			1.83						9.89	10.65		
R A					3.73	4.41		8.91				
I L				4.06		6.07	7.52			9.84		
T U				4.07		5.85	7.67			9.79		
I E			2.76			5.12			9.55	10.14		
C S				5.29	5.68		7.38	9.73		9.90		
A			3.33		4.42			9.26	10.06	10.37		
L				4.57	5.89	6.61			10.15	10.25		
	1.33								10.19	10.73	11.51	11.61
		3.76				7.01		9.00	10.37	10.15		
			4.23			6.13	8.17	9.60		10.29		
			5.74		6.98	7.21		9.29	9.89	10.66		
	2.34					6.98			10.23	10.75	11.18	11.30

the absolute change, rather than the percentage change, increase as the eigenvalues increase. No divisor in $\hat{\lambda}_k^r, r = 0,\ldots 1$, was found to give critical levels that were constant over the different eigenvalues. This also follows from Table 4.6.5 as the critical values for a given eigenvalue vary, even when p is fixed. See for example, $\hat{\lambda} = 1.5$ for $p = 5$. The critical values for the smallest eigenvalues tend to increase when other eigenvalues are close to them, but the largest eigenvalues tend to get larger critical values when the other eigenvalues are further away.

As p increases the critical value for a given eigenvalue increases, see for example, the column for $\lambda = 2.5$. As p decreases the percentage of variance, λ/p, accounted for by that eigenvalue increases. This means that a larger proportion of the data may be confirming this direction, so omitting one point may have less affect. Comparing the critical values for $p = 3$ and $p = 6$ we see that as the eigenvalues decrease the steadier the critical values are over p. There is a three-fold increase in the critical value from $p = 3$ and $p = 6$ for $\hat{\lambda} = 2.5$, a doubling for $\hat{\lambda} = 1.5$, but the critical values for the smallest remain steady or possibly even decrease as p increases.

The sample and empirical critical values follow the same pattern but the empirical values are smaller.

4.7. Influence in a Dataset of Rock-Chip Samples

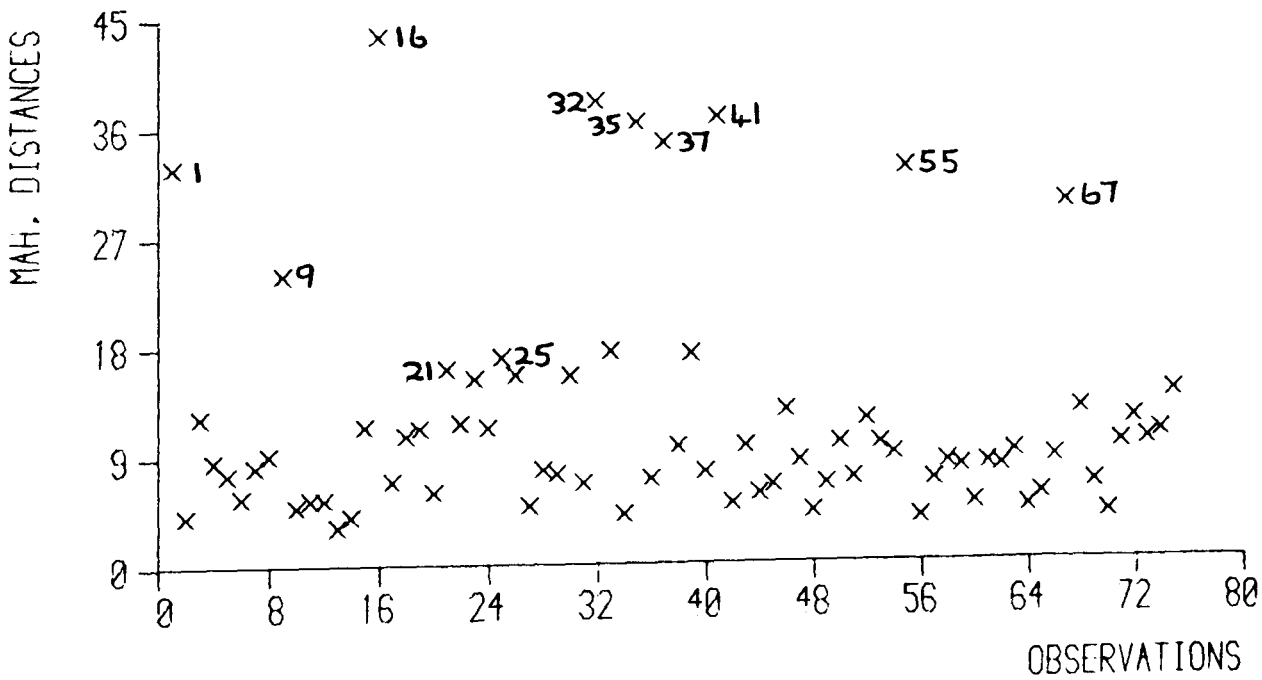
In this example we shall specifically be examining the differences and similarities of outliers and influential observations. We shall also consider whether the detection of influential observations on the individual correlation coefficients has much bearing on what is influential in the principal component analysis of this correlation matrix. This dataset consists of 12 variables which are recordings of the trace elements present in 75 rock chip samples. The data have been standardised, so we will only be considering the principal components from the correlation matrix. The data is published in Hawkins and Fatti (1984) and Table 4.7.1 gives the observations in the dataset found to be outliers by Hawkins and Fatti, using three types of measures. The first measure was just the original standardised variables, the second was the principal component scores computed from the D matrix, which has the eigenvectors standardised by their variances, and the third measure consisted of the scores computed from a varimax rotation of this D matrix. Hawkins and Fatti's criterion for an observation to be an outlier was that at least one of the twelve variables/components had an absolute value greater than 3.34, which corresponded to an overall Bonferroni significance level of 1% for the twelve variables/scores. This testing was done by assuming normality. The Mahalanobis distance, which was used in § 2.3.3, has a special form when we have multivariate normal data. It can be written in terms of the difference in the log-likelihoods for the full dataset and for the dataset with the ith observation omitted, when the true form of the covariance matrix is known, and as the ratio of the two log-likelihoods if the covariance matrix is not known. Fig. 4.7.1 is a plot of the Mahalanobis distances against observation number, and we see that all the observations in Table 4.7.1 have large Mahalanobis distances, except perhaps for observation 25 which was

Table 4.7.1

Observations Found to be Outliers by Hawkins and Fatti and the Techniques that Found them as Outliers

Obs	Techniques		
1		2	3
16	1	2	3
25			3
32		2	3
35		2	3
37		2	3
41	1		3
55		2	3
67		2	3

Figure 4.7.1 Plot of the Mahalanobis Distances for the Dataset on Rock Chip Samples.



only highlighted by the varimax procedure. Critical values for the Mahalanobis distance only exist up to $p=5$ so we will examine such plots informly when discussing our datasets rather than performing any formal test.

We will first analyse this dataset as if we were carrying out a principal component analysis with the aim to reduce the dimensionality of the problem. Using the 0.7 rule of Jolliffe (1972) we would retain the first three dimensions which account for 83% of the total variance. One question of interest is to what extent are the outliers found by Hawkins and Fatti, which are found to be outlying on the minor principal components, influential on our principal component analysis when we are only extracting the first three dimensions. We discussed in § 3.8.2 that as λ_k decreases its influence, as revealed by the theoretical expressions, becomes dominated by its own score z_k, but since the eigenvalues sum to p it is possible that these outliers with large (minor) scores will be influential on the early components. However, as λ_k becomes smaller so do the component scores, z_{ki}, so although the change may be large compared to λ_k it may not be large compared to the earlier eigenvalues.

Table 4.7.2 gives the three most influential observations on the first three and last two eigenvalues and eigenvectors. To the left of the observation number we have its ranked position on the principal scores in that direction. On the right we have the sample influence which is recorded as the percentage change in the eigenvalues and the angle between the original and perturbed eigenvector. Using an informal gap test one could conclude observation 32 was influential but it only leads to a perturbed second eigenvalue of 1.36 which seems reasonably unimportant (however, we noted in our simulation results of § 4.6 that we would expect smaller percentage changes in the early components). Observation 21 causes $\hat{\lambda}_3$ to change by more than 10% and the perturbed eigenvalue is, $\hat{\lambda}_3 = 0.703$, so it has nearly fallen below the 0.7 rule

Table 4.7.2

$\lambda_1 = 7.86$			$\lambda_2 = 1.29$			$\lambda_3 = 0.8$		
ranked PC score	Obsn.	Sam. Infl.	ranked PC score	Obsn.	Sam. Infl.	ranked PC score	Obsn.	Sam. Infl.
2	24	1.5%	20	32	-5.6%	1	21	12%
53	55	-1.2%	2	52	3.2%	2	22	9.3%
71	1	1.1%	1	33	3.2%	3	23	7.7%

$\underline{\alpha}_1$			$\underline{\alpha}_2$			$\underline{\alpha}_3$		
ranked PC score	Obsn.	Sam. Infl.	ranked PC score	Obsn.	Sam. Infl.	ranked PC score	Obsn.	Sam. Infl.
1	32	1.2°	22	26	5.1°	3	23	8.9°
4	46	0.9°	29	23	4.3°	5	1	8.9°
53	55	0.9°	1	35	4.3°	1	21	8.9°

$\lambda_{11} = 0.06$			$\lambda_{12} = 0.03$		
ranked PC score	Obsn.	Sam. Infl.	ranked PC score	Obsn.	Sam. Infl.
1	16	39.4%	2	9	11.8%
2	37	24.7%	1	43	11.6%
3	35	20.5%	6	1	9.7%

$\underline{\alpha}_{11}$			$\underline{\alpha}_{12}$		
ranked PC score	Obsn.	Sam. Infl.	ranked PC score	Obsn.	Sam. Infl.
2	37	31.3°	6	1	7.65°
18	67	28.5°	40	35	5.23°
3	35	24.8°	26	67	4.93°

used above for the retention of components. Comparing the change in $\hat{\lambda}_3$ when observation 21 is omitted against simulated critical values, obtained by simulating normal data from the same correlation matrix and the same sample size as this dataset, we find observation 21 is just influential at the 5% level. Thus, the only influential observation on all of the first three eigenvalues is not included in the list of outliers found by Hawkins and Fatti. However, we do find that the score for observation 21 on the third axis for the second outlier technique is 3.19 (compared to the level 3.34 which it was tested against) which was higher than any of the scores on this technique for observations 25 and 41 which were found to be outliers with the varimax approach. However, observation 41, and less so observation 25, do come close to the top ranked scores on several dimensions rather than just one like observation 21. The most influential observations on the second and third eigenvectors are not among the list of outliers in Table 4.7.1. Observation (outlier) 32 is at the top of ranked influences on $\hat{\underline{\alpha}}_1$ but an angle of 1° will not make the eigenvector change in appearance. A question of interest is just what does a 5° or 9° change look like in an eigenvector, as given by observation 26 on $\hat{\underline{\alpha}}_2$ and 23 on $\hat{\underline{\alpha}}_3$ respectively. These are given in Table 4.7.3 and we see that neither of these angles leads us to change our interpretation of the eigenvector. We do usually find that eigenvectors corresponding to large well separated eigenvalues, for example the first eigenvector from highly correlated data, are stable as most of the data tend to be supporting its direction. However, this is not true for the second eigenvector from the correlation principal component analysis of the Student dataset that we will discuss in § 4.8 and first mentioned in § 4.3.2. Observations 26 and 23 would not be classed as 'highly influential' when using the gap test, and since our interpretation of the relevant eigenvectors

Table 4.7.3

5% change		9% change	
$\underline{\alpha}_2$	$\underline{\alpha}_2$ (26 deleted)	$\underline{\alpha}_3$	$\underline{\alpha}_3$ (23 deleted)
-0.13	-0.16	-0.40	-0.43
-0.17	-0.18	-0.14	-0.20
0.28	0.29	0.05	0.06
0.37	0.39	0.24	0.22
0.05	0.04	-0.25	-0.19
0.02	0.04	0.38	0.31
-0.47	-0.42	0.48	0.53
-0.21	-0.22	-0.06	-0.05
-0.18	-0.13	0.38	0.41
-0.14	-0.15	-0.22	-0.19
-0.00	-0.03	-0.32	-0.31
0.65	0.66	0.12	0.06

Table 4.7.4

Obsn. deleted	16	37	35	67
$\tilde{\kappa}_{11}$	0.035	0.044	0.046	0.055
$\tilde{\underline{\alpha}}_{11}$	0.71	0.50	0.52	0.60
	-0.42	-0.20	-0.54	-0.57
	-0.00	-0.00	-0.03	-0.08
	0.12	0.07	0.11	0.24
	0.01	0.09	0.00	0.06
	0.05	-0.13	-0.08	0.11
	0.09	0.11	-0.02	0.01
	0.14	0.15	0.52	0.23
	0.01	0.14	0.02	-0.09
	-0.51	-0.78	-0.31	-0.17
	-0.11	0.14	-0.22	-0.33
	-0.07	-0.01	-0.04	-0.20

has not changed we will ignore their affects.

Although they were not found to be 'highly influential' (by the use of the gap test or simulated critical values) out of the nine outliers found by Hawkins and Fatti four appear in the top three rankings for the first three eigenvalues and eigenvectors. This is despite the low component scores most of these have in these dimensions. This may be due to their large affects in the other dimensions since we need to maintain the sum $\sum_{k=1}^{p} \hat{\lambda}_k = p$ and the orthogonality of the eigenvectors.

Using the simulated critical values, as discussed earlier, we find the following observations are influential on the latter eigenvalues,

λ_6	...	...	λ_9	λ_{10}	λ_{11}
55*			35*	67*	16*
32*			32*	1*	37
					35

(this is after we have taken into account the swops in λ_6 and λ_7 when observations 55 and 32 are omitted, individually). The * means that the observation was also found to be an outlying along this direction by Hawkins and Fatti when using the component scores calculated from the D matrix. Observations 37 and 35 were not found to be outlying using the 11th eigenvector from the D matrix but they did have large scores for it. This supports the theoretical result that as λ_k decreases its influence is dominated by its own score. The only two observations that were classed as outliers that have not come out as influential on some eigenvalue are 25 and 41 which were detected by the varimax procedure only. Observation 41 does come out in the top three ranked influences on the fourth, seventh, eighth and ninth eigenvectors and this may be due to the number of large (but not largest) component scores it had. This was seen to be important in the covariance

eigenvectors from the theoretical influence function (3.8.2) where observations with just one large score would not be so influential on the eigenvectors. We could not actually show this for the correlation eigenvectors but we did note that one of the terms in the theoretical influence function for the eigenvectors was the same as the function for the covariance eigenvectors. Observations 25 and 41 are the second and third most influential observations on $\hat{\lambda}_8$ with changes of 9% and 7.5% respectively. However, the most influential on $\hat{\lambda}_8$ is observation 39 with 12.7% and this was not included in Hawkins and Fatti's list of outliers.

We saw in § 4.4.2 that when the eigenvalues are closer together, which often occurs when they are small, we can get very large changes in the eigenvectors, so we need to be careful about our interpretations. However, when eigenvalues are small we may wish to examine the eigenvectors since they define near-constant relationships among the variables. However, if we have many close eigenvalues we may not obtain large angles as the large contributions to the influence functions can cancel each other out. This could occur in this dataset since the smallest three eigenvalues are 0.08, 0.06 and 0.03. From Table 4.7.2 we can see the angular changes for $\hat{\underline{\alpha}}_{12}$ are fairly small. Since they define nearly constant relationships, variables with large coefficients in the latter eigenvectors will often have large correlations with each other. We would thus expect that observations that undermine or enhance these correlations should be influential on the relevant eigenvalues and eigenvectors. This appears to be supported by the theoretical influence function for the correlation eigenvalues which is

$$TIC_R(\underline{x}, \lambda_k) = \sum_{\substack{j=1 \\ j \neq t}}^{p} \alpha_{kj} TIC(\underline{x}, \rho_{jt}) \alpha_{kt} \quad , \tag{4.7.1}$$

since if α_{kj} and α_{kt} are large then the changes in ρ_{jt} will be given a lot of

weight in the influence function. The influence function for the correlation eigenvectors can also be written in terms of the influence function for the bivariate correlation, see(3.8.8), but the coefficients multiplying it are from different eigenvectors and there are other complicated terms in the influence function. Hence, it is not immediate that observations affecting the large correlations will also be the most influential on the eigenvectors defining the near constant relationships. Below, where we consider the sample changes, we will see that it can be difficult to understand influence in principal component analysis through influence on the correlations in R, although there are obviously strong links between the two.

The last two eigenvectors are given below and the most influential observations on the eigenvectors and corresponding eigenvalues are given in Table 4.7.2.

$\hat{\underline{\alpha}}_{11}$	$\hat{\underline{\alpha}}_{12}$
0.65	-0.01
-0.49	0.03
0.07	0.70
0.11	-0.67
0.03	-0.01
-0.01	0.18
0.06	0.00
0.22	-0.02
0.06	-0.03
-0.50	-0.09
0.10	-0.14
0.07	0.05

The last eigenvector has large coefficients in x_3 and x_4 and the correlation $r_{34} = 0.94$. No observation could have undermined this correlation by much and from the low influences it appears the high correlation was not due to one discrepant value. We will confine our attention to the 11th dimension which is more interesting.

Using simulated critical values, or the gap test, since the fourth largest influence is 4.6%, we would find all three observations on $\hat{\lambda}_{11}$ as influential.

Since the fourth largest angle of change is 13.5 ° we may also pick out the top three influences on the eigenvector as well. We can see from Table 4.7.2 that observations 37 and 35 come out on both the eigenvalue and eigenvector but not observations 16 and 67. Eigenvector $\hat{\underline{\alpha}}_{11}$ is a contrast of variable x_1 against x_2 and x_{10}. When observation 16 is omitted all correlations involving x_1 increase and when observation 37 is omitted all correlations involving x_{10} increase. Both lead to a decrease in λ_{11} which corresponds to the closer relationships among the variables. This supports the earlier comment, made from looking at the theoretical influence function, that observations influential on the last few eigenvalues will tend to be influential on the correlations involving the variables with high coefficients in the corresponding eigenvector. However, only observation 37 is influential on the corresponding eigenvector. Table 4.7.4 gives the perturbed eleventh eigenvector when certain observations are omitted. When observation 37 is omitted the coefficient of x_{10} has increased at the expense of x_1 and particularly x_2. Observation 16 causes the coefficient of x_1 to increase by only a small amount and the other coefficients hardly change. The difference of the two observations on the eigenvector may be attributed to observation 37 changing the structure of the eigenvector, which becomes more of a contrast between x_1 and x_{10} when it is omitted, whereas omitting observation 16 leads to a closer relationship between the variables without changing the form of the original relationship.

When observation 35 is omitted all the correlations involving x_8 increase and from Table 4.7.4 we see that the 8th coefficient more than doubles. Observation 35 is thus highly influential on λ_{11} without affecting correlations involving x_1, x_2 and x_{10} (except with x_8). Expression (4.7.1) also shows that this can happen since if the change in a set of correlations is large enough

they can affect the eigenvalues irrespective of the sizes of the coefficients. It may not be clear what eigenvalue they may affect though. Finally, observation 67 does not affect $\hat{\lambda}_{11}$ when it is omitted but it does have a large influence on the eigenvector. When it is omitted all the correlations decrease except those involving x_{11} and x_{12} and from Table 4.7.4 we see the coefficients of these variables increase at the expense of x_{10}. Why it is at the expense of x_{10} and not x_1 or x_2 is not clear and this shows why influence is not easy to follow through when looking at the correlation coefficients. The theoretical influence function for the eigenvectors shows the relationship between the influences on the correlation and those on the eigenvectors is complicated. Even the simpler expression for the eigenvalues shows that there can be cancelling out affects.

In summary, we have used this dataset to illustrate several points. First, we saw that if we were interested in just the first few dimensions of the principal component analysis we would not need to worry about the numerous outliers found in this dataset. The only observation that was influential in these dimensions was not one of these outliers. We then saw that all the outliers found by Hawkins and Fatti were influential somewhere in the analysis, except perhaps for observation 25. These outliers tended to come out on the dimensions they were discrepant on, which coincided with the result from the theoretical influence function that as the eigenvalues become smaller the influence of an observation depends more on its score in that dimension. Finally, we considered investigating influence indirectly through looking at influence on the bivariate correlations in R. Although there are links with influence on the bivariate correlations, there is no better way to find influential observations in principal component analysis other than carrying out the influence procedure on the eigenvalues and eigenvectors directly.

4.8. Influence in the Dataset of Anatomical Measurements on Students at the University of Kent.

This dataset was introduced in § 2.5.3 and in § 4.3 we saw that there were problems with swops in the eigenvalues/eigenvectors when certain observations were omitted. This is an interesting dataset so we will consider influence in detail for this dataset in this section. Both the covariance and correlation principal component analyses will be considered. We shall also consider influence on the principal component scores. Fig. 4.8.1 is a plot of the Mahalanobis distances, against observation number, for these data. This plot suggests that a few observations could be outliers and from the plots of the data we can see the reasons why they are outlying. A plot of variables 3 and 7 is given by Fig. 2.5.1 and we see, as discussed in § 2.5.3, that observation 30 is clearly outlying on these two variables. We also noted in § 2.5.3 that it is possible that the two measurements had been written down in the wrong order since when they are swopped the observation is no longer outlying. Fig. 4.8.2 shows that observation 24 is undermining the correlation, r_{56}, and less so observation 19. However, we usually find that observation 19 has a similar correlation structure to the rest of the data but corresponds to a 'large' person. Observation 18 has a large Mahalanobis distance but is not shown to be unusual on any of the pairwise plots of the variables.

The three largest influences on each eigenvalue and eigenvector for the covariance and correlation matrices are given in Tables 4.3.3, 4.3.4,4.3.7 and 4.3.8. First, we will consider the changes in our analysis if only the first two dimensions were of interest. For the covariance and correlation matrices these correspond to 91% and 82% of the total variance respectively. Out of all the observations with large Mahalanobis distances only observation 19, with an 18.1 percentage change on λ_1, has much affect on the first two principal

Figure 4.8.1 Plot of the Mahalanobis Distances for the Dataset on Anatomical Measurements of Students.

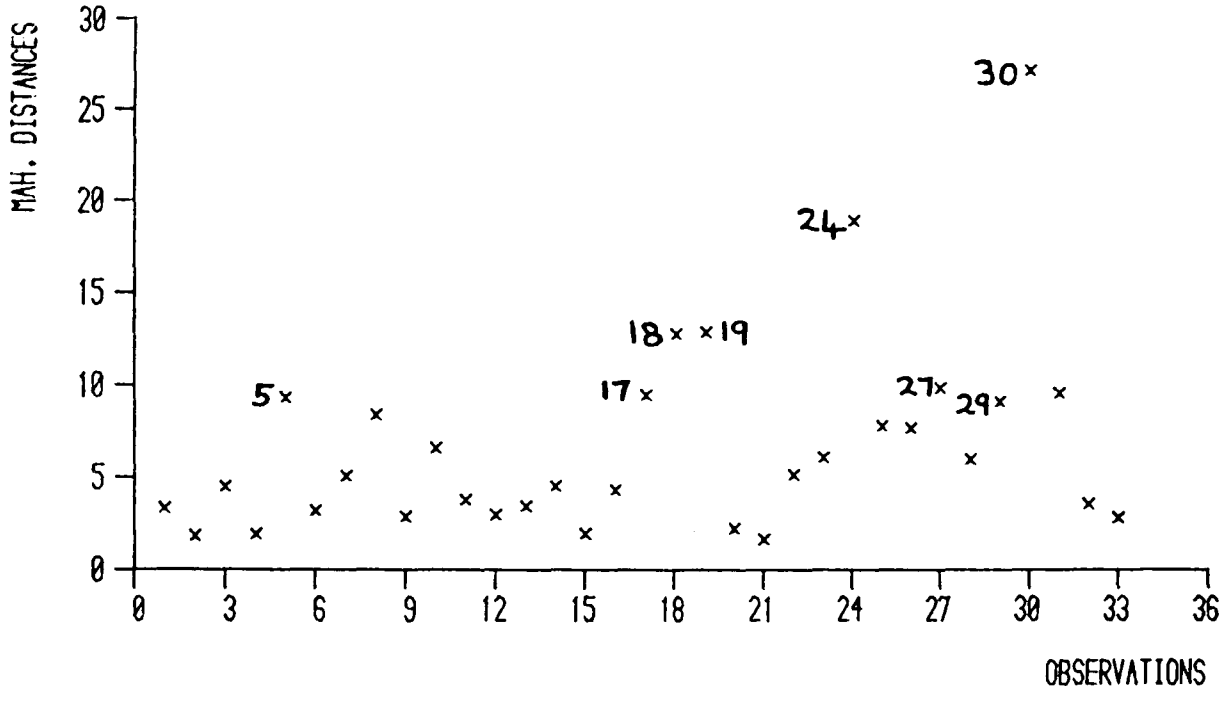
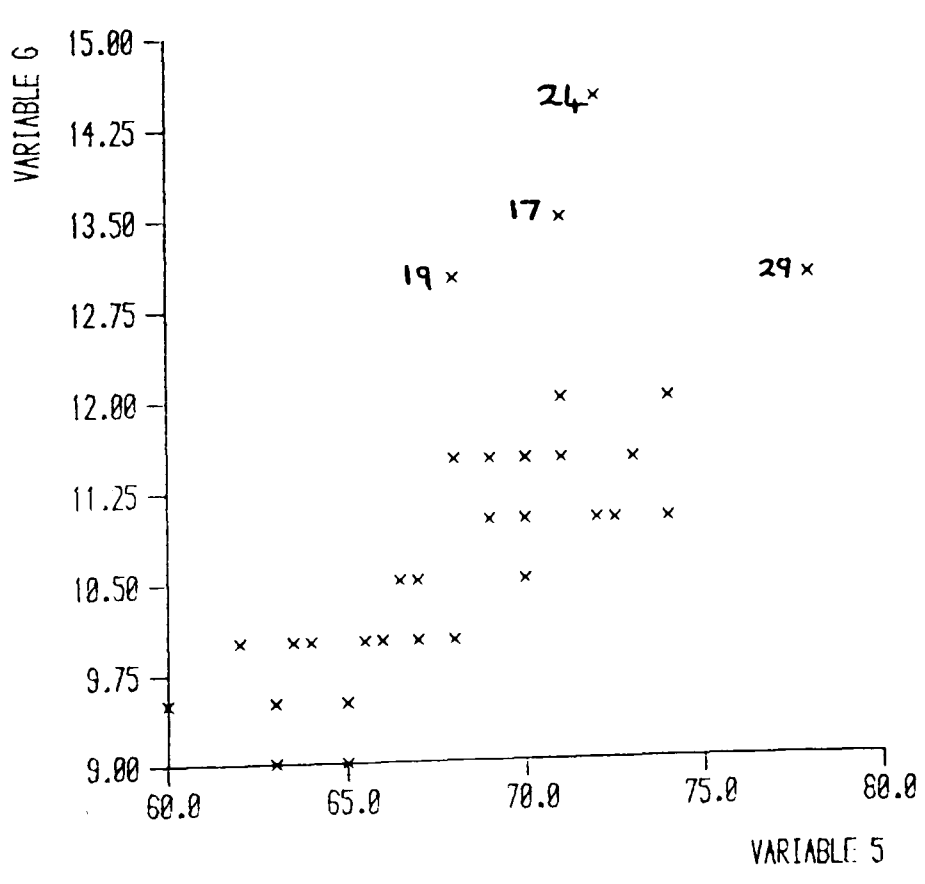


Figure 4.8.2 Plot of Variables x_5 and x_6 for the Dataset on Anatomical Measurements on Students.



components from the covariance matrix. When compared against simulated critical values (see § 4.6 for details) this was not significant at the 5% level. However, using the gap test one would probably conclude observation 19 as 'influential' as the next largest percentage change is 7.4%. Observation 30 leads to a 14° change in $\hat{\underline{\alpha}}_2$ when it is omitted but, as we will see later, our interpretation of $\hat{\underline{\alpha}}_2$ does not change much so we shall not regard it as influential (of course, an influence which is significant need not lead to a change in our interpretations etc, particulary as n increases). So observation 19 is the only one that has much affect on the first two components from the covariance matrix. When it is omitted the first eigenvalue falls from 35.31 to 28.91 which suggests that observation is inflating the variances rather than one countering the multivariate structure of the data.

The situation is quite different if we consider the first two eigenvalues and eigenvectors from the correlation matrix. Although observation 19 is in the top groups of influence on these eigenvalues and eigenvectors it has little affect. However, observation 30 has a substantial affect on the first two components. Eigenvalue $\hat{\lambda}_2$ falls to 0.47 from 0.93 and $\hat{\lambda}_1$ increases to 5.38 from 4.80. Using the 0.7 rule, discussed in the previous section, for the retention of components we would no longer consider the second dimension. In the original data we find that the second dimension has been determined almost entirely by observation 30 which has a score of $z_{2,30} = 4.88$ compared to the next largest absolute score of 1.14. The original second eigenvector is a contrast of variables 3 and 7 which are the variables we found observation 30 to be unusual on. When observation 30 is omitted we find the angle of 70.25 in Table 4.3.8 is not so much caused by rotation or swopping of the eigenvectors as a change in the structures (this will be seen below).

The above comments show how different our influence procedure can be

on the covariance and correlation matrices. We would not need to worry about the presence of certain outliers, especially observation 30, when looking at the first two components from the covariance matrix. However, outliers can be highly influential on the first few dimensions, as shown by observation 30 on the correlation matrix analysis. We will now consider the other dimensions and in particular the affect of observation 30 across the whole analyses.

The original and perturbed eigenvalues and eigenvectors when observation 30 is omitted for the covariance matrix are given in Tables 4.8.1 and 4.8.2 respectively. The 14° change in $\hat{\underline{\alpha}}_2$, mentioned above, has not changed our original interpretation of this eigenvector. The coefficients of variables x_3 and x_7 have decreased but they were not large originally. The original eigenvector $\hat{\underline{\alpha}}_4$ is a contrast of variables x_3 and x_7, $\hat{\underline{\alpha}}_5$ is dominated by x_6 and $\hat{\underline{\alpha}}_6$ by x_4. When observation 30 is omitted $\hat{\underline{\alpha}}_4^*$ and $\hat{\underline{\alpha}}_5^*$ are similar to $\hat{\underline{\alpha}}_5$ and $\hat{\underline{\alpha}}_6$ respectively. Eigenvectors $\hat{\underline{\alpha}}_6^*$ and $\hat{\underline{\alpha}}_7^*$ are dominated by variables $\underline{x}_3$ and x_7 respectively. Hence the original fourth eigenvector is no longer present in the analysis. Only by looking at the coefficients for all the eigenvectors is this obvious, so after finding influential observations, using scalar measures of influence, one should consider in detail the affects of these observations by looking at the individual coefficients of the eigenvectors. Since the original fourth dimension does not exist in the perturbed problem this raises difficulties in that it should not be compared with any of perturbed eigenvectors. The sample curve will just compare it with the new fourth eigenvector, which is why we obtain the large angles. The theoretical thus appears to be the best curve to use in such circumstances since it only considers changes along that given direction. Unfortunately, there are difficulties with this curve as well. We noted, in § 4.4, that when we included the second order terms into the estimated angle we did not get a valid range

Table 4.8.1

Original Eigenvalues and Eigenvectors from the Covariance Matrix for the Student Dataset

$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_4$	$\hat{\lambda}_5$	$\hat{\lambda}_6$	$\hat{\lambda}_7$
35.31	2.78	1.41	0.86	0.68	0.38	0.25
$\hat{\underline{\alpha}}_1$	$\hat{\underline{\alpha}}_2$	$\hat{\underline{\alpha}}_3$	$\hat{\underline{\alpha}}_4$	$\hat{\underline{\alpha}}_5$	$\hat{\underline{\alpha}}_6$	$\hat{\underline{\alpha}}_7$
0.42	0.48	0.74	0.16	-0.06	-0.06	0.09
0.57	0.46	-0.60	0.08	0.26	-0.02	0.16
0.15	-0.31	-0.08	0.67	0.21	0.02	-0.62
0.12	-0.01	-0.00	-0.00	-0.19	0.97	-0.01
0.65	-0.66	0.14	-0.33	0.07	-0.08	0.13
0.17	-0.07	-0.18	0.32	-0.88	-0.20	0.14
0.10	0.17	-0.17	-0.56	-0.26	-0.07	-0.74

Table 4.8.2

Eigenvalues and Eigenvectors from the Covariance Matrix for the Student Dataset when Observation 30 is Omitted

$\hat{\lambda}_1^*$	$\hat{\lambda}_2^*$	$\hat{\lambda}_3^*$	$\hat{\lambda}_4^*$	$\hat{\lambda}_5^*$	$\hat{\lambda}_6^*$	$\hat{\lambda}_7^*$
36.44	2.49	1.45	0.72	0.41	0.32	0.09
$\hat{\underline{\alpha}}_1^*$	$\hat{\underline{\alpha}}_2^*$	$\hat{\underline{\alpha}}_3^*$	$\hat{\underline{\alpha}}_4^*$	$\hat{\underline{\alpha}}_5^*$	$\hat{\underline{\alpha}}_6^*$	$\hat{\underline{\alpha}}_7^*$
0.42	0.53	0.72	-0.11	-0.02	0.12	-0.02
0.57	0.46	-0.67	0.23	-0.02	-0.10	0.11
0.15	-0.16	-0.12	-0.11	0.38	0.86	0.20
0.12	-0.04	0.00	-0.20	0.86	-0.38	-0.23
0.65	-0.69	0.18	0.17	-0.14	-0.14	0.03
0.17	-0.06	-0.19	-0.92	-0.25	-0.09	0.08
0.09	-0.01	-0.13	-0.02	-0.16	0.25	-0.94

for the cosine when observation 30 was omitted. In fact we find the theoretical does not behave correctly either when just using the first order term, for this example. In Table 4.2.7 we see that the theoretical gives angles 49.96° and 46.79° for the fourth and fifth eigenvectors respectively. Although these are valid angles the resulting eigenvectors no longer have elements whose sums of squares are 1, and their cross multiplication with some of the other eigenvectors are no longer zero. The empirical influence function leaves the first three and last two eigenvectors virtually unchanged when observation 30 is omitted. The perturbed fourth and fifth eigenvectors given when using the empirical, become dominated by $x_3 + x_6$ and $x_3 vs\ x_6 + x_7$ respectively. The coefficient for x_6 in the perturbed fourth eigenvector by the empirical has a coefficient similar to that in the perturbed sample eigenvector at 0.97. However, the coefficient for x_3 also remains large, close to the original at 0.65, which explains why we get a sum of squares larger than 1. Thus, the perturbed eigenvectors in the fourth and fifth dimensions, when using the empirical, have become combinations of the original coefficients that dominated these directions. It is interesting to look at how the deleted empirical curve, discussed in Chapter 1, deals with the emergence of a dimension (i.e. the fourth) when it is not in the original set of eigenvectors, which are now based on the dataset without observation 30. We find that this curve fairs no better than the empirical influence function above. The first four dimensions using the deleted empirical curve remain virtually unchanged. This means the fourth dimension remains dominated by x_6 rather than becoming a contrast of variables x_3 and x_7, as in the sample case, for the dataset including observation 30. In the remaining three perturbed eigenvectors given by the deleted empirical curve, when observation 30 is added, there is one coefficient over 1 for either x_3, x_4 or x_7, and often a large

coefficient under 1 as well. These variables are those that were large originally (in the deleted dataset) for these dimensions. The fifth dimension remains dominated by x_4, but its coefficient is above 1. The contrast of x_3 against x_7 does appear in the perturbed eigenvectors from the deleted empirical curve but for the seventh eigenvector (rather than the fourth as for the sample perturbed eigenvector based on the full model). However, the coefficient in x_3 is above 1 in this eigenvector. The perturbed sixth dimension has large coefficients in all of x_3, x_4 and x_7.

Tables 4.8.3 and 4.8.4 give the original and perturbed eigenvalues and eigenvectors from the correlation matrix when observation 30 is omitted. Like $\hat{\underline{\alpha}}_4$ from the covariance matrix , $\hat{\underline{\alpha}}_2$ from the correlation matrix is a contrast of variables x_3 and x_7. The second dimension seems to disappear when observation 30 is omitted, but unlike the covariance matrix we do not have problems with the theoretical giving invalid eigenvectors. This may be due the perturbed 5th eigenvector which is reasonably similar to the original second dimension. However, the theoretical does not imply even a switch in the eigenvectors when observation 30 is omitted. Looking at the changes in the eigenvalues it gives that the original $\hat{\lambda}_2$ falls down to 0.54, without a swop. This would isolate observation 30 as highly influential but it does not reflect the sample changes well. So we again see that when an observation is extremely discrepant, the empirical will indicate this but not in a way comparable to the sample influence function. In Tables 4.8.3 and 4.8.4 the middle eigenvectors seem to change a lot, rather than just moving up in their ranked position as in the covariance case. However, all the same variables seem to involved in the original third, fourth and fifth eigenvectors as in the perturbed second, third and fourth eigenvectors. Hence, the changes could be due to rotation, particularly as their eigenvalues are all close. The angles

Table 4.8.3

Original Eigenvalues and Eigenvectors from the Correlation Matrix for the Student Dataset

$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_4$	$\hat{\lambda}_5$	$\hat{\lambda}_6$	$\hat{\lambda}_7$
4.80	0.93	0.40	0.37	0.28	0.13	0.10
$\hat{\underline{\alpha}}_1$	$\hat{\underline{\alpha}}_2$	$\hat{\underline{\alpha}}_3$	$\hat{\underline{\alpha}}_4$	$\hat{\underline{\alpha}}_5$	$\hat{\underline{\alpha}}_6$	$\hat{\underline{\alpha}}_7$
-0.40	0.12	-0.28	0.44	-0.56	-0.04	0.49
-0.43	0.11	-0.01	0.27	-0.07	0.52	-0.67
-0.33	-0.63	0.15	0.19	0.47	0.30	0.35
-0.37	0.00	-0.65	-0.64	0.13	0.08	0.04
-0.43	-0.12	-0.01	0.18	0.20	-0.79	-0.32
-0.38	-0.12	0.63	-0.51	-0.44	-0.03	0.03
-0.29	0.74	0.28	-0.00	0.46	0.04	0.29

Table 4.8.4

Eigenvalues and Eigenvectors from the Correlation Matrix for the Student Dataset when Observation 30 is omitted

$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_4$	$\hat{\lambda}_5$	$\hat{\lambda}_6$	$\hat{\lambda}_7$
5.38	0.47	0.40	0.35	0.18	0.13	0.09
$\hat{\underline{\alpha}}_1^*$	$\hat{\underline{\alpha}}_2^*$	$\hat{\underline{\alpha}}_3^*$	$\hat{\underline{\alpha}}_4^*$	$\hat{\underline{\alpha}}_5^*$	$\hat{\underline{\alpha}}_6^*$	$\hat{\underline{\alpha}}_7^*$
-0.37	0.14	0.65	0.24	-0.38	0.24	0.40
-0.40	-0.09	0.32	0.07	0.26	0.21	-0.78
-0.38	-0.09	-0.36	-0.60	-0.42	0.42	-0.02
-0.35	0.79	-0.20	-0.07	0.43	0.03	0.16
-0.41	0.02	0.05	-0.18	-0.27	-0.85	-0.11
-0.36	-0.14	-0.55	0.72	-0.18	0.04	-0.01
-0.38	-0.57	0.03	-0.13	0.57	-0.04	0.43

between $\hat{\underline{\alpha}}_3$ and $\hat{\underline{\alpha}}_2^*$, $\hat{\underline{\alpha}}_4$ and $\hat{\underline{\alpha}}_3^*$, $\hat{\underline{\alpha}}_5$ and $\hat{\underline{\alpha}}_4^*$, are 35.58°,43.82° and 32° respectively, indicating there could be some rotation. The last two eigenvectors mostly have the same large coefficients as they did originally, although some of the sizes of the coefficients have altered.

We will now consider some of the other observations in this dataset that had large Mahalanobis distances. Observation 24 is highly influential on $\hat{\lambda}_5$ from the covariance matrix and the angles of change in $\hat{\underline{\alpha}}_5$ and $\hat{\underline{\alpha}}_6$ are almost equal. We noted in the § 4.3.2 that the perturbed fifth and sixth eigenvalues were closer than the originals and the sample and empirical differed as this can cause rotation in the perturbed eigenvectors which the empirical does not take account of when the original eigenvalues were not too close. The original fifth and sixth eigenvectors were dominated by variables x_6 and x_4 respectively, and when observation 24 is omitted $\hat{\underline{\alpha}}_5^*$ is a contrast of x_6 and x_4 and $\hat{\underline{\alpha}}_6^*$ is the sum of the two variables. This shows that the two eigenvectors have rotated as the perturbed, rather than the original eigenvalues, were close. Observation 24 is also highly influential on $\hat{\lambda}_5$ from the correlation matrix. We saw in § 4.3.1 that observations 5 and 29 caused a swop in the third and fourth eigenvectors when they were omitted. Observation 24 actually had a larger influence than observations 5 and 29 on the third eigenvalue, even after the swops were taken into account. Observation 24 lead to an increase of 17.7% in the third eigenvalue whereas observations 29 and 5 lead to decreases in $\hat{\lambda}_3$ of 15.96% and 8.64% respectively. This shows a swop need not correspond to the largest changes in an eigenvalue but it can just depend on the closeness of the two eigenvalues. Observation 18, which has a similar Mahalanobis distance to observation 19, is influential on the smallest eigenvalue from both the covariance and correlation matrices; it also has the largest scores in these two dimensions. Most of the correlation

increase when observation 18 is omitted but without any particular correlation changing drastically. Both of the smallest eigenvalues decrease, when observation 18 is deleted, possibly to reflect the stronger relationships indicated by the slightly higher correlations. However, observation 18 is not influential on either of the corresponding eigenvectors. This may occur as no one correlation was affected more than the others, so the basic structure remained the same.

In § 3.6.3 we derived the theoretical influence function for the principal component scores from the covariance matrix. We will use this dataset to examine which observations may be influential on the other principal component scores when it is deleted. Fig. 4.8.3 is an example of the movement of the first two scores, for nine of the 33 observations, when each observation is omitted from the dataset in turn (note, an observation cannot affect its own score when it is omitted). The scores do not seem to have moved much, with little change in their ranked positions along the two axes. The observations with the most positive and negative ranked influences on the first five scores (numbered by observation number) in the first dimension are given in Table 4.8.5. The ranked non-absolute influences for all scores are similar and these correspond to moving from left to right of the observations plotted along the first axis in Fig 4.8.3. Hence, observation 19 has the largest positive influence on all the scores in the first dimension and observation 10 the largest negative influence. This means that the positioning of an observation in the first dimension is almost determining its influence on all the other scores in that dimension. The changes in the 32 scores given by observation 19 range from 0.40 to 0.54, and those for observation 10 from -0.28 to -0.41. This means the affect of an observation is almost the same for all the scores. The reason for this can be seen from the theoretical influence

Figure 4.8.3 Plot of the Original Covariance Principal Component Scores from the First Two Dimensions and the Perturbed Scores for Nine Observations when the 33 Observations are Omitted Individually.

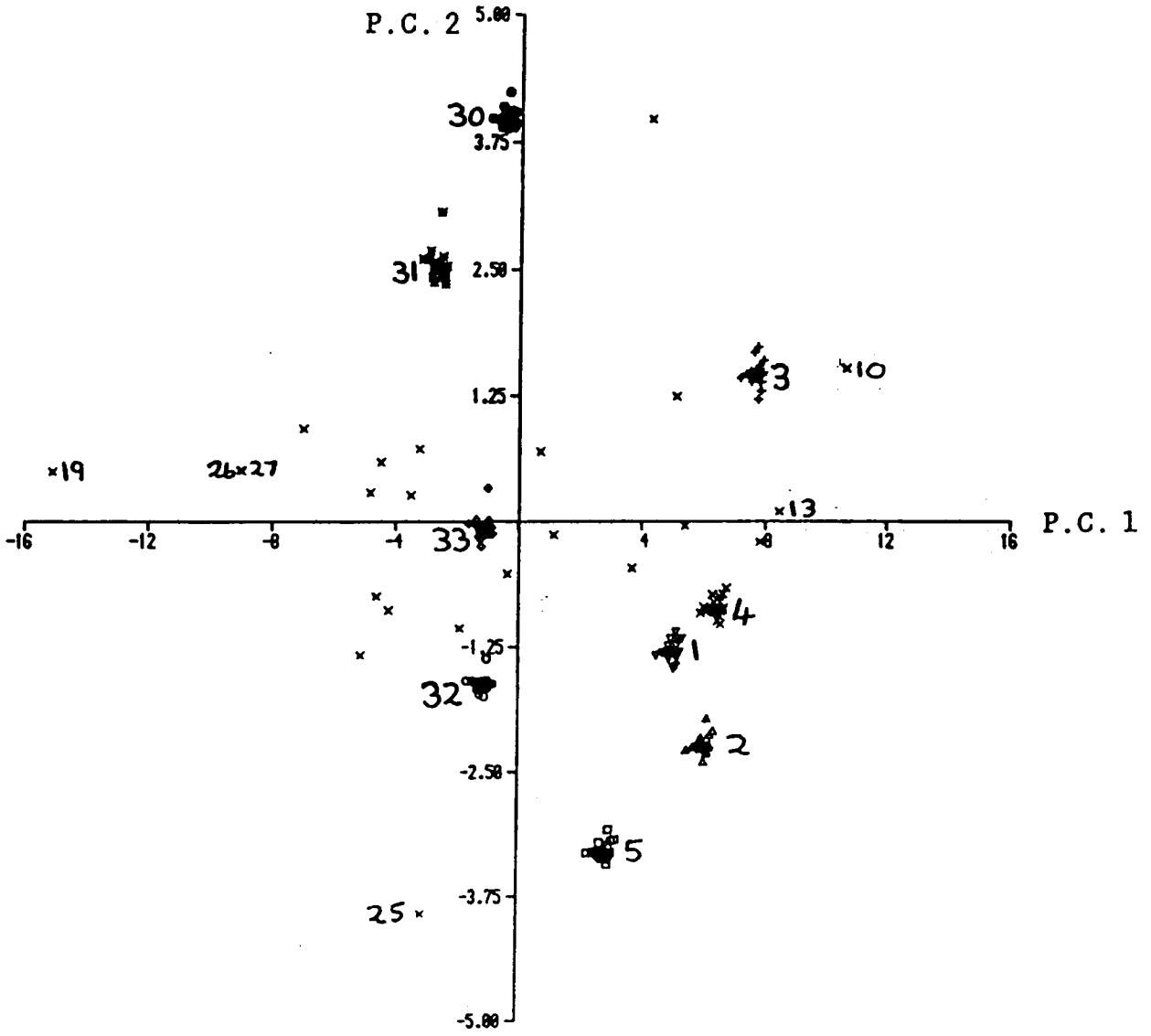


Figure 4.8.4 Plot of the Original Correlation Principal Component Scores for the First Two Dimensions.

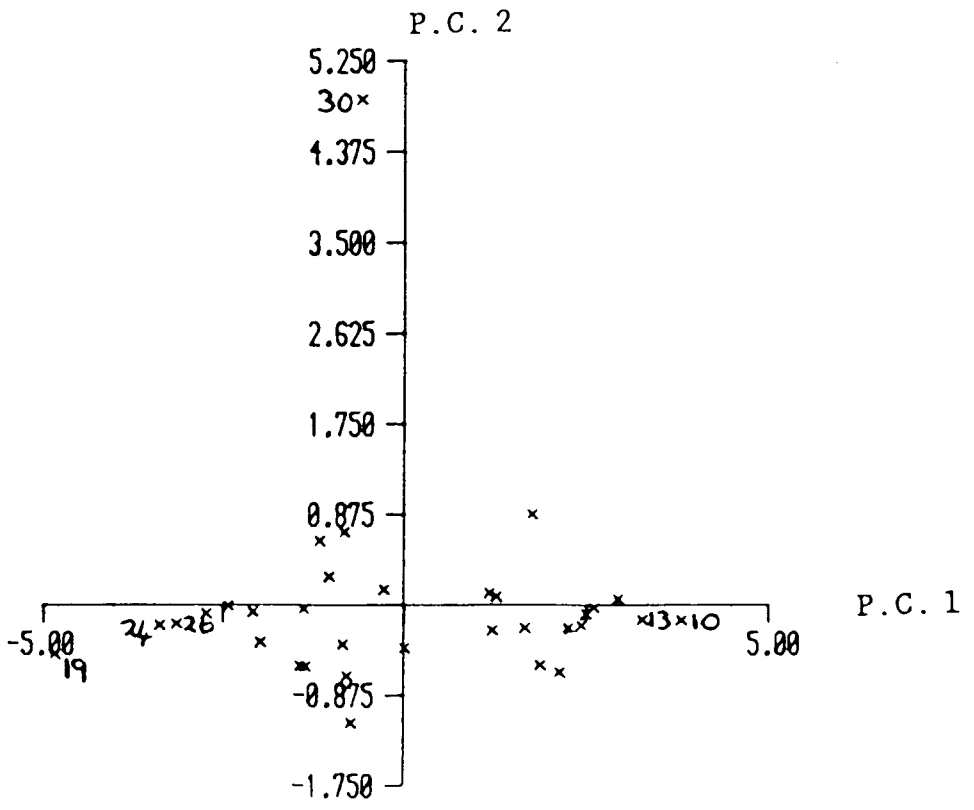


Table 4.8.5

Three Most Positive Influential Observations on the First Five Scores from the First Component of the Covariance Matrix

1	2	3	4	5
19	19	19	19	19
27	27	27	27	26
26	26	26	26	27

Three Most Negative Influential Observations on the First Five Scores from the First Component of the Covariance Matrix

1	2	3	4	5
10	10	10	10	10
13	13	13	13	3
3	3	11	3	13

function given by (3.8.3), which is

$$TIC_V(\underline{x}, Z_{kc}) = -Z_k \left[1 + \sum_{\substack{j=1 \\ j \neq k}}^{p} \underline{\alpha}_j (\lambda_j - \lambda_k)^{-1} Z_j Z_{jc} \right] \quad ,$$

where Z_k is the score for the point omitted and Z_{kc} is the score whose change we are interested in. When λ_k is large and distinct from the other eigenvalues, as $\hat{\lambda}_1$ is in this dataset, the second term in the above influence expression will be small and the term in $-Z_k$ will dominate. From the derivation of this influence function in § 3.6.3 the term $-Z_k$ comes from the change in the mean when the observation is deleted. We thus obtain similar changes in all the scores due to dominant term $-Z_k$, and this represents the mean of the principal axis moving. We usually use the theoretical results by substituting in the sample equivalents and dividing by $(n-1)$. This gives,

$$\frac{-Z_{1,19}}{32} = 0.46 \text{ and } \frac{-Z_{1,10}}{32} = -0.32$$

which are within the sample ranges of influence given above.

The pattern, observed in the first dimension, gradually disappears as $\hat{\lambda}_k$ becomes smaller and less distinct from the other eigenvalues. We can see why this happens from the theoretical influence function as the summation term will start to dominate. Although the observations with large principal component scores in these later dimensions tend to be those that are 'highly influential', there seems to be little pattern to which scores they are influential on, and their influences on some scores will be positive and negative on others.

Fig. 4.8.4 is a plot of the first two principal component scores from the correlation matrix. Observation 19, as for the covariance matrix analysis, has the largest Z_1 score but we do not find it is always the most (positive) influential observation on the other scores as above. Those observations that come out as most influential (with positive influence) are a combination of

16,22,30,26,19,15,17 and 24, all of which are to the left of the plot, except for observation 30. However, the largest influences over the scores do occur for observation 19 and this is on the scores that lie close to it in the first dimension, for example, on the scores for observations 24,26,22 and 27. Again, we can look at the theoretical influence function to explain why this is so. This is,

$$TIC_R(\underline{x}, Z_{kc}) = -\sum_{\substack{j=1 \\ j\neq k}}^{p} \underline{\alpha}_j(\lambda_j - \lambda_k)^{-1} Z_{jc} \sum_{s=1}^{p} \sum_{\substack{t=1 \\ t\neq s}}^{p} \alpha_{js}\alpha_{ks} TIC(\underline{x}, \rho_{st})$$
$$+\frac{1}{2}Z_{kc} - \frac{1}{2}\sum_{t=1}^{p}\alpha_{kt}Y_{ct}Y_t^2 - Z_k \quad .$$

The first term is small when λ_1 is large and distinct, as in this dataset, and (taking sign into account) when Z_{kc} and Z_k are large and close, terms (2) and (4) must cancel each other out to a certain extent and we find the third term is large and the same sign as Z_k. Conversely, when Z_{kc} and Z_k are large but opposite signs terms (2) and (4) will combine together but the third term is large and a different sign to Z_k. For example, the four terms when observation 19 is omitted for the influence function on $Z_{1,24}$ are,

(1) +0.64 (2) −1.68 (3) 7.03 (4) 4.84

and for the influence function on $Z_{1,10}$

(1) −0.14 (2) 1.90 (3) −8.04 (4) 4.84

For the other dimensions the first term will become more important, like the similar term in the influence function for the covariance scores, and no clear picture of influence emerges.

4.9. Influence on a Dataset for the Protein Consumption in Europe and Russia (Application to Covariance Biplot).

This dataset has been published in two books (see Greenacre(1984) and Gabriel (1981)) and it is also given in Table 4.9.1, since we shall be discussing the individual observations. This dataset is also discussed in Chapter 6. In this section we will consider influence on the covariance principal component analysis only, and then apply this to the covariance biplot. Fig 4.9.1 is a plot of the Mahalanobis distances for the 25 observations. The largest Mahalanobis distances are for observations 1 and 17, but the values are not particularly distinct from the others. Greenacre notes that observation 17 (Portugal) has a large score on the second principal component due to its generally low consumption of protein (a feature of size) but has a large score on the second correspondence analysis due to its high consumption of the variable fish (a feature of shape). Although this is true, we find that omitting observation 17 has a large affect on the covariance biplot due to this unusually high consumption of fish. Greenacre (1984, § 9.6) also considers the influence of observations in PCA for this dataset and uses the upper bounds for angles of rotation discussed in § 3.8.4. We shall consider these bounds in greater detail here. The most influential observations on the first two eigenvalues and eigenvectors are given in Table 4.9.2. The order of influence on the first two eigenvalues, even though n is as small as 25, is the same as that for the ranked principal component scores as suggested by our theoretical expression for the covariance eigenvalues. The angles for the observations on the first two eigenvectors were all less than the upper bounds given for the angles of rotation discussed in § 3.8.4 and outlined in Greenacre(1984, § 8.1). The only exception was observation 8 on the second eigenvector. This observation had the largest affect on the first

Table 4.9.1

Dataset on the Consumption of Protein in Europe and Russia

Country	MEAT	PIPL	EGGS	MILK	FISH	CERS	STAR	NUTS	FRVG	Total
ALBA	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7	71.2
AUST	8.9	14.0	4.3	19.9	2.1	28.0	3.6	1.3	4.3	86.4
BELX	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4.0	87.3
BULG	7.8	6.0	1.6	8.3	1.2	56.7	1.1	3.7	4.2	90.6
CZEC	9.7	11.4	2.8	12.5	2.0	34.3	5.0	1.1	4.0	82.8
DENM	10.6	10.8	3.7	25.0	9.9	21.9	4.8	0.7	2.4	89.8
EGER	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6	75.7
FINL	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1.0	1.4	90.4
FRAN	18.0	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5	98.2
GREE	10.2	3.0	2.8	17.6	5.9	41.7	2.2	7.8	6.5	97.7
HUNG	5.3	12.4	2.9	9.7	0.3	40.1	4.0	5.4	4.2	84.3
IREL	13.9	10.0	4.7	25.8	2.2	24.0	6.2	1.6	2.9	91.3
ITAL	9.0	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7	84.0
NETH	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7	84.7
NORW	9.4	4.7	2.7	23.3	9.7	23.0	4.6	1.6	2.7	81.7
POLA	6.9	10.2	2.7	19.3	3.0	36.1	5.9	2.0	6.6	92.7
PORT	6.2	3.7	1.1	4.9	14.2	27.0	5.9	4.7	7.9	75.6
RUMA	6.2	6.3	1.5	11.1	1.0	49.6	3.1	5.3	2.8	86.9
SPAI	7.1	3.4	3.1	8.6	7.0	29.2	5.7	5.9	7.2	77.2
SWED	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2.0	80.0
SWIT	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9	88.1
UK	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3	88.4
USSR	9.3	4.6	2.1	16.6	3.0	43.6	6.4	3.4	2.9	91.9
WGER	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8	79.3
YUGO	4.4	5.0	1.2	9.5	0.6	55.9	3.0	5.7	3.2	88.5
Total	245.7	197.4	73.4	427.8	107.1	806.2	106.9	76.8	103.4	2144.7

Abbreviations

ALBA	Albania	AUST	Austria	BELX	Belgium/Luxembourg
BULG	Bulgaria	CZEC	Czechoslovakia	DENM	Denmark
EGER	East Germany	FINL	Finland	FRAN	France
GREE	Greece	HUNG	Hungaria	IREL	Ireland
ITAL	Italy	NETH	Netherlands	NORW	Norway
POLA	Poland	PORT	Portugal	RUMA	Rumania
SPAI	Spain	SWED	Sweden	SWIT	Switzerland
UK	United Kingdom	USSR	Russia	WGER	West Germany
YUGO	Yugoslavia	-	-	-	-
MEAT	Meat(grazing)	PIPL	Pigs & Poultry	EGGS	Eggs
MILK	Milk	FISH	Fish	CERS	Cereals
STAR	Starch	NUTS	Nuts/Pulses	FRVG	Fruit/Vegetables

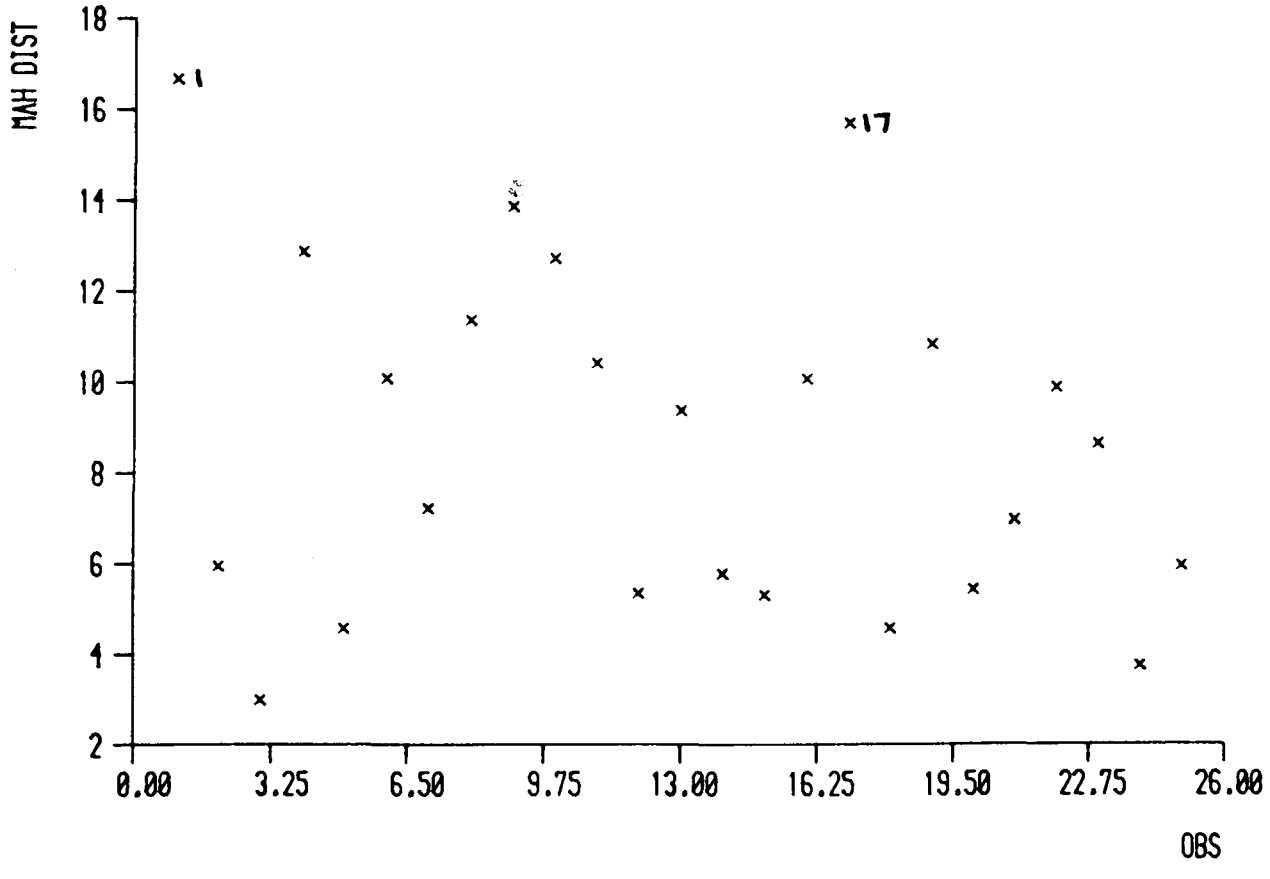
Table 4.9.2

Most Influential Observations on the First Two Dimensions of the Covariance PCA of the Protein Consumption Dataset

$\lambda_1 = 155.23$			$\lambda_2 = 30.70$		
ranked PC score	Obsn.	Infl.	ranked PC score	Obsn.	Infl.
1	4	15.4%	1	17	27.4%
2	25	14.8%	2	8	11.9%
3	18	6.2%	3	19	10.9%

	α_1			α_2	
ranked PC score	Obsn.	Infl.	ranked PC score	Obsn.	Infl.
8	8	3.2°	1	17	30.9°
1	4	2.9°	2	8	15.5°
2	25	2.7°	3	19	8.5°

Figure 4.9.1 Plot of the Mahalanobis Distances for the Dataset on Protein Consumption in Europe.



eigenvector, and it was noted that the bounds would not be so good for the kth axis if previous axes had rotated when the observation was omitted. The upper bounds given by 3.8.14 and 3.8.15 for the observations in Table 4.9.2 are given in 4.9.3. For the first axis the ordering of the observations by the refined bound was almost the same as by the order of sample influence, and the upper bound was always greater. The same is true for the second axis, apart from observation 8 and observations 25 and 4 that were given larger upper bounds than observation 19 but they only had the 7th and 8th largest influences respectively. The bounds were less accurate for the latter principal axes especially when, as noted by Greenacre, an observation had a large affect on the earlier components. For example, when observation 17 was omitted $\hat{\underline{\alpha}}_2$ and $\hat{\underline{\alpha}}_3$ changed by 30.9° and 34.1° respectively but the refined bound (and the simple bound was similar) for $\hat{\underline{\alpha}}_3$ was 4.1°. Observation 8 had an angular change of 19.4° on $\hat{\underline{\alpha}}_3$ and its bound is slightly above that for observation 17 at 4.8°. Most of the refined bounds in the latter dimensions were too small and often the simple bound was better here, but this was still to small if an observation had affected previous dimensions. For example, in the 6th dimension the ranked observations using the actual angle between original and perturbed eigenvectors were 1,9,10,19 and 23. The simple bound for observations 10 and 23 are larger than the actual angle, and these observations had not affected previous directions. However, the bounds for observations 1,9 and 19 were small. The refined bounds for observations 10 and 23 were the largest, but they were smaller than the actual angle. As noted in § 3.8.4, we cannot obtain the upper bounds for the last principal component which is likely to be of more interest than the principal components in the middle.

From Tables 4.9.2 we see that the first principal axis does not change

much when a single observation is omitted. The angular changes in the second axis are larger, and we will consider the affect that the two most influential observations on this axis have on the corresponding covariance biplot. The original and perturbed eigenvectors in the second dimension when observations 17 (Portugal) and 8 (Finland) are removed are given in Table 4.9.4. Each eigenvector has large coefficients for the variables 'milk' and 'cereal' (Note, that the first axis is a contrast of these variables), but when observation 'Finland' is removed these are contrasted more with the 'fish' variable. Gabriel (1981) notes that the fish variable is at about 90° from the animal and cereal sources of protein on the (original) covariance biplot (see Fig 4.9.2) and so concludes that 'fish' must be 'pretty uncorrelated' with these sources of protein. When the observation 'Finland' is removed the negative coefficient for fish becomes larger indicating that protein sources 'milk' and 'cereals' do not tend to be consumed by the same countries who have a high consumption of fish. Finland has the largest consumption of milk but its consumption of the other animal proteins, especially on pigs/poultry is quite low or average. In the original biplot Finland is positioned close to the milk marker, as are Ireland and Switzerland who also have a high consumption of milk. When Finland is removed the milk marker remains relatively unchanged but the angles between the four animal products become smaller. Otherwise the biplot is largely unchanged (see Fig 4.9.3).

When observation 17(Portugal) is omitted the 'fish' coefficient becomes smaller and less prominent in the interpretation of the second axis. This is the opposite affect to that when 'Finland' was omitted. However, the actual change in the 'fish' coefficient is larger than when 'Finland' was omitted. If we look at the data in Table 4.9.1 we see that after Portugal the next highest consumers of fish are Denmark and Norway, but like all the observations high

on the milk variable these two countries are placed far from the 'fish' marker. Fig 4.9.4 is the biplot when 'Portugal' is removed. The RHS of the plot looks different to the original biplot but the LHS is almost unaltered. The 'fish' marker in Fig. 4.9.4 has moved closer to 'Denmark' and 'Norway' and is close to the 'meat' marker. This alters the interpretation for the original biplot given by Gabriel, and outlined above. The 'fish' protein, like the animal proteins is now associated with the Northern and Western countries and not the Mediterrenean countries as noted by Gabriel. In Fig. 4.9.4 the animal products have sprayed outwards with the 'pigs/poultry' and 'eggs' markers rotating clockwise, their coefficients in the second eigenvector having changed sign. The 'pigs/poultry' marker has almost moved to where the original 'fish' marker was, and countries 'Austria' and the 'Netherlands', which have a high consumption of pigs/poultry and eggs, have moved down.

The original and perturbed (for Portugal omitted) correlation matrices for the animal protein variables are given in Table 4.9.5. These help to explain the changes in the 'pigs/poultry' marker. 'Pigs/poultry' has a low correlation with all the other animal products, except with 'eggs' which may occur due to the poultry category (this suggests that 'pigs' and 'poultry' should not have been combined together to form a variable, but perhaps eggs and poultry may have been better). In the original biplot the 'pigs/poultry' marker was probably close to these other animal products, that it has a low correlation with, due to its negative correlation with the 'fish' variable, which meant it could not placed below the 'egg' marker. In the perturbed biplot where the 'fish' marker moves closer to the 'milk' marker, the marker for 'pigs/poultry' can rotate clockwise with less restraint and so becomes placed further from the 'meat' and 'milk' markers. The marker for 'eggs' also moves with the 'pigs/poultry' but less so, as it is still quite highly correlated with the other

Table 4.9.3

Upper Bounds for the Angle of Rotation in the Eigenvectors

$\underline{\alpha}_1$			$\underline{\alpha}_2$		
Obsn.	Simple	Refined	Obsn.	Simple	Refined
8	3.35°	3.3°	17	35.7°	31.9°
4	7.13°	3.2°	8	16.1°	12.3°
25	6.88°	2.9°	19	11.3°	10.8°

Table 4.9.4

Original and Perturbed Second Eigenvector

Variable	Original $\underline{\alpha}_2$	When Finland Omitted	When Portugal Omitted
MEAT	0.13	0.21	0.07
PIPL	0.04	0.24	-0.32
EGGS	0.02	0.06	-0.05
MILK	0.83	0.75	0.85
FISH	-0.29	-0.41	0.07
CEREALS	0.41	0.35	0.38
STAR	-0.08	-0.08	-0.06
NUTS	-0.07	-0.09	0.00
FRVG	-0.17	-0.15	-0.11

Table 4.9.5

Original and Perturbed (for 'Portugal' Omitted) Correlation Matrix

Between the Animal Products

	PIPL	EGGS	MILK	FISH
MEAT	0.15	0.59	0.50	0.06
PIPL		0.62	0.28	-0.23
EGGS			0.58	0.07
MILK				0.14

	PIPL	EGGS	MILK	FISH
MEAT	0.11	0.56	0.46	0.26
PIPL		0.59	0.22	-0.12
EGGS			0.52	0.37
MILK				0.48

Figure 4.9.2 Original Covariance Biplot.

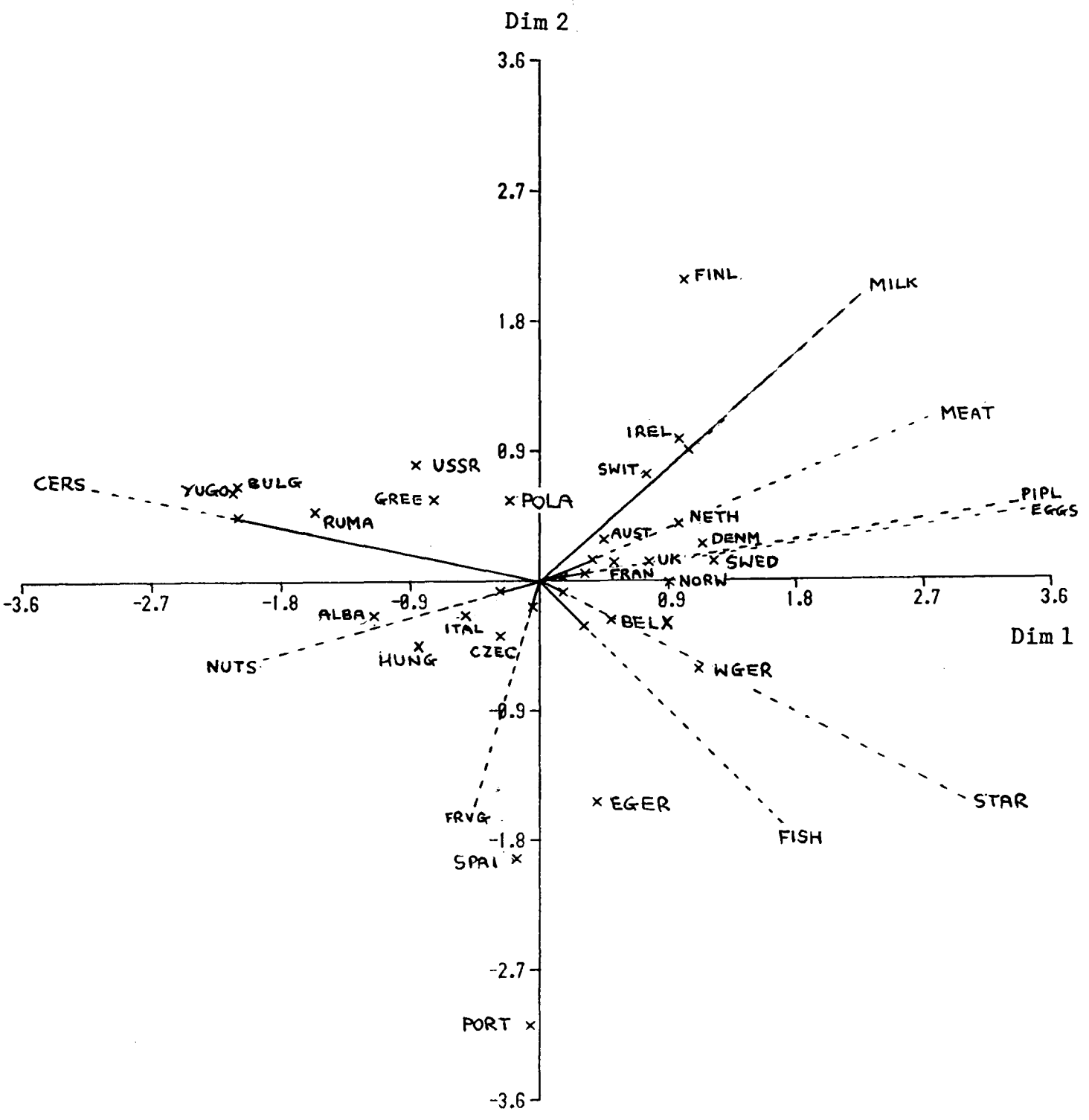


Figure 4.9.3 Covariance Biplot when Observation 8 (Finland) is Omitted.

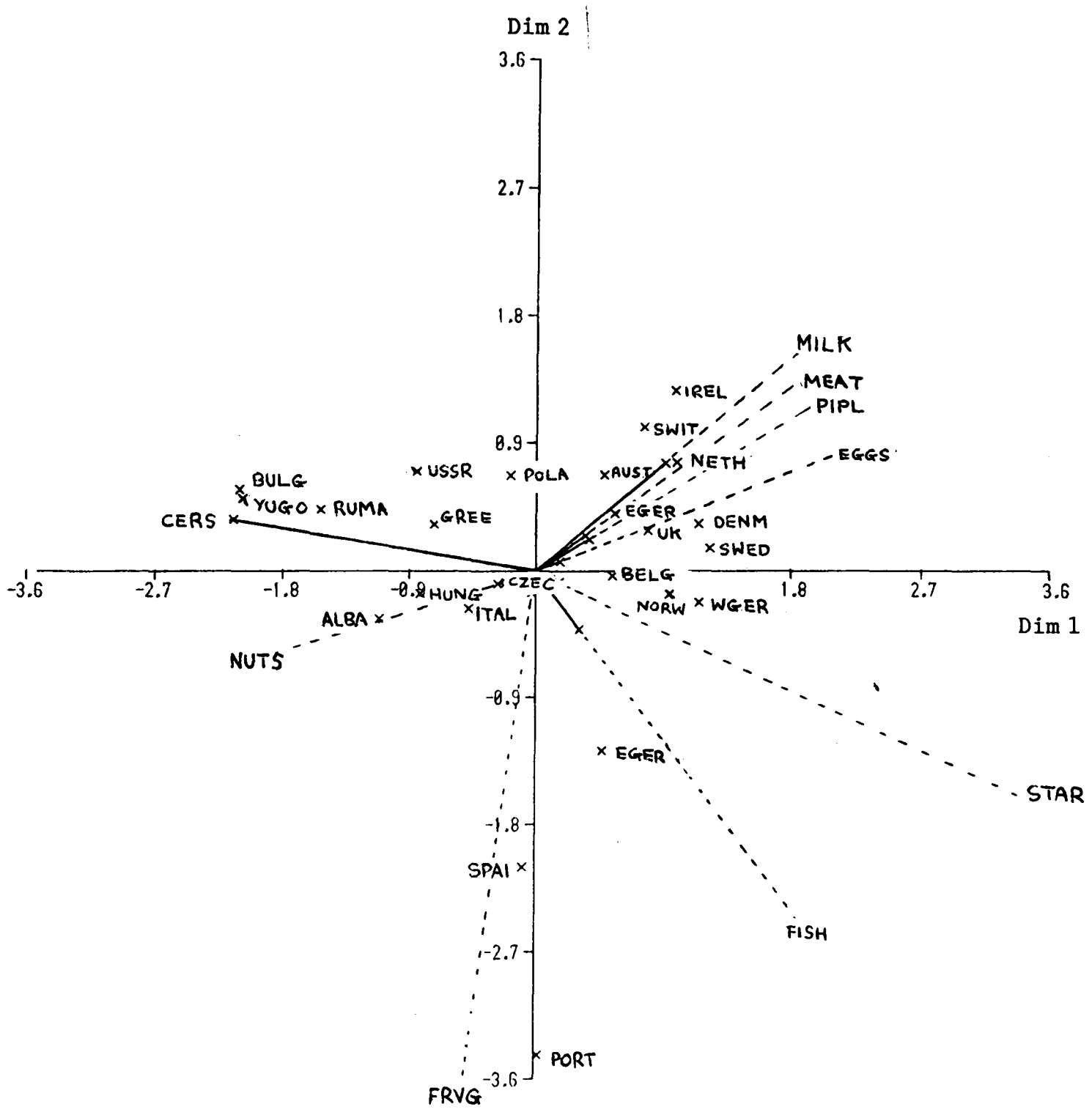
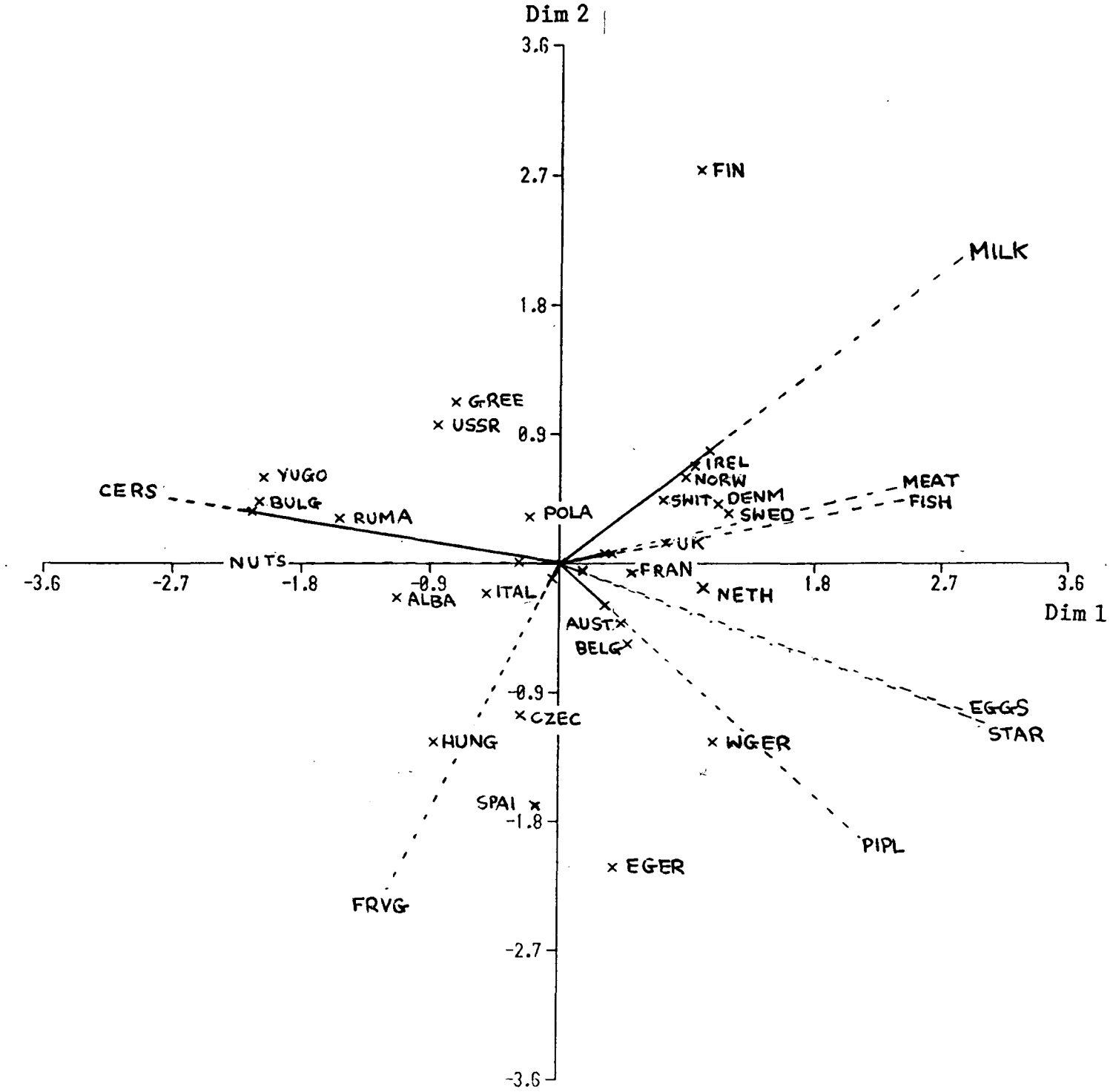


Figure 4.9.4 Covariance Biplot when Observation 17 (Portugal) is Omitted.



animal proteins.

The only noticeable change to the LHS of the biplot when 'Portugal' is omitted is the position of 'Hungary', that has moved downwards close to the 'fruit/vegetables' marker. From the original data in Table 4.9.1 we see that Hungary has a high consumption of pigs/poultry but a low consumption of fish. The perturbed biplot reflects this better than the original one does.

It would be interesting to see the position of a 'pigs' marker on its own as it is possible that it positioning, even in the biplot when 'Portugal' is omitted, is due to a high correlation of 'poultry' and 'eggs'. The positioning of the combined marker for 'pigs/poultry' in the perturbed biplot does seem to reflect its relationships with the other variables better than in the original. We will analyse this data further in § 6.3 using correspondence analysis, and we will see that it the pigs/poultry mark remains quite steady in the resulting two dimensional display when 'Portugal' is again omitted.

Chapter 5: Derivation of Influence Functions in Canonical Correlation Analysis and Correspondence Analysis

5.1. Introduction

5.1.1. Canonical Correlation Analysis

In canonical correlation analysis we are interested in the relationships between two sets of variables $Y_{(n\,p_1)}$ and $X_{(n\,p_2)}$ which are collected on the same set of individuals, as opposed to PCA where we are concerned with relationships within one set of variables. This is an extension of multiple linear regression and, when $p_1 = 1$, the only non-zero canonical correlation is R^2, the multiple correlation coefficient discussed in § 2.4. We form new variables, namely the linear combinations $\underline{a}_k'\underline{y}$ and $\underline{b}_k'\underline{x}$, such that these variables have maximum possible correlation $\lambda_k^{1/2}$ where $\lambda_1^{1/2} \geq \lambda_2^{1/2} \geq \ldots\ldots \geq \lambda_{p_1}^{1/2}$ (assuming $p_1 < p_2$) and the $\underline{a}_k$ and $\underline{b}_k$ are orthogonal, $k = 1, 2, \ldots\ldots, p_1$. The correlation between $\underline{a}_1'\underline{y}$ and $\underline{b}_1'\underline{x}$ is found by maximising

$$\frac{\underline{a}_1'\Sigma_{yx}\underline{b}_1}{\left(\underline{a}_1'\Sigma_{yy}\underline{a}_1\right)^{1/2}\left(\underline{b}_1'\Sigma_{xx}\underline{b}_1\right)^{1/2}} \tag{5.1.1}$$

or alternatively maximising $\underline{a}_1'\Sigma_{yx}\underline{b}_1$ subject to $\underline{a}_1'\Sigma_{yy}\underline{a}_1 = \underline{b}_1'\Sigma_{xx}\underline{b}_1 = 1$. We obtain $\underline{a}_k$, $\underline{b}_k$ and $\lambda_k^{1/2}$ from the eigen-relationships

$$\left(\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} - \lambda_k I\right)\underline{a}_k = 0 \tag{5.1.2}$$

$$\left(\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} - \lambda_k I\right)\underline{b}_k = 0 \tag{5.1.3}$$

(see Mardia *et al.*, 1979, Chapter 10). Assuming $p_1 < p_2$, we would solve (5.1.2) for $\underline{a}_k$ and obtain $\underline{b}_k$ from

$$\underline{b}_k = \frac{1}{\lambda_k^{1/2}}\Sigma_{xx}^{-1}\Sigma_{xy}\underline{a}_k \tag{5.1.4}$$

which also prevents problems with the arbitrariness of sign from using both (5.1.2) and (5.1.3). If $p_1 > p_2$ we would use (5.1.3) so that the eigenanalysis

is performed on the smaller matrix.

The matrix in (5.1.2) is not symmetric but we can use our usual symmetric eigenvalue routines by forming the matrix C such that

$$\Sigma_{yy}^{-1} = CC'$$

where C can be lower triangular with positive diagonal elements, as used by Radhakrishnan and Kshirsagar (1981), or alternatively C can be $\Sigma_{yy}^{-1/2}$ calculated by noting, if Σ_{yy} has spectral decomposition $\Gamma\Lambda\Gamma'$ then Σ_{yy}^{-1} has spectral decomposition $\Gamma\Lambda^{-1}\Gamma'$ and $\Sigma_{yy}^{-1/2}$ has spectral decomposition $\Gamma\Lambda^{-1/2}\Gamma'$.

We find the eigenvalues and eigenvectors γ_k and $\underline{\alpha}_k$ from the symmetric matrix $C'\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}C$ so that,

$$\left(C'\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}C - \gamma_k I\right)\underline{\alpha}_k = 0 \qquad (5.1.5)$$

and, by multiplying on the left by C, we obtain

$$\left(\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} - \gamma_k I\right)C\underline{\alpha}_k = 0 \quad . \qquad (5.1.6)$$

Comparing (5.1.6) with (5.1.2) we have $\lambda_k = \gamma_k$ and $\underline{a}_k = C\underline{\alpha}_k$.

The above approach will be used in finding the influence functions in this chapter.

5.1.2. Correspondence Analysis of a Two-Way Contingency Table

This technique was developed by Benzecri in the 1960s and many articles in French have been written on the subject. Below is a brief outline of the method which is required to understand the algebra in this chapter. For a detailed account, one is referred to Greenacre (1984).

Correspondence analysis, dual (or optimal) scaling and reciprocal averaging all require similar calculations, but the rationale behind each method, and so the presentation of results, differs. The aim of correspondence analysis is to obtain a joint display of the rows and columns of the

contingency table such that row (or column) co-ordinates that are close together have similar row (column) profiles, which are the rows (columns) divided by their sums, and a row co-ordinate will tend to be close to a column co-ordinate that is prominent in its profile.

Correspondence analysis is a special form of canonical correlation analysis with dummy variables (see below) and it is for this reason both techniques have been considered in the same chapter. One of our influence techniques for correspondence analysis is derived as a special form of the canonical correlation influence functions derived in § 5.2.

Let P be a two-way contingency table, N, divided through by the grand total $n_{..}$, and $\underline{r}$ and $\underline{c}$ be column vectors of the row and column totals of P respectively. Then the matrix $P - \underline{r}\underline{c}'$, whose $(i\,j)$th element is $(n_{ij} - n_{i.}\,n_{.j}/n_{..})/n_{..}$, $i = 1,2,....,I$, $j = 1,2,....,J$ (i.e the table has I rows and J columns), is the matrix of residuals from fitting a model of independence between the rows and columns, divided by through by $n_{..}$. If D_r (D_c) is the diagonal matrix of $\underline{r}$ $(\underline{c})$ then the chi-square statistic calculated on our contingency table, divided by $n_{..}$, is

$$\frac{1}{n_{..}}\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij} - r_i c_j)^2}{r_i c_j}$$

$$= \text{trace}\left(D_r^{-1}(P - \underline{r}\underline{c}')D_c^{-1}(P - \underline{r}\underline{c}')'\right)$$

$$= \text{trace}\left(D_c^{-1}(P - \underline{r}\underline{c}')'D_r^{-1}(P - \underline{r}\underline{c}')\right)$$

The eigenvalues and eigenvectors for these matrices are found from solving

$$\left(D_r^{-1}(P - \underline{r}\underline{c}')D_c^{-1}(P - \underline{r}\underline{c}')' - \lambda_k I\right)\underline{f}_k = 0 \qquad (5.1.7)$$

$$\left(D_c^{-1}(P - \underline{r}\underline{c}')'D_r^{-1}(P - \underline{r}\underline{c}') - \lambda_k I\right)\underline{g}_k = 0 \qquad (5.1.8)$$

with normalisation $\underline{f}_k'D_r\underline{f}_k = \underline{g}_k'D_c\underline{g}_k = \lambda_k$. If $J < I$ we would only solve (5.1.8) and obtain $\underline{f}_k$ from the transition formula

$$\underline{f}_k = \frac{1}{\lambda_k^{1/2}} D_r^{-1} P \underline{g}_k \quad . \tag{5.1.9}$$

We shall assume $J < I$ throughout this chapter. We do not need to centre the P matrix, but can solve

$$\left(D_c^{-1} P' D_r^{-1} P - \lambda_k I\right) \underline{g}_k = 0 \tag{5.1.10}$$

and use the same transition formula, (5.1.9). This results in a trivial dimension of $\lambda_1 = 1$ and $\underline{f}_1 = \underline{1}_I$ and $\underline{g}_1 = \underline{1}_J$. This trivial dimension is useful in checking our algebraic results, which should give zero influences. Equations (5.1.10) and (5.1.9) are similar to (5.1.2) and (5.1.4), and we shall return to this below.

The row (column) profiles are the rows (columns) of the contingency table divided by the row (column) totals. The larger is $\chi^2/n_{..}$, the more association there is between the rows and columns, and so the more our row (column) profiles differ. The sum of our eigenvalues $\sum_{k=2}^{J} \lambda_k = \chi^2/n_{..}$, and if the first two (non-trivial) eigenvalues (principal inertias) account for most of the total inertia, $(\chi^2/n_{..})$, then a plot of the co-ordinates $(f_{2i}\, f_{3i})$, $i = 1, 2, \ldots, I$, will reveal which of the row categories have the most similar and dissimilar row profiles by their distances apart. We also plot, on the same graph, the co-ordinates $(g_{2j}\, g_{3j})$, $j = 1, 2, \ldots, J$, and the same relationship between the plotted column categories holds.

The relationship defined between the rows and columns are determined by the transition formula (5.1.9). This shows that a row co-ordinate will tend towards a column co-ordinate that is prominent in its profile. The distances on the plot between any two rows (or columns) are defined as chi-squared distances, but although we interpret rows and columns that are close together on the plot, the actual distances between them are less meaningful.

The above may seem unusual in that we are actually plotting the eigenvectors rather than any sort of principal component score. In fact, the F co-ordinates (likewise the G co-ordinates) can be calculated as the principal component scores in the metric D_c^{-1} (D_r^{-1}) from the eigenvectors of the matrix of row (column) profiles. Alternatively, they can both be found as principal co-ordinates of the matrix P (or $P - \underline{r}\underline{c}'$). If the generalised singular value decomposition (SVD) of P is

$$P = AD_{\lambda^{1/2}}B \qquad A'D_r^{-1}A = B'D_c^{-1}B = I$$

then, in the metrics D_c^{-1} and D_r^{-1} respectively,

$$F = D_r^{-1}AD_{\lambda^{1/2}} \quad \text{and} \quad G = D_c^{-1}BD_{\lambda^{1/2}}$$

(for further details see Greenacre, 1984, chapter 4).

Above we noted the similarity between the eigen-equations (5.1.10) and (5.1.2) and the transition formulae (5.1.9) and (5.1.4). If we perform a canonical correlation analysis on two sets of dummy variables $Y_{(n\,J)}$ and $X_{(n\,I)}$ where

$$Y_{Lj} = \begin{cases} 1 & \text{If the } L\text{th individual belongs} \\ & \text{to the } j\text{th column category} \\ 0 & \texttt{Otherwise} \end{cases}$$

$$X_{Li} = \begin{cases} 1 & \text{If the } L\text{th individual belongs} \\ & \text{to the } i\text{th row category} \\ 0 & \texttt{Otherwise} \end{cases} \quad , \tag{5.1.11}$$

$L = 1,2,\ldots,n$, $i = 1,2,\ldots,I$, $j = 1,2,\ldots,J$, but we do not centre our variables, we obtain from the sample version of (5.1.1)

$$\left((Y'Y)^{-1}(Y'X)(X'X)^{-1}(X'Y) - \lambda_k I\right)\underline{a}_k = 0$$

but

$$\frac{1}{n_{..}}Y'Y = D_c \qquad \frac{1}{n_{..}}X'X = D_r \qquad \frac{1}{n_{..}}X'Y = P \quad , \tag{5.1.12}$$

which gives us (5.1.10). Similarly, substituting (5.1.12) into (5.1.4) gives (5.1.9). If we centre our variables then $S_{XY} = P - \underline{r}\underline{c}'$ but S_{YY} and S_{XX} would be singular. Because we have dummy variables, our observations are $\underline{z}_j' = (\underline{y}_L' \, \underline{x}_L') = (00..010..00 \;\; 00..010..00)$, where the "1"s are in the Y_{Lj}th and X_{Li}th places of $\underline{y}_L$ and $\underline{x}_L$ respectively. Therefore, our canonical scores, $\underline{y}'\underline{a}_k$ and $\underline{x}'\underline{b}_k$, just pick out the jth and ith elements of our eigenvectors $\underline{a}_k$ and $\underline{b}_k$ respectively. In correspondence analysis these scores are then displayed for the first two, non-trivial, dimensions on the same plot. For a more detailed account of the relationship between canonical and correspondence analysis see Greenacre (1984, pp108-116).

5.1.3. Multiple Correspondence Analysis

Let Z be the $n_{..} \times (I + J)$ bivariate indicator matrix $[X \; Y]$ where X and Y are made up of variables, as defined in (5.1.11), and the contingency table $N = n_{..}P = X'Y$. The correspondence analysis of the matrix Z yields a plot of the row and column co-ordinates which is a re-scaled version of the plotted row and column co-ordinates from a correspondence analysis of the contingency table N. The eigenvalues (principal inertias) of the indicator matrix are related to those of the contingency table by

$$\lambda_k^Z = \frac{1}{2}(1 \pm \lambda_k^{N\ 1/2}) \quad . \tag{5.1.13}$$

This results in double the number of eigenvalues for the indicator matrix analysis and there are a further $I - J$ (assuming $J < I$) inertias, $\lambda^Z = 1/2$. The percentages of inertia are thus much smaller and there are smaller differences in successive (ranked) inertias. However, these extra dimensions are unimportant, as are those in (5.1.13) that are less than $1/2$, which result from the negative roots. This justifies the proposal of Benzecri (1979) to recalculate the percentages of inertia based only on those λ^Z exceeding $1/2$ and on the

values $\lambda^Z - 1/2$.

The correspondence analysis of a multivariate indicator matrix, $Z = [Z_1 Z_2 .... Z_Q]$, for Q variables having J_j categories, $j = 1, 2, ..., Q$, yields a plot of the column co-ordinates which are a re-scaled version of the column (or row) co-ordinates of the symmetric Burt matrix. The Burt matrix, $B_{(J \times J)} = Z'Z$, where $J = \sum_{j=1}^{Q} J_j$, contains on its off-diagonal all the two-way contingency tables between the Q sets of variables. On its diagonal, it has diagonal matrices of the category sums, i.e.

$$B = \begin{pmatrix} N_{11} & N_{12} & .. & .. & .. & .. & N_{1Q} \\ N_{12}' & N_{22} & .. & .. & .. & .. & N_{2Q} \\ .. & .. & .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. & .. & .. \\ N_{1Q}' & N_{2Q}' & .. & .. & .. & .. & N_{QQ} \end{pmatrix} .$$

Since it is symmetric, the row and column co-ordinates are the same and its eigenvalues have the relationship

$$\lambda_k^B = (\lambda_k^Z)^2$$

with the eigenvalues from the multivariate indicator matrix. When $Q = 2$, the eigenvalues from the Burt matrix, using (5.1.13), are therefore

$$\lambda_k^B = \frac{1}{4}(1 \pm \lambda_k^{N\ 1/2})^2 \quad .$$

Again we obtain unimportant extra dimensions and Greenacre (1984, p144-145) gives a formula for recalculating the percentages of inertia. If we let H denote the column (or row) co-ordinates from the Burt matrix correspondence analysis, then H can be divided into the categories for each variable. Then, the expression for the co-ordinates, on the kth axis, for the categories of the qth variable, in terms of the other variables, is

$$\underline{h}_{kq} = \frac{1}{Q\lambda_k^{B\ 1/2} - 1} \sum_{q' \neq q}^{Q} D_q^{-1} P_{qq'} \underline{h}_{kq'}$$

where $R_{qq'} = D_q^{-1} P_{qq'}$ is the matrix of row profiles for the contingency table $N_{qq'}$. This expression is an extension of the transition formula (5.1.9). Thus, the category co-ordinates of one variable will tend towards category co-ordinates of other variables that are dominant in any of its row profiles across the $(Q-1)$ contingency tables involving it.

5.1.4. Summary of Chapter

In § 5.2 we derive the theoretical influence functions for the eigenvalues and eigenvectors in canonical correlation analysis. Radhakrishnan and Kshirsagar (1981) derived the influence function for the eigenvalues and the same approach is used here, but the expression for the eigenvalues is simplified and we derive the theoretical influence function for the eigenvectors. In § 5.2.3 we consider the specialisation of the influence function for the canonical eigenvalues to the multiple correlation coefficient. This influence function was derived in § 2.4.2, by an alternative method. In § 5.3 the influence functions derived in § 5.2 are used to obtain the influence functions for correspondence analysis when we add an extra observation, so that a cell of the contingency table is incremented by one. This uses the relationship between canonical correlation analysis and correspondence analysis discussed in § 5.1.2.

In § 5.4 we derive expressions for the changes in the eigenvalues and the G and F co-ordinates in correspondence analysis when we add a row to a contingency table. These expressions are derived in a slightly different way to the previous ones. They are derived as an asymptotic result by expanding out terms to $o(1/(n_{..}+m))$ where m is the sum of the new row. If we take $\epsilon = 1/(n_{..}+m)$ we can follow the usual derivations of the theoretical influence functions for the eigenvalues and eigenvectors. Since the expressions are derived as $n \to \infty$ we will refer to the final expressions immediately as the

empirical influence curve. A similar approach is used for the our final type of perturbation, which is adding m identical observations so a cell of a multiway table is incremented by m. This leads to Q^2 elements of the Burt matrix, introduced in § 5.1.3, increasing by m. Since we can do an analysis on the Burt matrix when we have just two variables we would expect some similarities of influence by perturbation of the Burt matrix, as by the perturbation of a cell of a contingency table. This is discussed in detail in § 6.4.

The second type of perturbation, adding a row to a contingency table, is different from our usual influence procedures in that we obtain an extra co-ordinate in the perturbed problem for the row added. The term in ϵ^0 for this co-ordinate in the perturbed problem is the same expression as that used by Greenacre (1984, p70-74) to add a supplementary point to an existing display. If we add an extra column rather than a row where $J < I$ we would obtain an additional dimension in the perturbed analysis. We will not consider this problem in this chapter, but in § 6.3.5 we shall examine the extension of the formulae in § 5.4 to deal with addition (deletion) of columns.

5.2. Influence Functions in Canonical Correlation Analysis

5.2.1. Influence Function for the Eigenvalues

We wish to find the influence function for the eigenvalues from the symmetric matrix $C'\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}C$, see (5.1.5), which are the same as the eigenvalues from the unsymmetric matrix $\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$, where $\Sigma_{yy}^{-1} = CC'$. Since the matrix $C'\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}C$ is symmetric, if its eigenvectors are $\underline{\alpha}_k$, such that $\underline{\alpha}_k'\underline{\alpha}_k = 1$ then, from (3.5.3)

$$TIC(\underline{z}, \lambda_k) = \underline{\alpha}_k' TIC(\underline{z}, C'\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}C)\underline{\alpha}_k \quad , \tag{5.2.1}$$

where $\underline{z}$ is the added in point $(\underline{y}, \underline{x})$. From (2.4.4)

$$\tilde{\Sigma}_{yy}^{-1} = \Sigma_{yy}^{-1} + \epsilon\left(\Sigma_{yy}^{-1} - \Sigma_{yy}^{-1}(\underline{y} - \underline{\mu}_y)(\underline{y} - \underline{\mu}_y)'\Sigma_{yy}^{-1}\right) + o(\epsilon^2) \quad , \tag{5.2.2}$$

so that if $\Sigma_{yy}^{-1} = CC'$, we can see

$$\tilde{C} = C + \frac{\epsilon}{2}\left(C - \Sigma_{yy}^{-1}(\underline{y} - \underline{\mu}_y)(\underline{y} - \underline{\mu}_y)'C\right) + o(\epsilon^2) \quad , \tag{5.2.3}$$

will give us (5.2.2). Simple algebra using (5.2.2) and (2.3.3) and the product rule for influence functions discussed in Chapter 1 (or by multiplying together the perturbed terms) gives us,

$$\begin{aligned} TIC(\underline{z}, C'\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}C) = C' \Big[& -\Sigma_{yx}\Sigma_{xx}^{-1}(\underline{x} - \underline{\mu}_x)(\underline{x} - \underline{\mu}_x)'\Sigma_{xx}^{-1}\Sigma_{xy} \\ & + (\underline{y} - \underline{\mu}_y)(\underline{x} - \underline{\mu}_x)'\Sigma_{xx}^{-1}\Sigma_{xy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} \\ & + \Sigma_{yx}\Sigma_{xx}^{-1}(\underline{x} - \underline{\mu}_x)(\underline{y} - \underline{\mu}_y)' \Big] C \\ & + \frac{1}{2}\left(C' - C'(\underline{y} - \underline{\mu}_y)(\underline{y} - \underline{\mu}_y)'\Sigma_{yy}^{-1}\right)\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}C \\ & + \frac{1}{2}C'\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\left(C - \Sigma_{yy}^{-1}(\underline{y} - \underline{\mu}_y)(\underline{y} - \underline{\mu}_y)'C\right) \quad . \end{aligned} \tag{5.2.4}$$

Simplifying (5.2.4) gives

$$\begin{aligned} TIC(\underline{z}, C'\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}C) = & -C'\Sigma_{yx}\Sigma_{xx}^{-1}(\underline{x}-\underline{\mu}_x)(\underline{x}-\underline{\mu}_x)'\Sigma_{xx}^{-1}\Sigma_{xy}C \\ & + C'(\underline{y}-\underline{\mu}_y)(\underline{x}-\underline{\mu}_x)'\Sigma_{xx}^{-1}\Sigma_{xy}C \\ & + C'\Sigma_{yx}\Sigma_{xx}^{-1}(\underline{x}-\underline{\mu}_x)(\underline{y}-\underline{\mu}_y)'C \\ & - \frac{1}{2}C'(\underline{y}-\underline{\mu}_y)(\underline{y}-\underline{\mu}_y)'\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}C \\ & - \frac{1}{2}C'\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}(\underline{y}-\underline{\mu}_y)(\underline{y}-\underline{\mu}_y)'C \quad . \end{aligned} \tag{5.2.5}$$

From (5.1.6), $C\underline{\alpha}_k = \underline{a}_k$ and since a scalar is equal to its transpose, we can write (5.2.1), using (5.2.5), as

$$\begin{aligned} TIC(\underline{z}, \lambda_k) = & -\left[\underline{a}_k'\Sigma_{yx}\Sigma_{xx}^{-1}(\underline{x}-\underline{\mu}_x)\right]^2 \\ & + 2\underline{a}_k'(\underline{y}-\underline{\mu}_y)(\underline{x}-\underline{\mu}_x)'\Sigma_{xx}^{-1}\Sigma_{xy}\underline{a}_k \\ & - \underline{a}_k'(\underline{y}-\underline{\mu}_y)(\underline{y}-\underline{\mu}_y)'\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}\Sigma_{xy}\underline{a}_k \quad . \end{aligned}$$

Using (5.1.2), (5.1.4) and letting the canonical scores be

$$R_{ky} = (\underline{y}-\underline{\mu}_y)'\underline{a}_k \quad \text{and} \quad R_{kx} = (\underline{x}-\underline{\mu}_x)'\underline{b}_k \quad ,$$

then,

$$TIC(\underline{z}, \lambda_k) = -\lambda_k R_{kx}^2 + 2\lambda_k^{1/2} R_{kx} R_{ky} - \lambda_k R_{ky}^2 \quad . \tag{5.2.6}$$

The kth canonical correlation is $\lambda_k^{1/2}$ and its influence function is

$$TIC(\underline{z}, \lambda_k^{1/2}) = \frac{TIC(\underline{z}, \lambda_k)}{2\lambda_k^{1/2}} \quad , \tag{5.2.7}$$

since

$$\tilde{\lambda}_k = \lambda_k + \epsilon TIC(\underline{z}, \lambda_k) + o(\epsilon^2) \quad ,$$

so that

$$\tilde{\lambda}_k^{1/2} = \lambda_k^{1/2}\left[1 + \epsilon\frac{TIC(\underline{z}, \lambda_k)}{\lambda_k} + o(\epsilon^2)\right]^{1/2} .$$

Expanding the bracket, we obtain

$$\tilde{\lambda}_k^{1/2} = \lambda_k^{1/2} + \epsilon\frac{TIC(\underline{z}, \lambda_k)}{2\lambda_k^{1/2}} + o(\epsilon^2) \quad ,$$

and hence (5.2.7).

5.2.2. Influence Functions for the Canonical Vectors

First we shall obtain the influence function for the eigenvector $\underline{\alpha}_k$ from the symmetric matrix $C'\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}C$ and then for $\underline{a}_k = C\underline{\alpha}_k$. Using (3.5.9), we have

$$TIC(\underline{z}, \underline{\alpha}_k) = -\sum_{\substack{t=1 \\ t\neq k}}^{J} \underline{\alpha}_t(\lambda_t - \lambda_k)^{-1}\underline{\alpha}_t' TIC(\underline{z}, C'\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}C)\underline{\alpha}_k \quad .$$

Using (5.2.5), (5.1.2), (5.1.4) and $\underline{a}_k = C\underline{\alpha}_k$, we find

$$\begin{aligned}
\underline{\alpha}_t' TIC(\underline{z}, C'\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}C)\underline{\alpha}_k &= -\lambda_t^{1/2}\lambda_k^{1/2}R_{tx}R_{kx} \\
&\quad + \lambda_k^{1/2}R_{ty}R_{kx} + \lambda_t^{1/2}R_{tx}R_{ky} - \frac{1}{2}\lambda_k R_{ty}R_{ky} \\
&\quad - \frac{1}{2}\lambda_t R_{ty}R_{ky} \\
&= -(\lambda_t^{1/2}R_{tx} - R_{ty})(\lambda_k^{1/2}R_{kx} - R_{ky}) \\
&\quad + (1 - \frac{1}{2}\lambda_k - \frac{1}{2}\lambda_t)R_{ky}R_{ty} \quad ,
\end{aligned}$$

so

$$TIC(\underline{z}, \underline{\alpha}_k) = -\sum_{\substack{t=1 \\ t\neq k}}^{J} \underline{\alpha}_t(\lambda_t - \lambda_k)^{-1}\left[-(\lambda_t^{1/2}R_{tx} - R_{ty})(\lambda_k^{1/2}R_{kx} - R_{ky}) + (1 - \frac{1}{2}(\lambda_k + \lambda_t))R_{ky}R_{ty}\right] . \qquad (5.2.8)$$

From either using the product rule or multiplying together the perturbed $\tilde{C}\tilde{\underline{\alpha}}_k$, to $o(\epsilon)$,

$$TIC(\underline{z}, \underline{a}_k) = TIC(\underline{z}, C)\underline{\alpha}_k + C\, TIC(\underline{z}, \underline{\alpha}_k)$$

where,

$$\tilde{C} = C + \epsilon TIC(\underline{z}, C) + o(\epsilon^2)$$

and

$$\tilde{\underline{\alpha}}_k = \underline{\alpha}_k + \epsilon TIC(\underline{z}, \underline{\alpha}_k) + o(\epsilon^2) \quad .$$

Using (5.2.8) and (5.2.3) we have that,

$$TIC(\underline{z}, \underline{a}_k) = \frac{1}{2}\underline{a}_k - \frac{1}{2}\Sigma_{yy}^{-1}(\underline{y} - \underline{\mu}_y)R_{ky}$$
$$- \sum_{\substack{t=1 \\ t \neq k}}^{J} \underline{a}_t(\lambda_t - \lambda_k)^{-1}\Big[-(\lambda_t^{1/2}R_{tx} - R_{ty})(\lambda_k^{1/2}R_{kx} - R_{ky}) +$$
$$(1 - \frac{1}{2}(\lambda_k + \lambda_t))R_{ky}R_{ty}\Big] \quad . \qquad (5.2.9)$$

The influence function $TIC(\underline{z}, \underline{b}_k)$ can be obtained using the relationship in (5.1.4) and applying the product rule for influence discussed in Chapter 1.

5.2.3. Specialisation to the Squared Multiple Correlation Coefficient

We can derive the theoretical influence function of P^2, given by expression (2.4.8), as a special case of (5.2.6) when we take $p_1 = 1$, i.e. we have only one $\underline{y}$ variable. Since $p_1 = 1$ there is only one canonical correlation $\lambda_1^{1/2}$ from (5.1.1), and this is the maximum correlation between $\underline{y}$ and a linear combination of the X variables which is the definition for the multiple correlation coefficient P. We obtain

$$P = \frac{(\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy})^{1/2}}{\sigma_{yy}^{1/2}}$$

(see expression (2.4.3)) from (5.1.1) by taking

$$a_1 = \frac{1}{\sigma_{yy}^{1/2}}$$
$$\underline{b}_1 = \frac{\Sigma_{xx}^{-1}\Sigma_{xy}}{(\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy})^{1/2}} = \frac{\beta}{(\underline{\beta}'\Sigma_{xx}\underline{\beta})^{1/2}} \quad .$$

The forms for the scalar a_1 and $\underline{b}_1$ occur due to normalisations imposed in canonical analysis i.e.

$$\underline{a}_1'\Sigma_{yy}\underline{a}_1 = \underline{b}_1'\Sigma_{xx}\underline{b}_1 = 1 \quad .$$

Substituting

$$\lambda_1^{1/2} = P$$
$$R_{1y} = (y - \mu_y)a_1 = \frac{y - \mu_y}{\sigma_{yy}^{1/2}}$$
$$R_{1x} = (\underline{x} - \underline{\mu}_x)'\underline{b}_1 = \frac{(\underline{x} - \underline{\mu}_x)'\underline{\beta}}{P\sigma_{yy}^{1/2}}$$

into (5.2.6), and letting $\zeta = (\underline{x} - \underline{\mu}_x)'\underline{\beta}$, we arrive at

$$TIC(\underline{z}, \lambda_1) = TIC(\underline{z}, P^2) = \frac{1}{\sigma_{yy}}\left[-\zeta^2 + 2(y - \mu_y)\zeta - P^2(y - \mu_y)^2\right] \quad .$$

This expression is the same as (2.4.6) which then leads to result (2.4.8).

5.3. Influence Functions in Correspondence Analysis When we Add a Single Observation to the (i j)th Cell

As discussed in § 5.1.4, the influence functions for adding to the (i j)th cell of a contingency table in correspondence analysis can be derived as a special form of the canonical correlation functions. Adding 1 to a cell is the same as adding an observation $\underline{z}' = (\underline{y}' \; \underline{x}')$ of dummy variables in canonical correlation analysis. However, we do need to note two differences in the proofs.

First, we do not wish to centre our variables due to the singularity problems discussed in § 5.1.2. This does not noticeably affect our working as everything has a similar form.

$$\Sigma_{xx} = \mathrm{E}\left((\underline{x} - \underline{\mu}_x)(\underline{x} - \underline{\mu}_x)'\right) = \int (\underline{x} - \underline{\mu}(F))(\underline{x} - \underline{\mu}(F))' \, \mathrm{d}F$$

taking $\tilde{F} = (1 - \epsilon)F + \epsilon\delta_{\underline{x}}$ (see Chapter 1), then

$$\begin{aligned} \tilde{\Sigma}_{xx} &= \int (\underline{x} - \underline{\mu}(\tilde{F}))(\underline{x} - \underline{\mu}(\tilde{F}))' \, \mathrm{d}\tilde{F} \\ &= (1 - \epsilon)\Sigma_{xx} + \epsilon(\underline{x} - \underline{\mu}_x)(\underline{x} - \underline{\mu}_x)' + o(\epsilon^2) \quad . \end{aligned}$$

Similarly if

$$V_{xx} = \mathrm{E}(\underline{x}\underline{x}') = \int \underline{x}\underline{x}' \, \mathrm{d}F$$

then

$$\begin{aligned} \tilde{V}_{xx} &= \int \underline{x}\underline{x}' \, \mathrm{d}\tilde{F} \\ &= (1 - \epsilon)V_{xx} + \epsilon\underline{x}\underline{x}' \end{aligned}$$

(we have no second order term when we do not centre). So our influence expressions have the same form but there is no centering.

Second, we normalise our vectors differently. In canonical analysis we let

$$\underline{a}_k' \Sigma_{yy} \underline{a}_k = \underline{b}_k' \Sigma_{xx} \underline{b}_k = 1 \quad ,$$

but in correspondence analysis we let

$$\underline{g}_k' D_c \underline{g}_k = \underline{f}_k' D_r \underline{f}_k = \lambda_k \quad .$$

The only consequence of this is that if $V_{yy}^{-1} = CC'$, then we take $\underline{g}_k = \lambda_k^{1/2} C \underline{\alpha}_k$,

whereas in canonical correlation analysis we take $\underline{a}_k = C\underline{\alpha}_k$ where $\Sigma_{yy}^{-1} = CC'$. The canonical scores were defined as

$$R_{ky} = (\underline{y} - \underline{\mu}_y)'\underline{a}_k \quad , \quad R_{kx} = (\underline{x} - \underline{\mu}_x)'\underline{b}_k \quad ,$$

so that our scores in correspondence analysis would just be

$$g_{kj} = \underline{y}'\underline{g}_k \quad , \quad f_{ki} = \underline{x}'\underline{f}_k$$

since $\underline{z}' = (\underline{y}', \underline{x}') = (00..010..000..010..00)$, where the "1"s are in the jth and ith rows of $\underline{y}$ and $\underline{x}$ respectively, and because $\underline{g}_k = \lambda_k^{1/2}\underline{a}_k$ (where $\underline{a}_k$ is based on the uncentred model), we replace R_{ky} in (5.2.6) by $g_{kj}/\lambda_k^{1/2}$ to give,

$$TIC(\underline{z}, \lambda_k) = -f_{ki}^2 + \frac{2}{\lambda_k^{1/2}} f_{ki} g_{kj} - g_{kj}^2 \qquad (5.3.1)$$

As we have not centred our variables, i.e. we have P rather than $P - \underline{r}\underline{c}'$ in correspondence analysis, this means that we have the trivial dimension in the original and perturbed analyses. Thus,

$$\lambda_1 = \tilde{\lambda}_1 = g_{1j} = \tilde{g}_{1j} = f_{1i} = \tilde{f}_{1i} = 1$$

and, substituting these into (5.3.1), we obtain

$$TIC(\underline{z}, \lambda_1) = 0$$

as required.

Noting $\underline{g}_k = \lambda_k^{1/2}\underline{a}_k$, we have,

$$\begin{aligned} TIC(\underline{z}, \underline{g}_k) &= TIC(\underline{z}, \lambda_k^{1/2})\underline{a}_k + \lambda_k^{1/2} TIC(\underline{z}, \underline{a}_k) \\ &= TIC(\underline{z}, \lambda_k^{1/2})\frac{\underline{g}_k}{\lambda_k^{1/2}} + \lambda_k^{1/2} TIC(\underline{z}, \underline{a}_k) \quad , \end{aligned}$$

Using (5.2.7), (5.2.9) and (5.3.1) and substituting the correspondence analysis forms for R_{ky}, R_{kx} and $V_{YY}^{-1} = D_c^{-1}$ for Σ_{yy}^{-1} gives,

$$\begin{aligned} TIC(\underline{z}, \underline{g}_k) = \frac{1}{2\lambda_k} \Bigg[-f_{ki}^2 + \frac{2}{\lambda_k^{1/2}} f_{ki} g_{kj} - g_{kj}^2 \Bigg] \underline{g}_k + \frac{1}{2}\underline{g}_k - \frac{1}{2} D_c^{-1} \underline{y} g_{kj} \\ - \lambda_k^{1/2} \sum_{\substack{t=1 \\ t \neq k}}^{J} \frac{g_t}{\lambda_t^{1/2}} (\lambda_t - \lambda_k)^{-1} \Bigg[- \left(f_{ti} - \frac{g_{tj}}{\lambda_t^{1/2}} \right) \left(f_{ki} - \frac{g_{kj}}{\lambda_k^{1/2}} \right) \\ + \left(\frac{1}{\lambda_t^{1/2}\lambda_k^{1/2}} - \frac{1}{2} \left(\frac{\lambda_k^{1/2}}{\lambda_t^{1/2}} + \frac{\lambda_t^{1/2}}{\lambda_k^{1/2}} \right) \right) g_{tj} g_{kj} \Bigg] \end{aligned}$$

$$TIC(\underline{z}, \underline{g}_k) = \frac{1}{2\lambda_k}\left[-f_{ki}^2 + \frac{2}{\lambda_k^{1/2}} f_{ki} g_{kj} - g_{kj}^2\right]\underline{g}_k + \frac{1}{2}\underline{g}_k - \frac{1}{2} D_c^{-1} \underline{y} g_{kj}$$
$$- \sum_{\substack{t=1 \\ t \neq k}}^{J} \frac{g_t}{\lambda_t}(\lambda_t - \lambda_k)^{-1}\left[-\left(\lambda_t^{1/2} f_{ti} - g_{tj}\right)\left(\lambda_k^{1/2} f_{ki} - g_{kj}\right)\right.$$
$$\left. + \left(1 - \frac{1}{2}\left(\lambda_k + \lambda_t\right)\right) g_{tj} g_{kj}\right] \qquad (5.3.2)$$

As for $TIC(\underline{z}, \lambda_k)$, since we have not centred our variables, we can show that $TIC(\underline{z}, \underline{g}_1) = \underline{0}$, for the trivial dimension. Now,

$$TIC(\underline{z}, g_1) = 0 + \frac{1}{2}\underline{1} - \frac{1}{2} D_c^{-1} \underline{y}$$
$$- \sum_{\substack{t=1 \\ t \neq 1}}^{J} \frac{g_t}{\lambda_t}(\lambda_t - 1)^{-1}\left[0 - \frac{1}{2}(\lambda_t - 1) g_{tj}\right]$$
$$= \frac{1}{2}\underline{1} - \frac{1}{2} D_c^{-1} \underline{y} + \frac{1}{2} \sum_{\substack{t=1 \\ t \neq 1}}^{J} \frac{1}{\lambda_t} \underline{g}_t \underline{g}_t' \underline{y}$$
$$= \frac{1}{2}\underline{1} - \frac{1}{2} D_c^{-1} \underline{y} + \frac{1}{2} C \sum_{t=1}^{J} \underline{\alpha}_t \underline{\alpha}_t' C' \underline{y} \ .$$

Now,

$$\sum_{\substack{t=1 \\ t \neq 1}}^{J} \underline{\alpha}_t \underline{\alpha}_t' = \Gamma\Gamma' - \underline{\alpha}_1 \underline{\alpha}_1' = I - \underline{\alpha}_1 \underline{\alpha}_1' \ , \qquad (5.3.3)$$

so that

$$TIC(\underline{z}, \underline{g}_1) = \frac{1}{2}\underline{1} - \frac{1}{2} D_c^{-1} \underline{y} + \frac{1}{2}\left[CC' - C \underline{\alpha}_1 \underline{\alpha}_1' C\right] \underline{y}$$

since $CC' = D_c^{-1}$ and $C \underline{\alpha}_1 = \underline{a}_1 = \frac{g_1}{\lambda_1^{1/2}} = \underline{1}$.

$$= \frac{1}{2}\underline{1} - \frac{1}{2} \underline{1}\underline{1}' \underline{y}$$
$$= 0$$

because $\underline{y}' = (0\,0\,..\,0\,1\,0\,..\,0\,0)$, the "1" being in the jth row of the vector The influence function for $\underline{f}_k$ can be obtained using the relationship (5.1.9).

5.4. Influence Functions for Adding a Row to a Contingency Table

5.4.1. Influence Function for the Eigenvalues

When we add a row to the contingency table, we are actually adding m observations such that our matrices of dummy variables for Y and X become

$\tilde{Y}$	$\tilde{X}$
column entries	row entries
1.....J	1.....I I+1
$Y_{(n\_ J)}$	$X_{(n\_ I)}$ $\underline{0}_n$
$Y^*_{(m\ J)}$	$0_{(m\ I)}$ $\underline{1}_m$

Using (5.1.12), we have

$$\tilde{D}_r = \frac{1}{n_{..}+m}\tilde{X}'\tilde{X} = \frac{1}{n_{..}+m}\begin{pmatrix} n_{..}D_r & \underline{0}_J \\ \underline{0}_J' & m \end{pmatrix}$$

$$= \begin{pmatrix} (1-\epsilon m)D_r & \underline{0}_J \\ \underline{0}_J' & \epsilon m \end{pmatrix},$$

where we let $\epsilon = 1/(n_{..}+m)$ and, since $\tilde{D}_r$ is diagonal, to $o(\epsilon)$ we have

$$\tilde{D}_r^{-1} = \begin{pmatrix} (1+\epsilon m)D_r^{-1} & \underline{0}_J \\ \underline{0}_J' & 1/\epsilon m \end{pmatrix} \tag{5.4.1}$$

$$\tilde{D}_c = \frac{1}{n_{..}+m}\tilde{Y}'\tilde{Y} = \frac{1}{n_{..}+m}\left(n_{..}D_c + Y^{*\prime}Y^*\right)$$

$$= (1-\epsilon m)D_c + \epsilon Y^{*\prime}Y^* \quad .$$

Since Y^* is a matrix, we need to use the following matrix result to obtain $\tilde{D}_c^{-1}$ (see Mardia *et al.*, 1979, p459).

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \quad . \tag{5.4.2}$$

In the above we take $A = (1-\epsilon m)D_c$, $B = \sqrt{\epsilon}Y^{*\prime}$, $C = I$ and $D = \sqrt{\epsilon}Y^*$.

Omitting the details (we actually need to use (5.4.2) twice), we obtain to $o(\epsilon)$,

$$\tilde{D}_c^{-1} = (1 + \epsilon m)D_c^{-1} - \epsilon D_c^{-1}Y^{*\prime}Y^{*}D_c^{-1} \quad .$$

provided m is small compared to $n_{..}$.
Let $D_{r^*} = Y^{*\prime}Y^{*}$ be the diagonal matrix of the added row $\underline{r}^{*}$ (since each row of Y^{*} only contains one entry of 1 for the column that the observation falls in) then, to $o(\epsilon)$,

$$\tilde{D}_c^{-1} = (1 + \epsilon m)D_c^{-1} - \epsilon D_c^{-1}D_{r^*}D_c^{-1} \quad . \tag{5.4.3}$$

We wish to find the influence function for the eigenvalues from the symmetric matrix $C'P'D_r^{-1}PC$, which are the same as those from the matrix $D_c^{-1}P'D_r^{-1}P$, where $D_c^{-1} = CC'$. Taking (to $o(\epsilon)$),

$$\tilde{C} = C + \frac{\epsilon}{2}\left(m I - D_c^{-1}D_{r^*}\right)C \tag{5.4.4}$$

(which is diagonal), we see that $\tilde{C}\tilde{C}' = \tilde{D}_c^{-1}$.

$$\tilde{P} = \frac{1}{n_{..} + m}\tilde{X}'\tilde{Y} = \begin{pmatrix}(1 - \epsilon m)P \\ \epsilon \underline{1}'Y^{*}\end{pmatrix} = \begin{pmatrix}(1 - \epsilon m)P \\ \epsilon \underline{r}^{*\prime}\end{pmatrix} , \tag{5.4.5}$$

where $\underline{r}^{*\prime}$ is the added row.

Using (5.4.5) and (5.4.1),

$$\tilde{P}'\tilde{D}_r^{-1}\tilde{P} = \begin{pmatrix}(1 - \epsilon m)P' & \epsilon \underline{r}^{*}\end{pmatrix}\begin{pmatrix}D_r^{-1}P \\ \underline{r}^{*\prime}/m\end{pmatrix} + o(\epsilon^2) \tag{5.4.6}$$

$$= (1 - \epsilon m)P'D_r^{-1}P + \frac{\epsilon}{m}\underline{r}^{*}\underline{r}^{*\prime} + o(\epsilon^2) \quad . \tag{5.4.7}$$

We can use the same formulae to obtain the influence functions for the eigenvalues and eigenvectors as used previously for the theoretical influence functions as the perturbed expressions above are similar. The only difference is it is an asymptotic result where we consider expansions to $o(\epsilon)$ where $\epsilon = 1/(n_{..} + m)$. Since, this is an asymptotic derivation we will refer to the influence functions immediately as the empirical. Using (5.4.7) and (5.4.4) by the product rule for influence,

$$EIC(\underline{r}^*, C'P'D_r^{-1}PC) = \frac{1}{2}C'\left(m\,I - D_r.D_c^{-1}\right)P'D_r^{-1}PC$$
$$+C'\left(-mP'D_r^{-1}P + \underline{r}^*\underline{r}^{*\prime}/m\right)C$$
$$+\frac{1}{2}C'P'D_r^{-1}P\left(m\,I - D_c^{-1}D_r.\right)C$$

$$= -\frac{1}{2}C'\left(D_r.D_c^{-1}P'D_r^{-1}P + P'D_r^{-1}PD_c^{-1}D_r.\right)C$$
$$+C'\left(\underline{r}^*\underline{r}^{*\prime}/m\right)C \quad . \tag{5.4.8}$$

The influence function using (3.5.3) (which would not be affected by this being derived asymptotically) and taking $\underline{g}_k = \lambda_k^{1/2}C\,\underline{\alpha}_k$ gives

$$EIC(\underline{r}^*, \lambda_k) = -\frac{1}{2\lambda_k}\underline{g}_k'\left(D_r.D_c^{-1}P'D_r^{-1}P + P'D_r^{-1}PD_c^{-1}D_r.\right)\underline{g}_k$$
$$+\frac{1}{m\lambda_k}\underline{g}_k'\underline{r}^*\underline{r}^{*\prime}\underline{g}_k \quad .$$

Using (5.1.10)

$$= -\underline{g}_k'D_r.\underline{g}_k + \frac{1}{m\lambda_k}\underline{g}_k'\underline{r}^*\underline{r}^{*\prime}\underline{g}_k \tag{5.4.9}$$

$$= -\sum_{j=1}^{J} r_j^* g_{kj}^2 + \frac{1}{m\lambda_k}\left[\sum_{j=1}^{J} r_j^* g_{kj}\right]^2 \quad . \tag{5.4.10}$$

Alternatively, from the transition formula (5.1.9), the ith row co-ordinate is

$$f_{ki} = \frac{1}{\lambda_k^{1/2}}\frac{1}{m_i}\underline{r}_i'\underline{g}_k$$

where m_i is the ith row sum. We can substitute into (5.4.9)

$$f_{ks} = \frac{1}{m\lambda_k^{1/2}}\underline{r}^{*\prime}\underline{g}_k \quad , \tag{5.4.11}$$

where f_{ks} is the plotting co-ordinate for the extra (supplementary) row with respect to the original $\underline{g}_k$ and λ_k. This is used by Greenacre (1984) to display a supplementary point (hence the subscript s) on the existing plot of rows and columns. As Greenacre notes, the point is not contributing to the display. We shall see later that this representation of the extra point is useful when we look at the influence function for $\underline{f}_k$.

So, letting $W_{kk} = \sum_{j=1}^{J} r_j^* g_{kj}^2$,

$$EIC(\underline{r}^{*}, \lambda_k) = -W_{kk} + mf_{ks}^2 \quad . \tag{5.4.12}$$

Since we are dealing with uncentred matrices, $\underline{g}_1 = \underline{1}_J$ and $\lambda_1 = 1$. Substituting these into (5.4.10) and noting that $\sum_{j=1}^{J} r_j^{*} = m$, the row sum, then

$$EIC(\underline{r}^{*}, \lambda_1) = 0$$

as required.

5.4.2. Influence Functions for the Co-ordinates

As with the influence function for the eigenvalues, result (3.5.9) will not be affected by our asymptotic approach to influence. Therefore,

$$EIC(\underline{r}^{*}, \underline{\alpha}_k) = -\sum_{\substack{t=1 \\ t \neq k}}^{J} \underline{\alpha}_t(\lambda_t - \lambda_k)^{-1}\underline{\alpha}_t' EIC(\underline{r}^{*}, C'P'D_r^{-1}PC)\underline{\alpha}_k \tag{5.4.13}$$

and since $\underline{g}_k = C\underline{\alpha}_k\lambda_k^{1/2}$, then

$$\begin{aligned} EIC(\underline{r}^{*}, \underline{g}_k) = EIC(\underline{r}^{*}, C)\underline{\alpha}_k\lambda_k^{1/2} + CEIC(\underline{r}^{*}, \underline{\alpha}_k)\lambda_k^{1/2} \\ + C\underline{\alpha}_k EIC(\underline{r}^{*}, \lambda_k^{1/2}) \end{aligned} \tag{5.4.14}$$

which occurs from considering the perturbed forms for C, $\underline{\alpha}_k$ and $\lambda_k^{1/2}$, or directly from the product rule.

Using (5.4.8),

$$\begin{aligned} \underline{\alpha}_t' EIC(\underline{r}^{*}, C'P'D_r^{-1}PC)\underline{\alpha}_k = -\frac{1}{2\lambda_t^{1/2}\lambda_k^{1/2}}\underline{g}_t' \left[D_r.D_c^{-1}PD_r^{-1}P + P'D_r^{-1}PD_c^{-1}D_r.\right]\underline{g}_k \\ + \frac{1}{m\lambda_t^{1/2}\lambda_k^{1/2}}\underline{g}_t'\underline{r}^{*}\underline{r}^{*}'\underline{g}_k \end{aligned}$$

Using (5.1.10),

$$\begin{aligned} \underline{\alpha}_t' EIC(\underline{r}^{*}, C'P'D_r^{-1}PC)\underline{\alpha}_k = -\frac{1}{2}\left(\frac{\lambda_k^{1/2}}{\lambda_t^{1/2}} + \frac{\lambda_t^{1/2}}{\lambda_k^{1/2}}\right)\underline{g}_t'D_r.\underline{g}_k \\ + \frac{1}{m\lambda_t^{1/2}\lambda_k^{1/2}}\underline{g}_t'\underline{r}^{*}\underline{r}^{*}'\underline{g}_k \quad . \end{aligned}$$

Using this and (5.4.11), (5.4.13) becomes

$$EIC(\underline{r}^{*}, \underline{\alpha}_k) = -\sum_{\substack{t=1 \\ t \neq k}}^{J} \underline{\alpha}_t (\lambda_t - \lambda_k)^{-1} \left(-\frac{1}{2\lambda_t^{1/2}\lambda_k^{1/2}} (\lambda_k + \lambda_t) W_{kt} + m f_{ts} f_{ks} \right) \quad , \tag{5.4.15}$$

where $W_{kt} = \sum_{j=1}^{J} r_j^{*} g_{tj} g_{kj}$. Using (5.4.15), (5.4.4), (5.2.7) and (5.4.12), (5.4.14) becomes,

$$\begin{aligned} EIC(\underline{r}^{*}, \underline{g}_k) = -\sum_{\substack{t=1 \\ t \neq k}}^{J} \frac{\lambda_k^{1/2}}{\lambda_t^{1/2}} \underline{g}_t (\lambda_t - \lambda_k)^{-1} \left(-\frac{1}{2\lambda_t^{1/2}\lambda_k^{1/2}} (\lambda_k + \lambda_t) W_{kt} + m f_{ts} f_{ks} \right) \\ + \frac{1}{2} m \underline{g}_k - \frac{1}{2} D_c^{-1} D_{r^{*}} \underline{g}_k \\ + \frac{1}{2\lambda_k} \underline{g}_k \left(-W_{kk} + m f_{ks}^2 \right) \end{aligned} \tag{5.4.16}$$

Again, we have worked with the uncentred matrix P, so that $EIC(\underline{r}^{*}, \underline{g}_1)$ should be zero. Substituting $\lambda_1 = 1$, $\underline{f}_1 = \underline{1}$ and $\underline{g}_1 = \underline{1}$, we obtain, in a similar way to adding to a cell,

$$EIC(\underline{r}^{*}, \underline{g}_1) = \frac{1}{2} \sum_{\substack{t=1 \\ t \neq 1}}^{J} \frac{1}{\lambda_t} \underline{g}_t \underline{g}_t' \underline{r}^{*} + \frac{1}{2} m \underline{1} - \frac{1}{2} D_c^{-1} \underline{r}^{*} + 0 \quad ,$$

because $W_{1t} = \sum_{j=1}^{J} r_j^{*} g_{tj} = \underline{g}_t' \underline{r}^{*}$ and,

$$m f_{ts} f_{1s} = m f_{ts} = \frac{1}{\lambda_t^{1/2}} \underline{g}_t' \underline{r}^{*} \quad .$$

Therefore,

$$EIC(\underline{r}^{*}, \underline{g}_1) = \frac{1}{2} C \sum_{t=2}^{J} \underline{\alpha}_t \underline{\alpha}_t' C' \underline{r}^{*} + \frac{1}{2} m \underline{1} - \frac{1}{2} D_c^{-1} \underline{r}^{*} \quad .$$

Using (5.3.3),

$$\begin{aligned} EIC(\underline{r}^{*}, \underline{g}_1) &= \frac{1}{2} C \left(I - \underline{\alpha}_1 \underline{\alpha}_1' \right) C' \underline{r}^{*} + \frac{1}{2} m \underline{1} - \frac{1}{2} D_c^{-1} \underline{r}^{*} \\ &= \frac{1}{2} D_c^{-1} \underline{r}^{*} - \frac{1}{2} \underline{a}_1 \underline{a}_1' \underline{r}^{*} + \frac{1}{2} m \underline{1} - \frac{1}{2} D_c^{-1} \underline{r}^{*} \end{aligned}$$

but $\underline{a}_1 = \underline{1}$ and $\underline{1}' \underline{r}^{*} = m$, the row sum. Therefore,

$$EIC(\underline{r}^{*}, \underline{g}_1) = 0$$

as required.

We obtain the influence function for $\underline{f}_k$ from the transition formula (5.1.9), which is

$$\underline{f}_k = \frac{1}{\lambda_k^{1/2}} D_r^{-1} P \underline{g}_k \quad . \tag{5.4.17}$$

However, $\frac{1}{\lambda_k^{1/2}} \underline{g}_k = \underline{a}_k$ and

$$EIC(\underline{r}^{*}, \underline{a}_k) = EIC(\underline{r}^{*}, C)\underline{\alpha}_k + CEIC(\underline{r}^{*}, \underline{\alpha}_k) \quad .$$

Because of (5.4.14), we can obtain this influence function by dividing (5.4.16) by $\lambda_k^{1/2}$ and, omitting the last term, this gives

$$EIC(\underline{r}^{*}, \underline{a}_k) = -\sum_{\substack{t=1 \\ t \neq k}}^{J} \frac{1}{\lambda_t^{1/2}} \underline{g}_t (\lambda_t - \lambda_k)^{-1} \left(-\frac{1}{2\lambda_t^{1/2}\lambda_k^{1/2}} (\lambda_k + \lambda_t) W_{kt} + m f_{tt} f_{kt} \right)$$

$$+ \frac{1}{2} m \underline{a}_k + \frac{1}{2} D_c^{-1} D_r \cdot \underline{a}_k \quad . \tag{5.4.18}$$

From (5.4.6),

$$\tilde{D}_r^{-1} \tilde{P} = \begin{pmatrix} D_r^{-1} P \\ \frac{1}{m} \underline{r}^{*}, \end{pmatrix} + o(\epsilon^2) \quad ,$$

i.e. it does not involve a first order term, so that

$$\tilde{\underline{f}}_k = \tilde{D}_r^{-1} \tilde{P} \tilde{\underline{a}}_k = \begin{pmatrix} D_r^{-1} P \\ \frac{1}{m} \underline{r}^{*}, \end{pmatrix} \left(\underline{a}_k + \epsilon EIC(\underline{r}^{*}, \underline{a}_k) \right) \quad . \tag{5.4.19}$$

Therefore, $\tilde{\underline{f}}_k' = (\underline{c}_1 \;\; c_2)$ where,

$$\underline{c}_1 = \underline{f}_k + \epsilon \left[-\sum_{\substack{t=1 \\ t \neq k}} \underline{f}_t (\lambda_t - \lambda_k)^{-1} \left(\frac{-(\lambda_t + \lambda_k)}{2\lambda_k^{1/2}\lambda_t^{1/2}} W_{kt} + m f_{tt} f_{kt} \right) \right.$$

$$\left. + \frac{1}{2} m \underline{f}_k - \frac{1}{2\lambda_k^{1/2}} D_r^{-1} P D_c^{-1} D_r \cdot \underline{g}_k \right]$$

$$c_2 = \frac{\underline{r}^{*\prime} g_k}{m\lambda_k^{1/2}} + \epsilon \left[- \sum_{t \neq k} \frac{\underline{r}^{*\prime} g_t}{m\lambda_t^{1/2}} (\lambda_t - \lambda_k)^{-1} \left(\frac{-(\lambda_t + \lambda_k)}{2\lambda_k^{1/2}\lambda_t^{1/2}} W_{kt} + m f_{ts} f_{ks} \right) \right.$$

$$\left. + \frac{\underline{r}^{*\prime} g_k}{2\lambda_k^{1/2}} - \frac{r^{*\prime}}{2m\lambda_k^{1/2}} D_c^{-1} D_r \cdot \underline{g}_k \right] \qquad (5.4.20)$$

Using (5.4.11) we can re-express c_2 as,

$$f_{k\,(I+1)} = f_{ks} + \epsilon \left[- \sum_{\substack{t=1 \\ t \neq k}} f_{ts} (\lambda_t - \lambda_k)^{-1} \left(- \frac{(\lambda_t + \lambda_k)}{2\lambda_k^{1/2}\lambda_t^{1/2}} W_{kt} + m f_{ts} f_{ks} \right) \right.$$

$$\left. + \frac{1}{2} m f_{ks} - \frac{r^*}{2m\lambda_k^{1/2}} D_c^{-1} D_r \cdot \underline{g}_k \right] .$$

The perturbed $\tilde{\underline{f}}_k$ vector is different to all the other eigenvector perturbations we have considered as it contains one more element than the original $\underline{f}_k$. Thus, we find that the first I rows of $\tilde{\underline{f}}_k$ can be written in the form

$$\tilde{\underline{f}}_k = \underline{f}_k + \epsilon EIC(\underline{r}^*, \underline{f}_k) + o(\epsilon^2)$$

and the term for the extra $(I+1)$th row is

$$\tilde{f}_{k\,(I+1)} = f_{ks} + \epsilon f_{k\,(I+1)}^{(1)} + o(\epsilon^2) \qquad (5.4.21)$$

i.e. it is the term for displaying the extra row on the original plot, as discussed earlier, plus higher order terms. We have not written $EIC(\underline{r}^*, f_{k\,(I+1)})$ since it was not in the original problem. However, the term in ϵ, $f_{k\,(I+1)}^{(1)}$, is similar to that for the other co-ordinates. Using f_{ks}, for displaying supplementary points on the original correspondence plot, is thus justified by the form for $\tilde{f}_{k\,(I+1)}$.

5.5. Adding m Identical Observations in Multiple Correspondence Analysis

5.5.1. Influence Function for the Eigenvalues

As discussed in § 5.1.3, we have a Q-variate indicator matrix Z such that the symmetric Burt matrix, which contains all the two-way contingency tables, is $B = Z'Z$. If there are J_q categories, $q = 1, 2, ..., Q$, on each variable then we will add m identical observations with the categories L_q, $q = 1, 2, ..., Q$. Then,

$$\tilde{B} = B + W \tag{5.5.1}$$

where W is a symmetric matrix consisting of m in the (L_i, L_j)th positions, $i, j = 1, 2, ..., Q$, and zeroes elsewhere. Thus, m is added to Q cells in the Q rows, corresponding to the L_q categories, of the Burt matrix, so that the grand total of $\tilde{B}$ is $\tilde{n}_{..} = n_{..} + Q^2 m$.

As $P = B / n_{..}$ is a symmetric matrix, the correspondence analysis of the Burt matrix results from finding the eigenvalues and eigenvectors of

$$(D_r^{-1} P D_r^{-1} P - \lambda_k I)\underline{h}_k = 0 \tag{5.5.2}$$

with normalisation $\underline{h}_k' D_r \underline{h}_k = \lambda_k$. However, the eigenvectors of $D_r^{-1}P$ are the same as those from $D_r^{-1} P D_r^{-1} P$ and the eigenvalues are square-rooted so that

$$(D_r^{-1} P - \lambda_k^{1/2} I)\underline{h}_k = 0 \quad . \tag{5.5.3}$$

From (5.5.1),

$$\tilde{P} = \frac{\tilde{B}}{\tilde{n}_{..}} = \frac{B + W}{n_{..} + Q^2 m} = \frac{n_{..} P}{n_{..} + Q^2 m} + \frac{W}{n_{..} + Q^2 m} \quad .$$

Letting $\epsilon = 1 / (n_{..} + Q^2 m)$, then

$$\tilde{P} = (1 - \epsilon Q^2 m) P + \epsilon W \quad . \tag{5.5.4}$$

If $\underline{s} = W' \underline{1}$, then $\underline{s}$ consists of entries, Qm, corresponding to the L_qth categories, and zeroes elsewhere. These non-zero entries occur in the V_q positions where

$$V_q = \sum_{j=1}^{q-1} J_j + L_q \quad . \tag{5.5.5}$$

From (5.5.4) we have,

$$\tilde{D}_r = (1 - \epsilon Q^2 m) D_r + \epsilon D_s$$

where D_s is a diagonal matrix of $\underline{s}$. Using (5.4.2) we obtain, in a similar manner to (5.4.3),

$$\tilde{D}_r^{-1} = (1 + \epsilon Q^2 m) D_r^{-1} - \epsilon D_r^{-1} D_s D_r^{-1}$$

and letting $D_r^{-1} = CC'$, similarly to (5.4.4) we have

$$\tilde{C} = C + \frac{\epsilon}{2} \left[m Q^2 I - D_r^{-1} D_s \right] C \quad . \tag{5.5.6}$$

The asymptotic form of (3.5.3) is,

$$\begin{aligned} EIC(m\underline{z}, \lambda_k^{1/2}) &= \underline{\alpha}_k' EIC(m\underline{z}, C'PC)\underline{\alpha}_k \\ &= \underline{\alpha}_k' \left[EIC(m\underline{z}, C')PC + C'EIC(m\underline{z}, P)C + C'P\, EIC(m\underline{z}, C) \right] \underline{\alpha}_k \end{aligned}$$

from multiplying $\tilde{C}'\tilde{P}\tilde{C}$ together, or from the product rule. Simple algebra (similar to previous sections), using (5.5.4) and (5.5.6), gives

$$EIC(m\underline{z}, \lambda_k^{1/2}) = \underline{\alpha}_k' C' \left[- D_s D_r^{-1} P + W \right] C \underline{\alpha}_k \quad .$$

We wish to normalise our vectors so that $\underline{h}_k' D_r \underline{h}_k = \lambda_k$, so we take $\underline{h}_k = C \underline{\alpha}_k \lambda_k^{1/2}$. Therefore, using (5.5.3),

$$EIC(m\underline{z}, \lambda_k^{1/2}) = - \frac{1}{\lambda_k^{1/2}} \underline{h}_k' D_s \underline{h}_k + \frac{1}{\lambda_k} \underline{h}_k' W \underline{h}_k \quad .$$

Since,

$$\tilde{\lambda}_k = \left[\lambda_k^{1/2} + \epsilon EIC(m\underline{z}, \lambda_k^{1/2}) + o(\epsilon^2) \right]^2$$

then

$$\begin{aligned} EIC(m\underline{z}\, \lambda_k) &= 2\lambda_k^{1/2} EIC(m\underline{z}, \lambda_k^{1/2}) \\ &= -2\underline{h}_k' D_s \underline{h}_k + \frac{2}{\lambda_k^{1/2}} \underline{h}_k' W \underline{h}_k \quad , \end{aligned} \tag{5.5.7}$$

$$\underline{h}_k' D_s \underline{h}_k = Qm \sum_{q=1}^{Q} h_k^2 v_q$$

(with v_q defined as in (5.5.5)), and

$$\underline{h}_k' W \underline{h}_k = m \sum_{q=1}^{Q} h_k^2 v_q + 2m \sum_{q=1}^{Q} \sum_{\substack{q'=1 \\ q' \neq q}}^{Q} h_k v_q h_k v_{q'} \quad ,$$

so that the influence function only involves the eigenvector co-ordinates for the categories of the m extra observations. Thus from (5.5.7)

$$EIC(m\underline{z}, \lambda_k) = \frac{2m}{\lambda_k^{1/2}}\left[1 - \lambda_k^{1/2}Q\right]\sum_{q=1}^{Q} h_k^2 v_q + \frac{4m}{\lambda_k^{1/2}}\sum_{q=1}^{Q}\sum_{q'>q}^{Q} h_k v_q h_k v_{q'} \quad . \tag{5.5.8}$$

Substituting $\lambda_1 = 1$ and $\underline{h}_1 = \underline{1}$ gives $EIC(m\underline{z}, \lambda_1) = 0$, for the trivial dimension. It can be shown that (5.5.8) specialises to the eigenvalue influence function for adding to a cell of a contingency table, derived in § 5.3, by using the relationship $\lambda_k^C = 4(\sqrt{\lambda_k^B} - 1/2)^2$. As for the two-way results, we see that influence depends only on the co-ordinates corresponding to the categories involved in the m observations. We can see that the affect of adding m, rather than unity, is just to multiply the influence by m (since all m observations are identical).

5.5.2. Influence Function for the Co-ordinates/Eigenvectors

The influence function for the eigenvector from the symmetric matrix $C'PC$, using (3.5.9), is

$$EIC(m\underline{z}, \underline{\alpha}_k) = -\sum_{\substack{t=1 \\ t\neq k}}^{J} \underline{\alpha}_t(\lambda_t^{1/2} - \lambda_k^{1/2})^{-1}\underline{\alpha}_t' EIC(m\underline{z}, C'PC)\underline{\alpha}_k \quad , \tag{5.5.9}$$

where J is the sum of the number of categories on each variable. Using (5.5.4) and (5.5.6),

$$\underline{\alpha}_t' EIC(m\underline{z}, C'PC)\underline{\alpha}_k = \underline{\alpha}_t' C'\left[-\frac{1}{2}D_s D_r^{-1}P + W - \frac{1}{2}PD_r^{-1}D_s\right]C\underline{\alpha}_k \quad .$$

Using (5.5.3) and $\underline{h}_k = C\underline{\alpha}_k\lambda_k^{1/2}$, then

$$\underline{\alpha}_t' EIC(m\underline{z}, C'PC)\underline{\alpha}_k = \frac{1}{\lambda_t^{1/2}\lambda_k^{1/2}}\underline{h}_t'\left[-\frac{1}{2}(\lambda_k^{1/2} + \lambda_t^{1/2})D_s + W\right]\underline{h}_k \quad . \tag{5.5.10}$$

As in (5.4.14),

$$EIC(m_{\underline{z}}, \underline{h}_k) = EIC(m_{\underline{z}}, C)\underline{\alpha}_k \lambda_k^{1/2} + CEIC(m_{\underline{z}}, \underline{\alpha}_k)\lambda_k^{1/2} + C\,\underline{\alpha}_k EIC(m_{\underline{z}}, \lambda_k^{1/2}) \quad .$$

Using (5.5.6) and substituting (5.5.10) into (5.5.9), then

$$\begin{aligned} EIC(m_{\underline{z}}, \underline{h}_k) = \frac{1}{2}(Q^2 ml - D_r^{-1} D_s)\underline{h}_k \\ - \sum_{\substack{t=1 \\ t \neq k}}^{J} \frac{1}{\lambda_t} \underline{h}_t (\lambda_t^{1/2} - \lambda_k^{1/2})^{-1} \underline{h}_t' \left[-\frac{1}{2}(\lambda_k^{1/2} + \lambda_t^{1/2}) D_s + W \right] \underline{h}_k \\ + \frac{1}{\lambda_k^{1/2}} \underline{h}_k EIC(m_{\underline{z}}, \lambda_k^{1/2}) \end{aligned} \qquad (5.5.11)$$

where

$$\underline{h}_t' D_s \underline{h}_k = Qm \sum_{q=1}^{Q} h_{t v_q} h_{k v_q}$$

$$\underline{h}_t' W \underline{h}_k = m \sum_{q=1}^{Q} \sum_{\substack{q=1 \\ q' \neq q}}^{Q} h_{t v_q} h_{k v_{q'}} \quad .$$

In exactly the same way to previous influence procedures, we can show $EIC(m_{\underline{z}}, \underline{h}_1) = 0$. The proof is omitted as the steps are almost identical to earlier work.

Expression 5.5.11 only involves the co-ordinates for the categories from the added points. This is similar to the expression for the changes in the co-ordinates when we add a single observation to a cell.

Chapter 6: Investigation of Influence in Correspondence Analysis by Application to Real Datasets

6.1. Introduction

In this chapter we examine influence in correspondence analysis for the three types of perturbations given in the last chapter. The three influence procedures are,

(1) adding in a single observation so that a cell of a contingency table is incremented by one.

(2) Deletion of a row from a contingency table. (No justification is given here, apart from the good comparisons of empirical and sample, but we find that it is justifiable as in other influence techniques to use the empirical expression for deletion as well as addition).

(3) Adding in a single observation to a multiway table so that Q^2 (where Q is the number of variables) cells of the Burt matrix increase by one.

These influence procedures are examined by application to real contingency tables. We investigate influence by looking at the patterns of sample influence and where possible we relate this back to our empirical expressions. The empirical expressions for the G and F co-ordinates in any of the perturbation schemes are not easy to interpret, but we can often gain insight from the eigenvalue expressions. If not stated otherwise the influence values given in this chapter refer to the sample function. We will refer to the first non-trivial dimension as the first dimension etc.

The first two types of perturbation are considered in detail, as well as the extension of (2) to the deletion of columns. Problems with the rotation and swopping of eigenvectors, as seen in PCA, will be observed in the first few dimensions from correspondence analysis under the perturbations in (2). This

is particularly seen when we delete columns rather than rows. In the latter influence method we point out the similarities with the patterns observed in (1) by considering the analysis for two and three variables.

In § 6.3.6 we consider an adaptation of the empirical influence function for the kth eigenvectors (G and F co-ordinates) that involves just the summation over the dimensions close to the kth dimension. For example, use the two dimensions either side of the one of interest or any dimensions whose eigenvalues are close. This is done to increase the speed of computation of the empirical influences and is justifiable as the terms in $(\lambda_j - \lambda_k)^{-1}$ will usually be largest for the closest eigenvalues.

6.2. Investigation of Influence When Adding in a Single Observation to the Cell of a Contingency Table

6.2.1. Influence for the Eigenvalues (Principal Inertias)

Investigation of such influence, for two contingency tables below, gives us interesting insights into the sensitivity of correspondence analysis. The results from the asymptotic theory for adding to a cell should hold quite well, provided $n_{..}$ is fairly large. The actual dimensions I and J should be less important in determining whether the asymptotic results hold well in practice.

The first contingency table we will consider is given in Table 6.2.1 and is taken from Greenacre (1984,p 259). The table is concerned with the worries of Israeli adults according to where in the world they live and where their father lived. The plot of the first two dimensions is given in Fig. 6.2.1 and the interpretation of the plot is given by Greenacre. The sample size is $n_{..} = 1554$, so we may expect our comparisons of actual sample and estimated change to be good. We will consider influence in the first two dimensions in detail, since we usually hope our analysis results in a useful two dimensional plot, and will

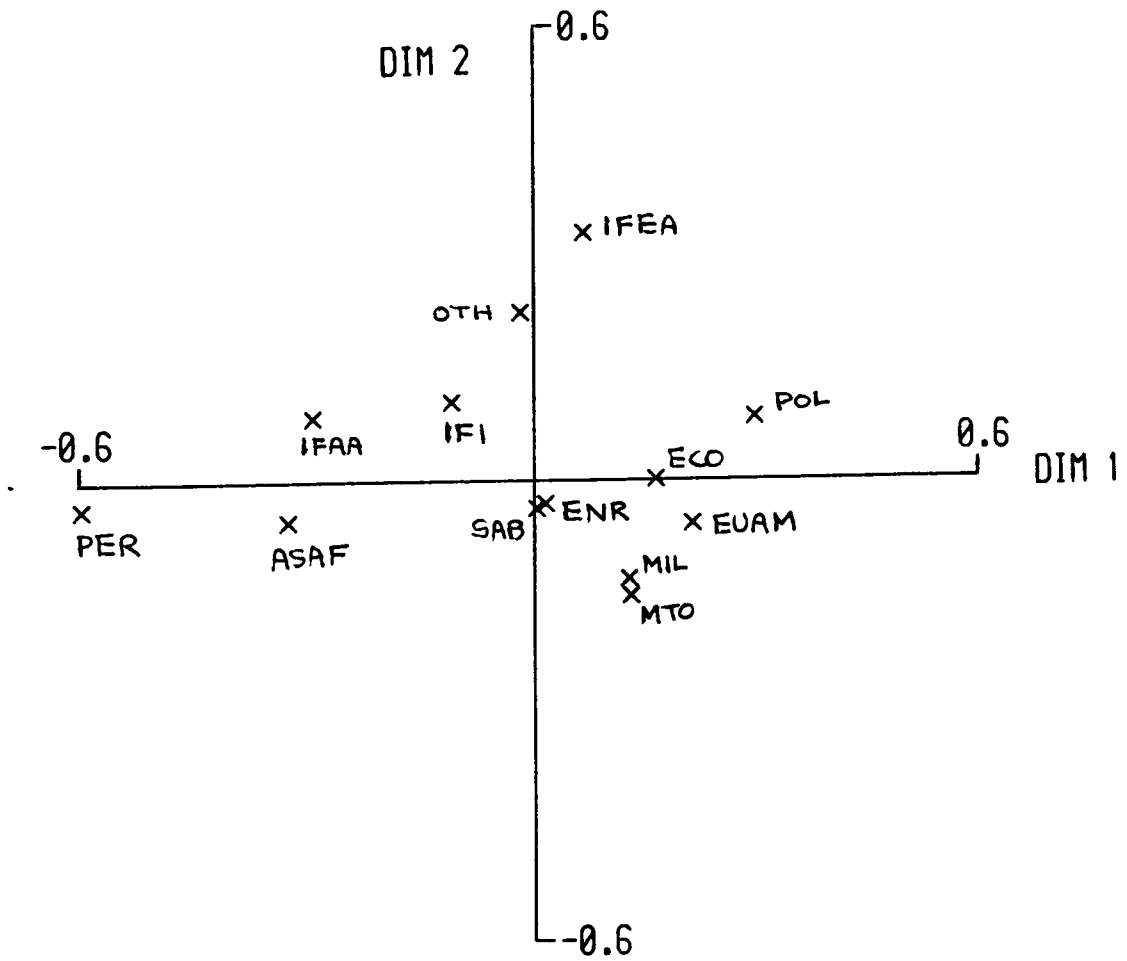
Table 6.2.1
Contingency Table of the Principal Worries of Israeli Adults

Worry	Demography ASAF	EUAM	IFAA	IFEA	IFI	Total
ENR	61	104	8	22	5	200
SAB	70	117	9	24	7	227
MIL	97	218	12	28	14	369
POL	32	118	6	28	7	191
ECO	4	11	1	2	1	19
OTH	81	128	14	52	12	287
MTO	20	42	2	6	0	70
PER	104	48	14	16	9	191
Total	469	786	66	178	55	1554.0

Abbreviations

ENR	Enlisted relative	SAB	Sabotage
MIL	Military situation	POL	Political situation
ECO	Economic situation	OTH	Other
MTO	More than one worry	PER	Personnal economics
ASAF	Asia/Africa	EUAM	Europe/America
IFAA	Israel,father Asia/Africa	IFEA	Israel,father Europe/America
IFI	Israel,father Israel	-	-

Figure 6.2.1 Plot of the First Two Dimensions from the Correspondence Analysis of the Dataset on Worries of Israeli Adults.



briefly outline the influence in later dimensions. The most influential observations/cells, ranked by the sample influence, and corresponding empirical estimated change, are given in Table 6.2.2. The change is recorded as the percentage change in the eigenvalue. The most influential observation is recorded as the cell that causes the largest change in the eigenvalues when its value is incremented by one.

Table 6.2.2.

Most influential Cells (when one is added to them) for the First Two Eigenvalues From Israeli Dataset

Cell	$\hat{\lambda}_1$ Actual Change	Estimated Change	Cell	$\hat{\lambda}_2$ Actual Change	Estimated Change
8,2	-1.5%	-1.6%	6,4	4.4%	4.3%
8,1	1.2%	1.2%	7,4	3.5%	4.0%
8,3	1.2%	0.9%	3,4	3.3%	3.5%

These changes are small but we are considering a small perturbation. The comparisons above are good and so we can interpret our asymptotic expression (5.3.1), to give us a clear picture of influence in practice. Expression (5.3.1) can be re-expressed as,

$$TIC(\underline{x}, \lambda_k) = -(g_{kj} - f_{ki})^2 + 2g_{kj}f_{ki}\left(\frac{1}{\lambda_k^{1/2}} - 1\right) \qquad (6.2.1)$$

$$\text{where } \frac{1}{\lambda_k^{1/2}} > 0 \text{ as } 0 < \lambda_k^{1/2} < 1 \quad .$$

Since we are considering the sample function for adding in points (so the cell entries are incremented by 1) the original parameter is subtracted from the perturbed. A positive influence thus corresponds to an increase in the parameter.

From (6.2.1) we obtain a large negative influence, i.e. λ_k decreases, when the co-ordinates g_{kj} and f_{ki} are far apart and have opposing signs. Adding in one to the $(i,j)^{th}$ cell increases the association between the corresponding categories. Thus, if g_{kj} and f_{ki} are at the opposite extremes of the dimension they should move inwards towards each other, and so the

variance decreases. The eigenvalue increases if g_{kj} and f_{ki} are close and at the extreme of the dimension. This occurs as the row and column are already highly associated with each other, and become more so when one is added to the cell entry. Thus, the two are relatively less associated with the rows and columns and so move further out together. The variance thus increases.

The comparisons between sample and empirical were good in this dataset, so we find the above patterns are exhibited for this contingency table. The influence of cell (8,2) on $\hat{\lambda}_1$ is negative and from Fig. 6.2.1 the co-ordinates for 'Personal economics' and 'Europe and America' are far apart at opposing sides of the axis. Conversely, the co-ordinates for 'Personal economics' and 'Asia/Africa' are close together and both large, and the affect of cell (8,1) on $\hat{\lambda}_1$ is positive. The same pattern occurs for the second axis with, for example, the affect of cell (6,4) being positive and the co-ordinates for the categories 'Other' and 'Israel father/Europe America' being close together and large in the second dimension.

The same pattern occurs for the second contingency table from Greenacre (1984, p 55), where $n_{..} = 193$, which is smaller than that for the above dataset. The contingency Table is given in Table (6.2.3) and the correspondence plot for the first two dimensions in Fig. 6.2.2. This is a set of artificial data, relating to the smoking habits of employees. The three most influential cells on the first two eigenvalues and the sign of the influence are,

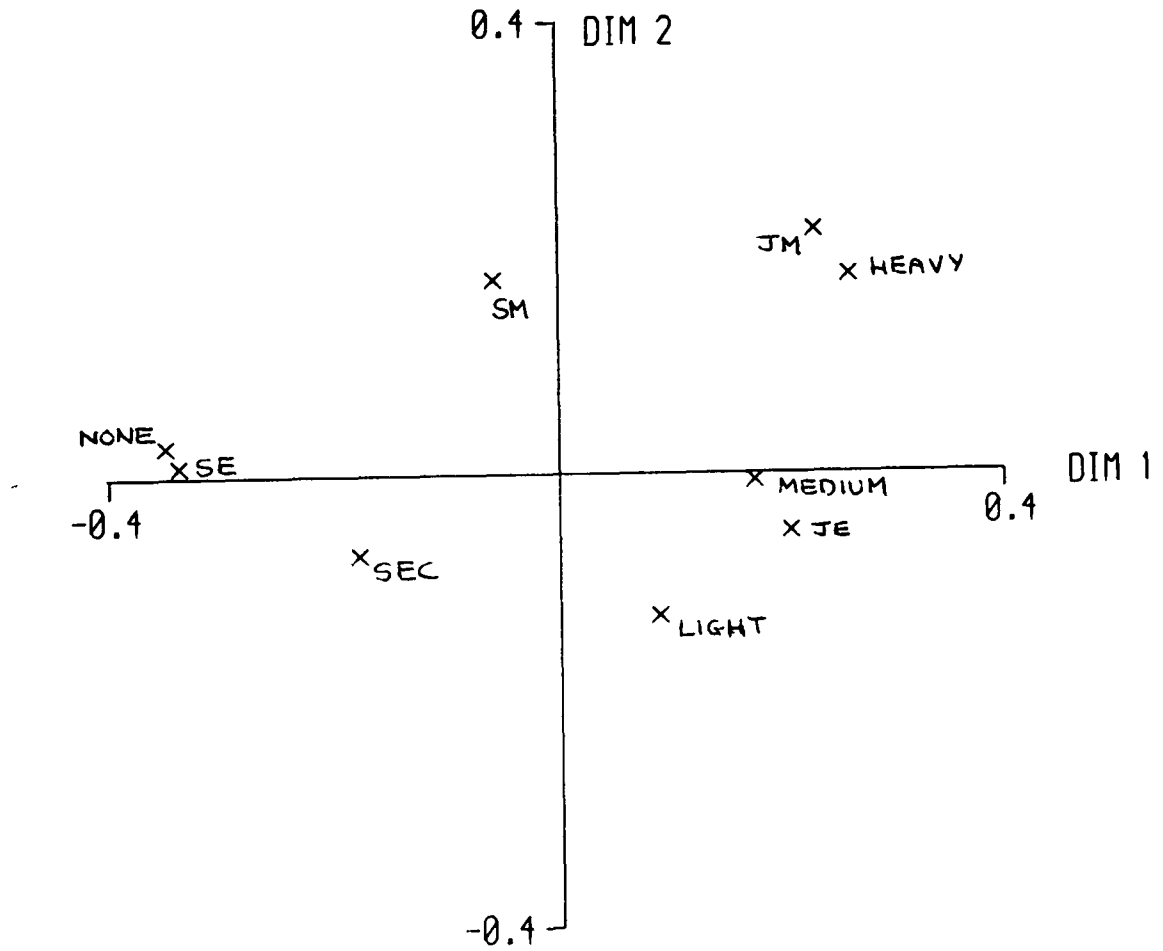
$\hat{\lambda}_1$		$\hat{\lambda}_2$	
Cell	Sign	Cell	Sign
3,4	-	2,4	+
2,1	-	2,2	-
4,1	-	1,4	+

The co-ordinates in the first dimension for 'Senior employees' and 'Heavy smoking' are far apart, and the influence when adding to cell (3,4) is negative, i.e. a decrease in the variance. The co-ordinates for 'Junior

Table 6.2.3
Artificial Contingency Table on the Smoking Habits of Personnel Staff

Staff	Smoking None	Light	Medium	Heavy	Total
Senior managers	4	2	3	2	11
Junior managers	4	3	7	4	18
Senior employees	25	10	12	4	51
Junior employees	18	24	33	13	88
Secretaries	10	6	7	2	25
Totals	61	45	62	25	193

Figure 6.2.2 Plot of the First Two Dimensions from the Correspondence Analysis of the Dataset on Smoking Habits of Employees.



managers' and 'Heavy smoking' in the second dimension are large and close together and the influence of cell (2,4) is positive.

The comparisons of the empirical and sample are better for the early than latter eigenvalues, but these may be of less interest than they are in PCA. Although not derived here, the second order terms for the eigenvalues will involve the influence function for the eigenvectors which is made up of terms in $(\lambda_j - \lambda_k)^{-1}$. When the eigenvalues are small and close together the more important the second order term is likely to be. We thus find that the empirical and sample influences differ more for these later dimensions. The two functions do pick out the same observations as influential but differ in the values. The most influential cells in these dimensions do tend to follow the patterns discussed above for the early dimensions. For example, cell (5,3) in the first dataset is the most influential on the fourth eigenvalue (the smallest eigenvalue) with a sample change of +8.8% and a smaller, but still largest change, of +4.3% for the empirical. Both these co-ordinates are large and close in this dimension, see Fig. 6.2.3.

There is a tendency for cells with low counts to be most influential on the later eigenvalues, especially when their row and column totals are small as well. The influences on $\hat{\lambda}_3$ for the first dataset are dominated by cells in column 5 and $\hat{\lambda}_4$ by cells in columns 3 and row 5. From Fig. 6.2.3 we can see that the co-ordinates of these categories are extreme in the dimension of the eigenvalue that they are influential on.

6.2.2. Influence on the Eigenvectors (*G* and *F* co-ordinates)

To compare the sample and empirical estimated changes we will take as our statistic the sums of squares of the individual changes in the G and *F* co-ordinates for each dimension. Table 6.2.4 gives the most influential cells on the *G* and *F* co-ordinates for the first two dimensions.

Figure 6.2.3 Plot of the Last Two Dimensions from the Correspondence Analysis of the Dataset on Worries of Israeli Adults.

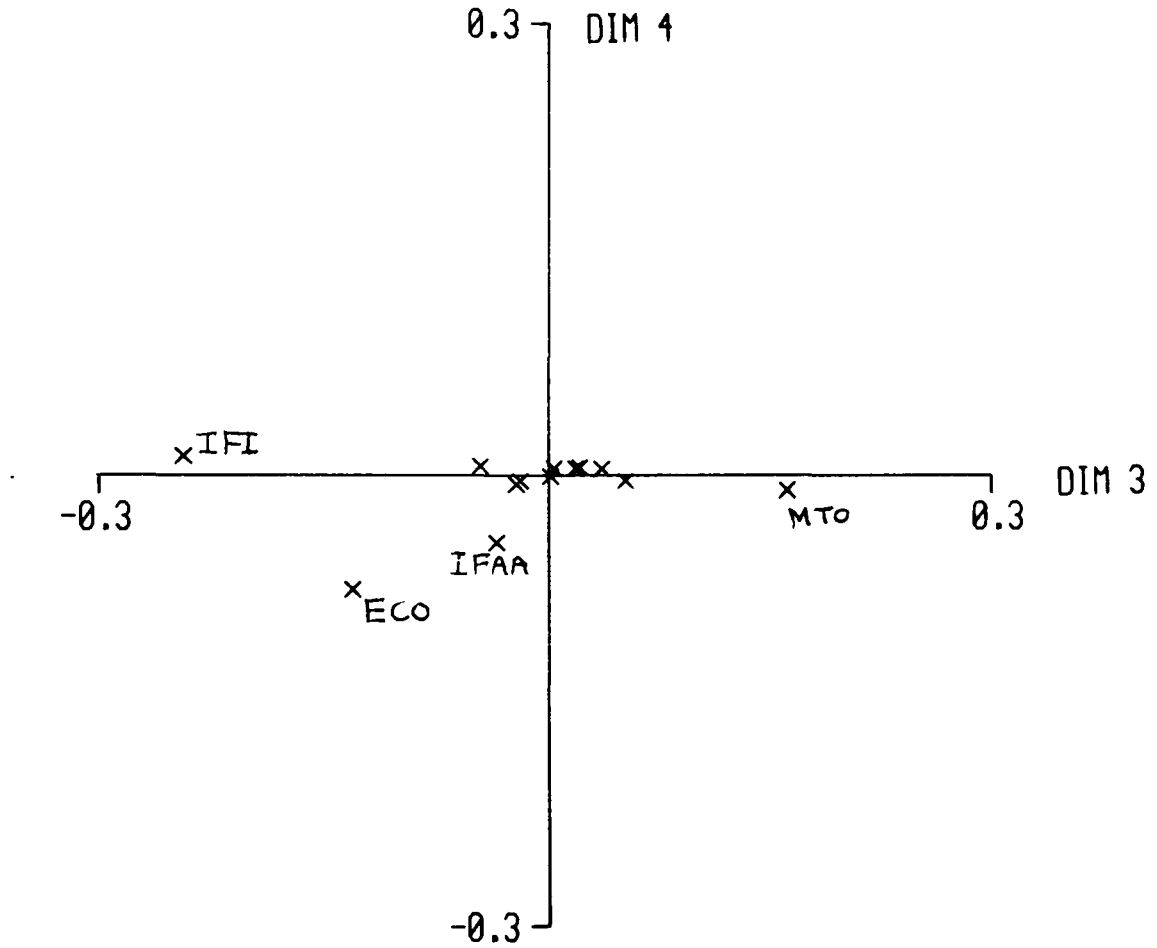


Table 6.2.4
Most Influential Observations on the G and F Co-ordinates

	$\underline{g}_1$			$\underline{g}_2$	
Obs	Sample	Empirical	Obs	Sample	Empirical
85	0.0018	0.0018	65	0.0011	0.0012
83	0.0009	0.0010	75	0.0008	0.0010
45	0.0006	0.0006	63	0.0005	0.0005

	$\underline{f}_1$			$\underline{f}_2$	
Obs	Sample	Empirical	Obs	Sample	Empirical
51	0.0056	0.0062	54	0.0170	0.0194
53	0.0043	0.0050	55	0.0021	0.0019
52	0.0012	0.0014	74	0.0017	0.0017

The changes are small partly because they are largely attributable to the observations affect on just one of the coefficients, see below. The comparisons are very close, especially if one looked at the individual changes in the co-ordinates. Again, the comparisons deteriorate for the latter dimensions although the order of ranked influences changes little. Unfortunately the asymptotic expressions for the influence on the G and F co-ordinates, see for example (5.3.2), are not simple and almost impossible to interpret. However, we can obtain a good insight into the influence on the co-ordinates from the numerical results and plots below.

The most influential cells, by the sums of squares of individual coefficient changes, on $\underline{f}_1$ and $\underline{f}_2$ are dominated by combinations with the row 5 and row 7 categories, i.e. (5,j) or (7,j) j=1,2...J. These rows have the smallest row totals. The column totals are much larger than the row totals and we find that the changes in $\underline{g}_1$ and $\underline{g}_2$ are less than for $\underline{f}_1$ and $\underline{f}_2$. However, the most influential cells on $\underline{g}_1$ and $\underline{g}_2$ are made up of cells involving the column categories 5 and 3, which have the smallest column totals. The most influential cells on the last two dimensions for both $\underline{g}$ and $\underline{f}$ are almost the same as each other. In the top ten ranked influences on $\underline{g}3$, $\underline{g}4$, $\underline{f}3$ or $\underline{f}4$ there is not one cell that does not involve one of the row numbers 5 or 7, or

one of the column numbers 3 or 5.

Table 6.2.5 gives the most influential cells on each of the individual co-ordinates, $f_{1i}\ i=1,....I$. If the ith row (or jth column) has a small total then the most influential cells on f_{1i} (or g_{1j}) tend to be those involving their own row (column) number. This does not occur for columns 1 and 2 which have large totals. The smaller the row (column) total the larger are the influences, i.e. the plotting positions are unstable. Although cell (5,1) was the most influential on $\underline{f}_1$ when we looked at the sums of squares of the individual changes, from Table 6.2.5 we see it is only due to its large affect on f_{15}. As well as cell numbers $(i,j), j=1,....J$ coming out as the most influential on f_{1i} we usually find that there is a pattern to the order of the column numbers. From Table 6.2.5 we see this order tends to be $(i,1), (i,3), (i,2)\ (i,5)$ and $(i,4)$, although often $(i,5)$ and particularly $(i,4)$ is not in the top 10 ranked influences on each f_{1i}. This order coincides with the ranked absolute g_{1j} co-ordinates, with g_{11} being the largest and g_{14} the smallest, see Fig. 6.2.1. Cells involving columns 1 and 3 lead to a decrease in f_{1i}, corresponding to the co-ordinate f_{1i} moving to the left, closer to $g_{1j}\ j=1,3$. Conversely, cell $(i,2)$ leads to an increase in f_{1i}. If f_{1i} increases, say, when we add one to cell (i,j) we usually find that all the other co-ordinates $f_{it}\ t\neq i$ decrease. This occurs since we have not only increased the association between row i and column j, so that the co-ordinates have moved closer, but we have relatively decreased the associations of the other rows with this column. This illustrates that correspondence analysis is a very useful technique at displaying the relationship between the rows and columns of a contingency table.

Whereas cells $(i,4)$ were the least influential out of the cells (i,j) on f_{1i} they are the most influential on $f_{2i}\ i=1,...7$ (though not f_{28}). The co-ordinate g_{2i}, corresponding to the category 'Israel: father Europe/America' is the

Table 6.2.5

Most Influential Cells on the Row Co-ordinates from the First Dimension of the Israeli Worries Contingency Table
Actual Sample Change × 100 in Parentheses

f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	f_{17}	f_{18}
11	21	31	41	51	84	71	82
(-0.59)	(-0.50)	(-0.34)	(-0.81)	(-7.46)	(-0.54)	(-2.04)	(0.69)
13	23	33	43	53	61	73	84
(-0.53)	(-0.45)	(-0.33)	(-0.64)	(-6.56)	(-0.47)	(-1.86)	(0.50)
12	22	84	45	52	63	72	81
(0.37)	(0.33)	(0.32)	(-0.25)	(3.50)	(-0.41)	(1.04)	(-0.29)
85	25	32	42	55	62	75	83
(0.22)	(-0.17)	(0.17)	(0.24)	(-2.73)	(0.32)	(-0.90)	(-0.26)
15	85	44	84	54	44	85	61
(-0.21)	(0.13)	(-0.17)	(-0.20)	(0.59)	(0.22)	(0.65)	(0.12)

Table 6.2.6

Most Influential Cells on the Row Co-ordinates from the First Dimension of the Israeli Worries Contingency Table When 10 is Added to Each Cell
Actual Sample Change × 10 in Parentheses

f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}	f_{17}	f_{18}
11	21	31	41	53	84	71	82
(-0.56)	(-0.48)	(-0.33)	(-0.76)	(-9.57)	(-0.45)	(-1.84)	(0.65)
13	23	84	43	55	61	73	53
(-0.48)	(-0.41)	(0.29)	(-0.36)	(-6.74)	(-0.44)	(-1.51)	(0.58)
12	22	33	42	51	63	72	83
(0.36)	(0.31)	(0.25)	(0.23)	(-5.11)	(-0.39)	(0.93)	(-0.44)
85	53	44	84	52	62	85	84
(0.26)	(0.22)	(-0.19)	(-0.21)	(2.43)	(0.30)	(0.55)	(0.35)
55	85	54	44	54	44	55	55
(0.25)	(0.18)	(-0.17)	(0.17)	(1.24)	(0.20)	(0.54)	(0.32)

largest column co-ordinate in the second dimension. This is a similar situation to the first dimension where the most influential cell involving the ith row was for the most extreme column co-ordinate in the first dimension.

For the latter dimensions the most influential cells on all the $\underline{f}$ and $\underline{g}$ co-ordinates include the rows and columns with small totals i.e. rows 5 and 7 and columns 3 and 5. From Fig. 6.2.3 we can see that the co-ordinates for these categories are extreme in the third or fourth dimension.

We will investigate influence on the G and F co-ordinates in the smaller contingency table on smoking habits, by two plots. These plots will illustrate some of the patterns observed above in the previous dataset. The plots give the new plotting positions, in the first two dimensions, when we add one to each cell individually in a given row. We have considered the first and fourth row in Table 6.2.3, which have the smallest and largest row totals respectively. In Fig. 6.2.4 we observe how the plotting positions change when we increment each cell in the first row, corresponding to the category 'Senior managers', by one. We can see how the co-ordinates in the first two dimensions for 'Senior managers' move towards the smoking category that one has been added to. This is nice to observe, given how we normally interpret our correspondence analysis. As discussed above for the 'Israeli' dataset, we see that whereas as the O symbol for 'Senior managers' moves to the left towards the 'No smoking' category, the O symbol on all the other employee categories have moved slightly to the right away from it. This can be observed for the other symbols as well. The variability in the 'Senior managers' co-ordinate is much greater than the changes in the smoking categories that have been added to. This happens because of the low row total for the first row. Adding to cells in the first row has little affect on the other row co-ordinates. The most influential cell in the first row overall seems to be (1,4), as the +

Figure 6.2.4 Plot of the First Two Dimensions from the Correspondence Analysis of the Dataset on Smoking Habits of Employees When 1 is Added to the Cells in the First Row.

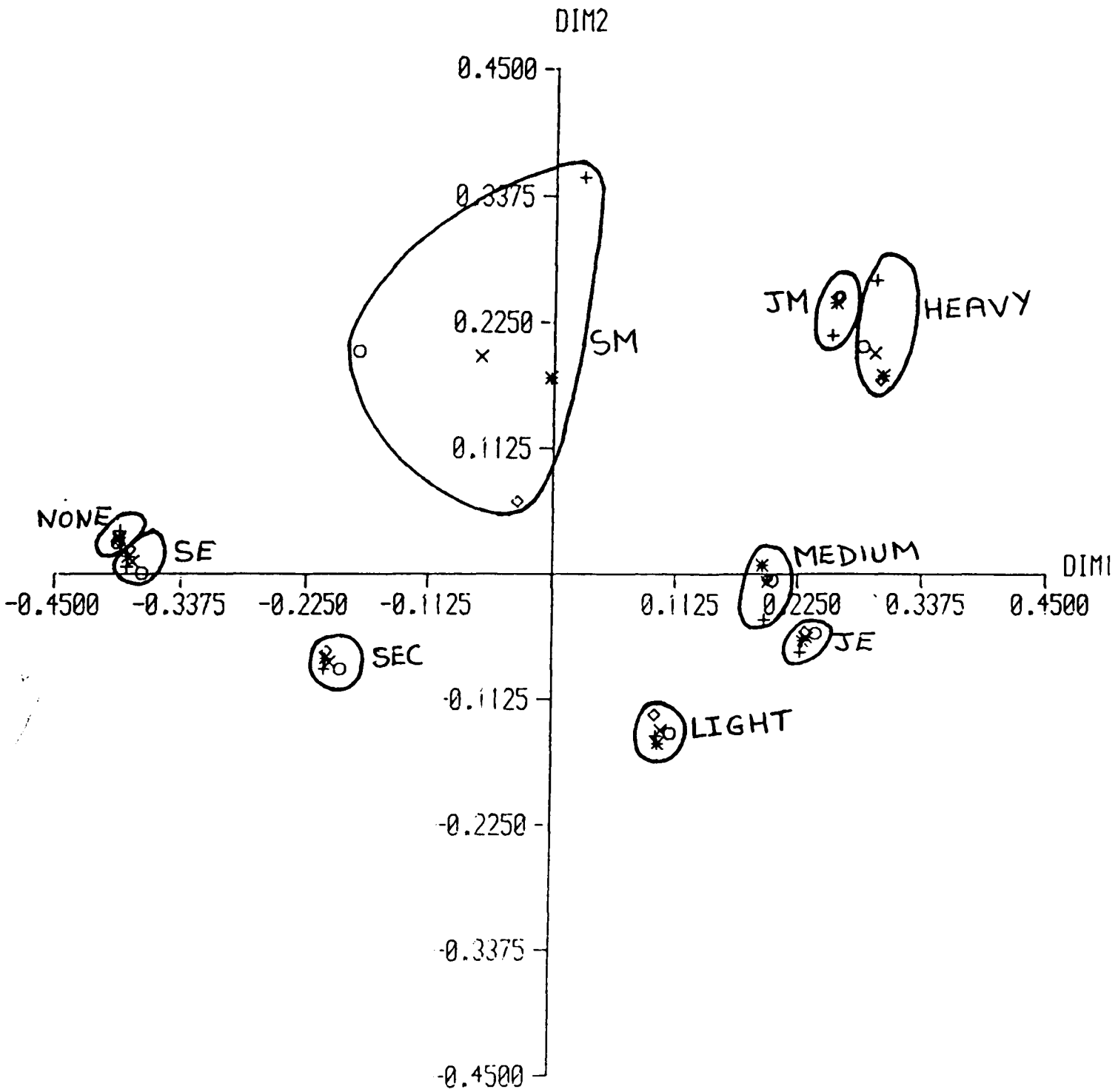
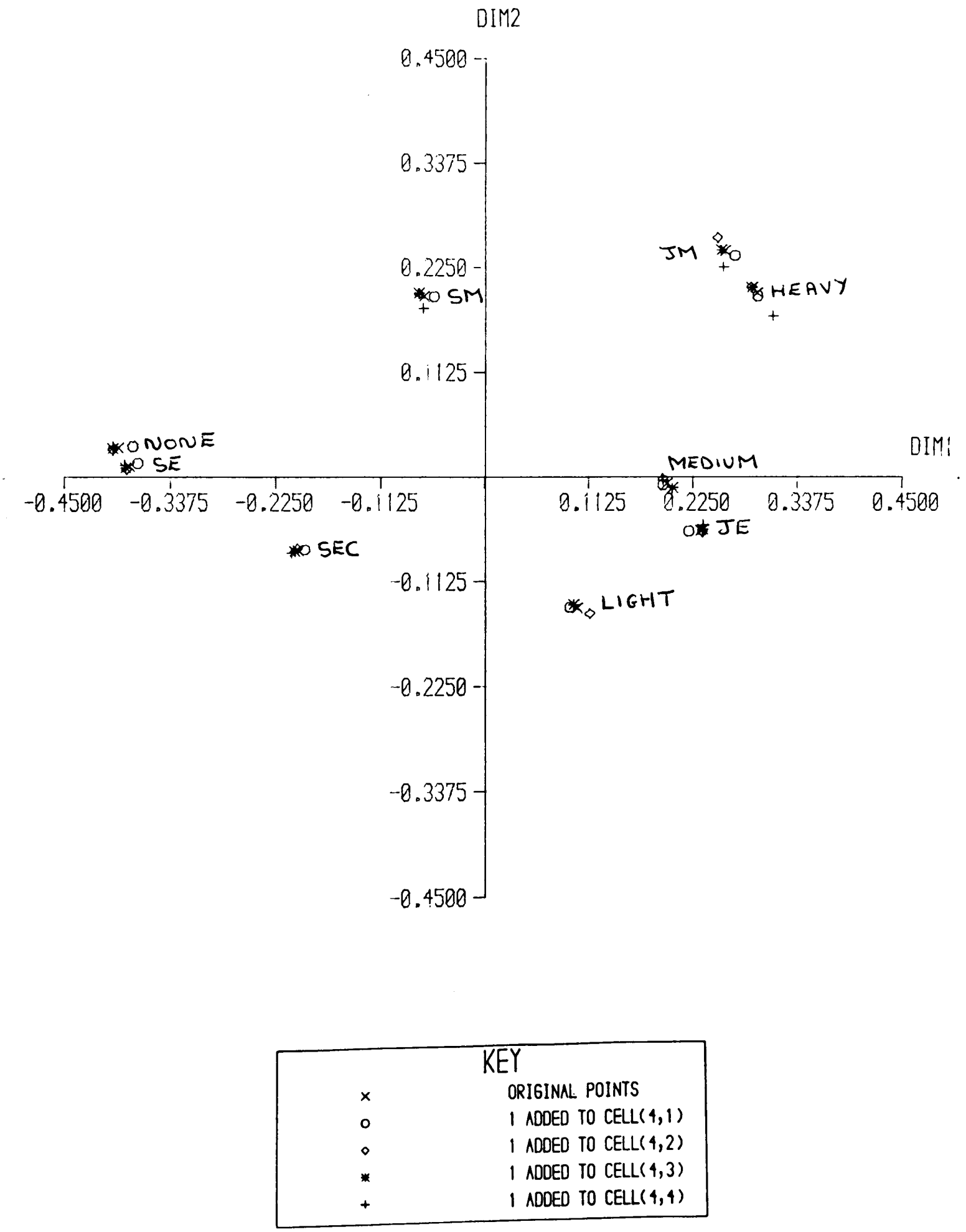


Figure 6.2.5 Plot of the First Two Dimensions from the Correspondence Analysis of the Dataset on Smoking Habits of Employees When 1 is Added to the Cells in the Fourth Row.



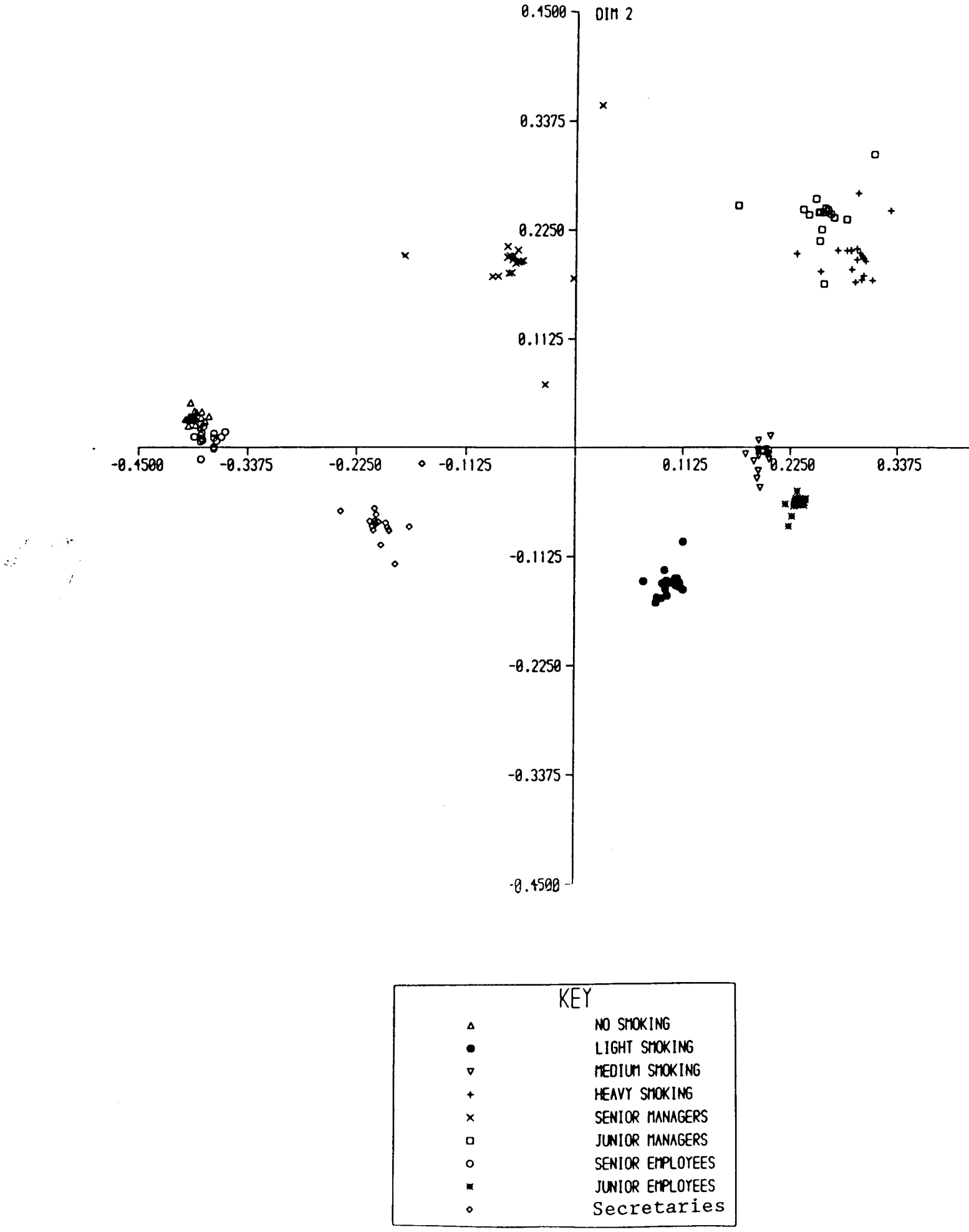
symbol seems to deviate the most for all the co-ordinates. Cell (1,4) is a small cell in the column with the lowest total.

The changes in Fig. 6.2.5 are very small. These correspond to adding one to the cells individually in the fourth row, which has the largest row total. The cells with the most influence are (4,1) and (4,4), which correspond to the O and + symbols respectively. However, these have as much affect, if not more, on the other categories as they do on the 4th row category 'Junior employees'. Also, the changes in the 'Junior employees' co-ordinates were as great in Fig. 6.2.4, when we were adding to cells in the first row. This was also observed in the 'Israeli' contingency table, with columns with large totals not having cells involving their own row numbers as the most influential. There is usually little pattern to what will be most influential, but most prevalent seems to be cells who have extreme co-ordinates in the dimension. However, these influences are not usually large.

6.2.3. Summary and Discussion

The above results show, perhaps unsurprisingly, that the most sensitive plotting positions in the first two dimensions correspond to the categories based on the least information, i.e. their row/column total is small. Plots such as Fig. 6.2.4 provides useful and clear information on the sensitivity of our analysis. One would not necessarily want the different symbols for each cell that is perturbed (this was done above to illustrate what direction the co-ordinates moved in) but to have a different symbol for each category co-ordinate and consider the changes for adding to each cell on one plot. See for example Fig. 6.2.6, which clearly indicates which categories are the least stable. If one felt that all the rows and columns of the contingency table had reasonably large totals then such an analysis may not be needed. The above also shows that in the first two dimensions the G and F co-ordinates tended to

Figure 6.2.6 Plot of the First Two Dimensions from the Correspondence Analysis of the Dataset on Smoking Habits of Employees When 1 is Added to Each Cell in Turn.



be most affected by cells in their own row/column, although there was little pattern to the most influential on the categories with large totals. However, the changes in these tended to be small anyway. In the latter dimensions the same cells tended to be the most influential on all aspects of the analysis, and these were usually cells with small entries in rows and columns with small totals.

The theoretical expression for the changes in the eigenvalues proved helpful in describing influence on the eigenvalues, and in particular what cells cause the eigenvalues (inertias) to increase or decrease. However, the theoretical expressions for the G and F co-ordinates are not very informative.

The specialisation of the Burt analysis empirical expressions, which were derived in § 5.5 for adding m observations, to the two way contingency table analysis implies that the affect of adding more than one to a cell is additive (i.e. the influence is multiplied up by m as we are considering the addition of m into the same cell). Table 6.2.6 gives the five most influential observations on f_{1i} when 10 is added to each cell of the 'Israeli' contingency table. These are almost the same as in Table 6.2.5, with the values of the influence being approximately ten times bigger, sometimes more and sometimes less. There are exceptions in the order of influence but it is difficult to say why cells go up or down the lists of ordered influence as m increases. Many of the influences in the smaller 'Smoking' dataset were also a factor of 10 greater when 10 rather than 1 was added to each cell.

We obtain a similar type of plot to Fig. 6.2.6 if we consider bootstrapping of the original $n_{..} = 193$ observations of dummy variables making up the 'Smoking' contingency table (see Greenacre (1984), Chapter 8). This involves drawing samples of size $n_{..} = 193$, by sampling from the original variables, and plotting the new G and F co-ordinates of the resultant

contingency tables. The original sample is thus being treated like the underlying population and we are looking at the possible contingency tables that could have arisen. The procedure is thus best if the original data was also drawn randomly since it will be more representative of the underlying distribution. However, as Greenacre notes most data is collected in a 'deliberate non-random fashion'. The generated contingency tables are thus greater perturbations of the original than the influence technique for adding in a single observation (i.e. adding 1 to a cell). Both techniques will reveal which plotting positions are the least stable, but as discussed above we may only need to consider such plots if we think some of our row or column totals are small.

6.3. Influence When Omitting a Row from a Contingency Table

We will investigate influence when omitting a row from a contingency table using three datasets. The first was introduced in Section 4.9 where we considered influence on the covariance biplot which we will compare with the influence on the correspondence analysis display. The contingency table for this dataset is given in Table 4.9.1 and the original correspondence analysis display for the first two dimensions in Fig. 6.3.1. The second dataset is on the worries of Israeli adults and was introduced in the previous section. Its contingency table is given in Table 6.2.1 and correspondence display in Fig. 6.2.1. The last dataset is taken from Gabriel and Zamir (1979) (but it is also in Greenacre (1984, p268)). Table 6.3.1 gives the contingency table and Fig. 6.3.2 the correspondence analysis display of the first two dimensions. This dataset is concerned with the science doctorates in the USA for different years. As before we will not consider the interpretation of the original plots, one is referred to Greenacre (1984), except where it is relevant in our influence studies. Note, we are again considering the deletion of rows so our

Figure 6.3.1 Plot of the First Two Dimensions from the Correspondence Analysis of the Dataset on Protein Consumption in Europe.

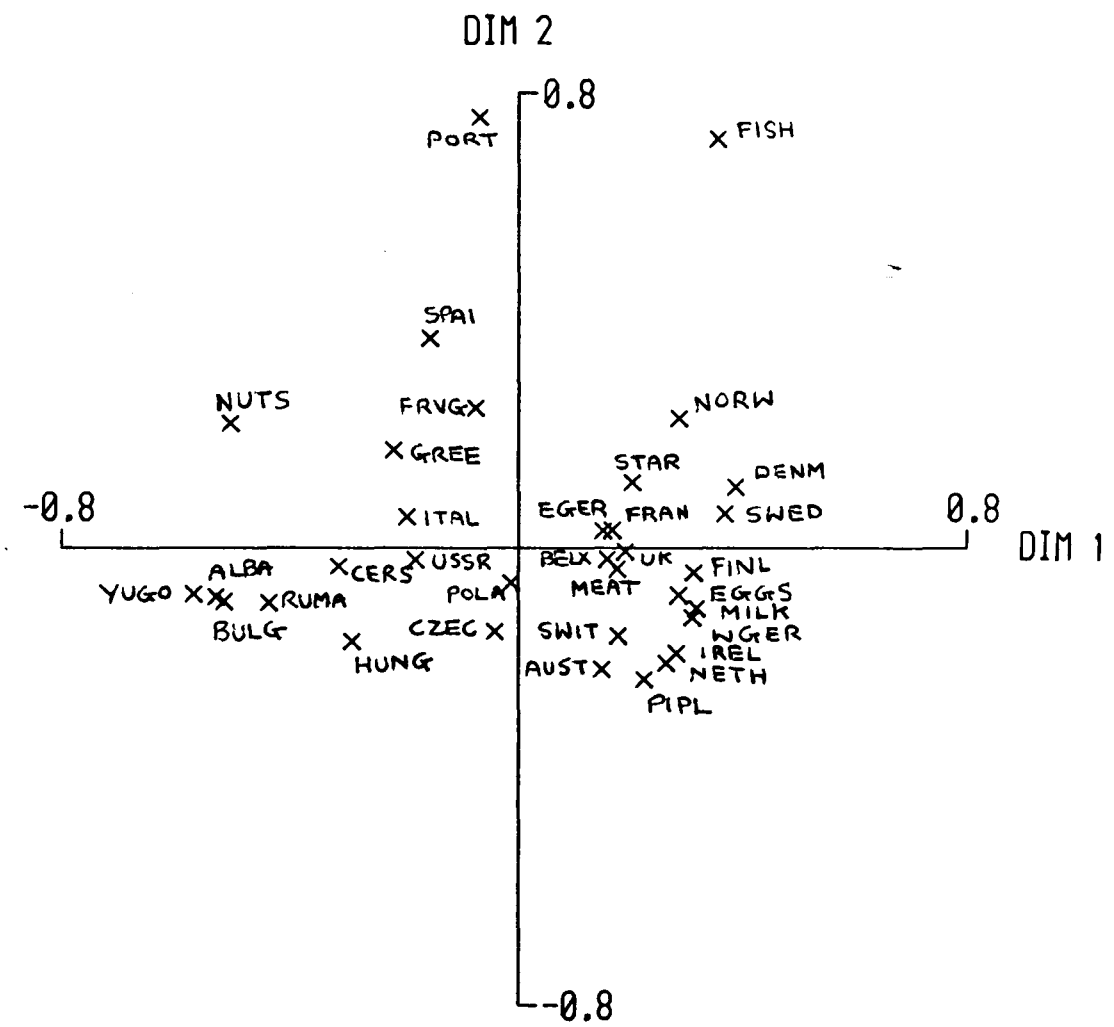


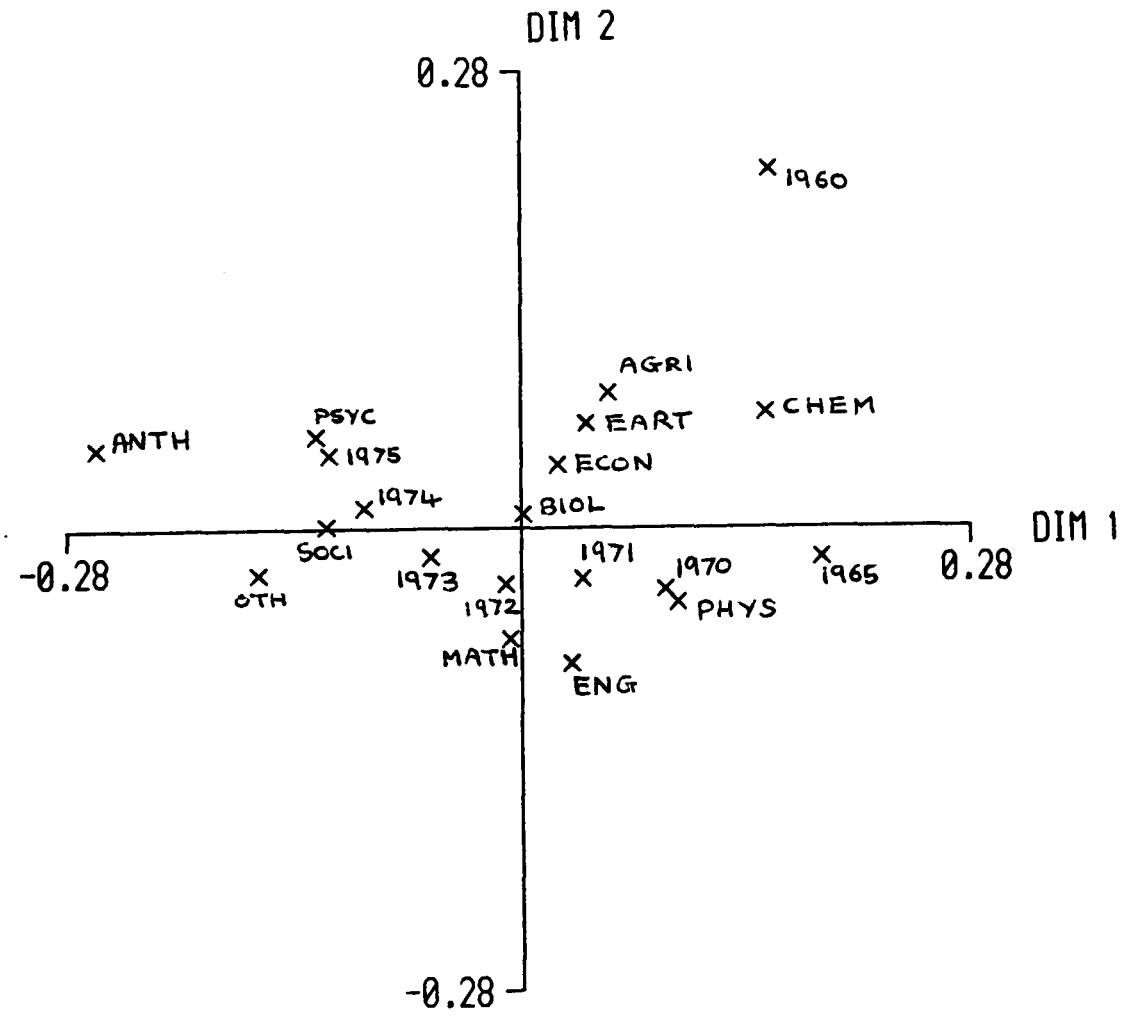
Table 6.3.1
Data on Science Doctorates in the USA

Subject	Year 1960	1965	1970	1971	1972	1973	1974	1975	Total
ENG	794	2073	3432	3495	3475	3338	3144	2959	22710
MATH	291	685	1222	1236	1281	1222	1196	1149	8282
PHYS	530	1046	1655	1740	1635	1590	1334	1293	10823
CHEM	1078	1444	2234	2204	2011	1849	1792	1762	14374
EART	253	375	511	550	580	577	570	556	3972
BIOL	1245	1963	3360	3633	3580	3636	3473	3498	24388
AGRI	414	576	803	900	855	853	830	904	6135
PSYC	772	954	1888	2116	2262	2444	2587	2749	15772
SOCI	162	239	504	583	638	599	645	680	4050
ECON	341	538	826	791	863	907	833	867	5966
ANTH	69	82	217	240	260	324	381	385	1958
OTH	314	502	1079	1392	1500	1609	1531	1550	9477
Totals	6263	10477	17731	18880	18940	18948	18316	18352	127907

Abbreviations

ENG	Engineering	MATH	Mathematics	PHYS	Physics
CHEM	Chemistry	EART	Earth sc.	BIOL	Biological sc.
AGRI	Agricultural sc.	PSYC	Psychology	SOCI	Sociology
ECON	Economics	ANTH	Anthropology	OTH	Other Social sc.

Figure 6.3.2 Plot of the First Two Dimensions from the Correspondence Analysis of the Dataset on Science Doctorates in the USA.



signs of influence will be different to those in the previous section for increases and decreases in the eigenvalues.

We will consider the comparisons of empirical and sample influences and discuss the usefulness (or lack of it) of the empirical expressions throughout. This will be summarised in Section 6.2.6, and in Chapter 7.

6.3.1. Influence on the Eigenvalues (Principal Inertias)

The empirical and sample influence functions give similar rankings for the eigenvalues from the first two dimensions of the Protein Consumption dataset. The top three most influential rows are given in Table 6.3.2, and for the Israeli dataset, where $n=8$, in Table 6.3.3. In Table 6.3.2 we can see the usual pattern of the empirical (estimated change) underestimating the large sample (actual change) influences. An exception to this is 'Portugal' on $\hat{\lambda}_2$. As the influences become smaller the empirical and sample values become closer, as occurs in most applications. The empirical and sample differ most on the second eigenvalue from the Israeli dataset, particularly on 'Personal economics'. This row is highly influential on $\hat{\lambda}_1$, for both functions, and we find when it is omitted that the first two dimensions rotate (and virtually switch), see Fig. 6.3.3. This explains why the empirical and sample disagree on its influence on $\hat{\lambda}_2$ as the former will not take this rotation, due to close perturbed eigenvalues, into account.

For both datasets it has been the rows with extreme co-ordinates that have been the most influential and they have positive influences, which represents a decrease in the variances along the relevant directions. A counter example to this is found in the Doctoral dataset. From Fig. 6.3.2 we see that 'Anthropology' has the largest row co-ordinate in the first dimension but it has only the sixth largest absolute influence (using either the sample or empirical influence function) on the first eigenvalue. The most influential are

Table 6.3.2

Three Most Influential Rows on the Eigenvalues in the First Two Dimensions of the Dataset on Protein Consumption and their Sample and Empirical Influences

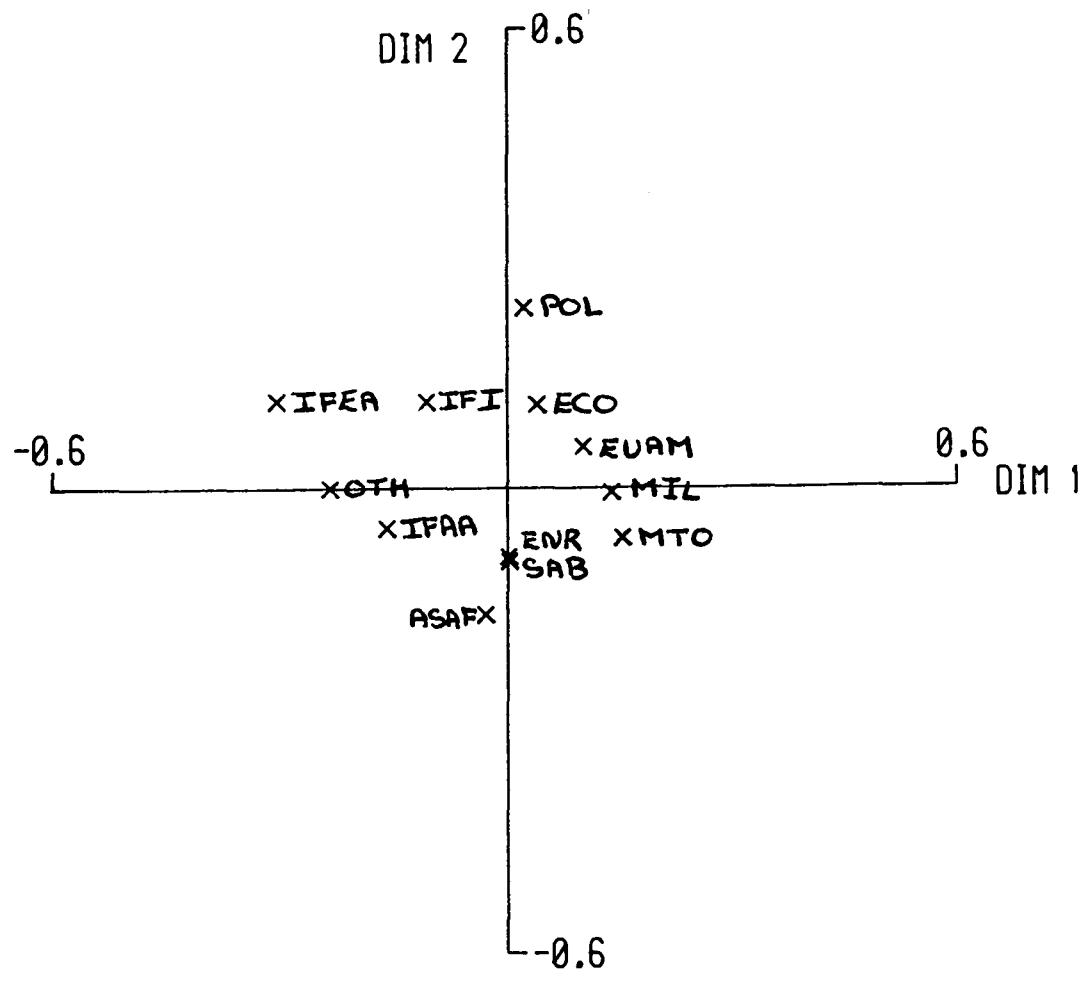
$\lambda_1 = 0.087$			$\lambda_2 = 0.039$		
Obsn.	Sam.	Emp.	Obsn.	Sam.	Emp.
Yugo	11.6%	10.8%	Port	36.4%	41.7%
Bulg	9.1%	8.5%	Spai	7.1%	6.8%
Alba	7.0%	7.0%	Denm	-6.6%	-6.3%

Table 6.3.3

Most Influential Observations on the Eigenvalues fron the First Two Dimensions of the Dataset on Worries of Israeli Adults and their Sample and Empirical Influences.

$\lambda_1 = 0.060$			$\lambda_2 = 0.015$		
Obsn.	Sam.	Emp.	Obsn.	Sam.	Emp.
Per	69.7%	57.5%	Oth	62.0%	32.0%
Oth	-20.9%	-17.2%	Per	29.6%	-8.8%
Mil	-17.4%	-14.8%	Mil	18.6%	8.4%

Figure 6.3.3 Plot of the First Two Dimensions from the Correspondence Analysis of the Dataset on Worries of Israeli Adults When 'Personal Economics' is Deleted.



'Biological sciences', 'Engineering' and 'Chemistry' which have negative,negative and positive influences respectively. Since both functions agree on the lack of influence for 'Anthropology' we can use our theoretical expression to find the reason. Extreme points, as seen in Tables 6.3.2 and 6.3.3 have positive influences and small co-ordinates negative influences, which represents a decrease and increase in the variance respectively, when they are omitted. Usually the number of negative influences outweigh the number of positive. The empirical influence function is

$$\frac{1}{n_{..}-m}EIC(\underline{x},\hat{\lambda}_k)=\frac{1}{n_{..}-m}\left[-\sum_{j=1}^{J}r_j^* g_{kj}^2+mf_{ki}^2\right] \qquad (6.3.1)$$

where r_j^* is the jth element of the row and m is the row sum. We thus find that 'Anthropology' has little influence because of its low row sum, or mass $\frac{m}{n_{..}}$, which makes the positive term small.

The influence measure of Escofier and Le Roux (1976), see Greenacre (1984, p211) which is derived in the same manner as that for the eigenvalue in a covariance PCA (see § 3.8.4), also depends on the mass. This influence measure is

$$I_{EF}=-\frac{m}{n_{..}-m}\hat{\lambda}_k+\frac{mn_{..}}{(n_{..}-m)^2}f_{ki}^2 \qquad (6.3.2)$$

The term f_{ki}^2 again obtains more weight than the first term when m is large, i.e. $(n_{..}-m)$ is small. The importance of the mass $\frac{m}{n_{..}}$ stems from the normalisation of our vectors which is

$$\underline{f}'_k D_r \underline{f}_k=\lambda_k \quad ,$$

where $\frac{m}{n_{..}}$ will be the ith diagonal element of D_r, which is a diagonal matrix of the row totals of P. The term mf_{ki}^2 is thus called the contribution to the kth inertia of the ith point. The first term in (6.3.2) is different to that in (6.3.1) with the former just involving λ_k and the latter a function of the G co-

ordinates. Since the G co-ordinates also have the normalisation

$$\underline{g}'_k D_c \underline{g}_k = \lambda_k \quad ,$$

and D_c is the diagonal matrix of the column sums of P, which involves $\frac{r^*}{n_{..}}$, the first term in (6.3.1) is like another expression of the current contribution of the ith row to the eigenvalue. Expression (6.3.2) behaves similarly to the empirical influence but tends to be larger in absolute terms. We also find that whereas (6.3.1) is zero for the trivial dimension, when $f_{ki} = g_{kj} = \lambda_k = 1$ expression (6.3.2) is not, but has the value $\frac{m^2}{(n_{..} - m)^2}$.

The masses in the Protein Consumption dataset are all about equal and we find that the order of non-absolute influence on $\hat{\lambda}_1$ coincides exactly with the ordering by f_{ki}^2 (or absolute f_{ki}) value. This was less true for the other dimensions and other datasets, but it does hold to a certain extent in the early dimensions of many datasets. This means that a point with a very small f_{ki}^2 value can have a larger affect than one with a larger f_{ki}^2 value, but its influence will be negative (i.e. increase $\hat{\lambda}_k$) rather than positive. For example, in the Protein Consumption 'Poland', see Fig. 6.3.1, has the smallest score on the first dimension but it has the fifth largest absolute influence, and the largest negative influence, on the first eigenvalue. Only 'Albania', 'Bulgaria', 'Rumania' and 'Yugoslavia' have a larger affect, but their influence is positive. This pattern is most interesting in the Israeli dataset. From Table 6.3.3 we see that the second most influential row in the first dimension is 'Other' and it has a small f_{1i} value, see Fig. 6.2.1. The category 'Political situation' has a large f_{1i}^2 value but it is only the sixth largest absolute influence.

6.3.2. Influence on the G Co-ordinates

In a practical situation it is unlikely that one would wish to routinely look at the individual influences on all the G and F co-ordinates as this would lead to many statistics to examine. However, in the next two sections we investigate the individual influences to see if we can detect any patterns, and in § 6.3.4 we will consider scalar measures of influence that one can calculate. If one finds a row to be influential from looking at the scalar measurements one may wish to examine how it has affected the individual co-ordinates. This is best examined by plotting the perturbed co-ordinates and comparing against the original correspondence analysis plot. Such plots will also be examined in the next two sections and we will see that even when we have quite large changes in our co-ordinates the interpretation of the correspondence analysis display may hardly change. The most influential observations on the $\underline{g}_1$ co-ordinates from the Protein Consumption dataset are,

g_{11}	g_{12}	g_{13}	g_{14}	g_{15}	g_{16}	g_{17}	g_{18}	g_{19}
Yugo	Hung	Alba	Finl	Port	Yugo	Bulg	Bulg	Alba
Alba	Alba	Yugo	Bulg	Yugo	Bulg	Alba	Denm	Yugo

and comparisons of the empirical and sample changes for the most influential on each co-ordinate are,

	g_{11}	g_{12}	g_{13}	g_{14}	g_{15}	g_{16}	g_{17}	g_{18}	g_{19}
SAM	0.037	-0.056	0.037	0.020	-0.090	-0.021	0.054	0.052	0.050
EMP	0.033	-0.054	0.035	0.018	-0.105	-0.016	0.049	0.045	0.047

These comparisons, and similar comparisons for the less influential rows, are close indicating that the empirical expressions reflect the sample influences well. Unfortunately, the empirical expressions are not easily interpretated. We thus need to examine observed results to detect any patterns.

From Fig. 6.3.1 we can see that the most influential rows on the individual $\underline{g}_1$ co-ordinates tend to be extreme in the first dimension, although

there are exceptions such as 'Portugal' on g_{15}. However, Portugal is extreme on the second dimension so this may be some carry on effect. 'Yugoslavia','Albania' and 'Bulgaria' are all close in the first dimension and are the most extreme rows, but often one of these will not be in the top group of influence on a g_{1j} co-ordinate even though the others are. Whether a row comes out in the top group of influence on the g_{1j} co-ordinate depends jointly on whether it is extreme in the dimension and on the size of its jth residual from the independence model. We thus need to consider the (i,j) th cell of the matrix $P-\underline{rc}'$, which is given in Table 6.3.4, multiplied through by $n_{..}$. We also find that the sign of the influence of omitting a row on a g_{kj} co-ordinate depends on the sign of residual. Countries 'Yugoslavia', 'Albania', 'Bulgaria' and 'Hungary', which lie on the LHS of the plot, have a positive influence on a g_{kj} co-ordinate when the residual is negative and vice versa. For example, 'Yugoslavia' has a large negative residual in cell (25,1) on the 'Meat' category and its influence on g_{11} is positive (i.e. g_{11} decreases when 'Yugoslavia' is omitted). Conversely, 'Yugoslavia' has a positive residual in cell (25,6) on the 'Cereals' category, and its influence on g_{16} is negative. Since 'Yugoslavia' and 'Meat' have a negative residual we find that they are positioned at opposing ends of the correspondence analysis display, whereas 'Cereals' is positioned on the left of the plot near 'Yugoslavia'. The above influences, when 'Yugoslavia' is omitted, thus represent the display moving inwards towards the origin since g_{11} decreases and g_{16} increases. Those pulled out close to 'Yugoslavia' and those at the other end of the plot, due to a large negative residual, move inwards when it is removed.

In a similar way, Finland which lies on the RHS of the plot has an influence whose sign coincides with that of the sign of the residual. This also represents the display of column co-ordinates moving inwards. For example,

Table 6.3.4

Table of Residuals from the Independence Model (Multiplied by $n_{..}$) for the Contingency Table on Protein Consumption

Country	MEAT	PIPL	EGGS	MILK	FISH	CERS	STAR	NUTS	FRVG
ALBA	1.94	-5.15	-1.94	-5.30	-3.36	15.54	-2.95	2.95	-1.73
AUST	-1.00	6.05	1.34	2.67	-2.21	-4.48	-0.71	-1.79	0.13
BELX	3.50	1.26	1.11	0.09	0.14	-6.22	1.35	-1.03	-0.21
BULG	-2.58	-2.34	-1.50	-9.77	-3.32	22.64	-3.42	0.46	-0.17
CZEC	0.21	3.78	-0.03	-4.02	-2.13	3.18	0.87	-1.87	0.01
DENM	0.31	2.53	0.63	7.09	5.42	-11.86	0.32	-2.52	-1.93
EGER	-0.27	4.63	1.11	-4.00	1.62	-3.86	2.73	-1.91	-0.05
FINL	-0.86	-3.42	-0.39	15.67	1.29	-7.68	0.59	-2.24	-2.96
FRAN	6.75	0.86	-0.06	-0.09	0.80	-8.81	-0.09	-1.12	1.77
GREE	-0.99	-5.99	-0.54	-1.89	1.02	4.97	-2.67	4.30	1.79
HUNG	-4.36	4.64	0.01	-7.12	-3.91	8.41	-0.20	2.38	0.14
IREL	3.44	1.60	1.58	7.59	-2.36	-10.32	1.65	-1.67	-1.50
ITAL	-0.62	-2.63	0.03	-3.06	-0.79	5.22	-2.09	1.29	2.65
NETH	-0.20	5.80	0.70	6.51	-1.73	-9.44	-0.02	-1.23	-0.38
NORW	0.04	-2.82	-0.10	7.00	5.62	-7.71	0.53	-1.33	-1.24
POLA	-3.72	1.67	-0.47	0.81	-1.63	1.25	1.28	-1.32	2.13
PORT	-2.46	-3.26	-1.49	-10.18	10.42	-1.42	2.13	1.99	4.26
RUMA	-3.76	-1.70	-1.47	-6.23	-3.34	16.93	-1.23	2.19	-1.39
SPAI	-1.74	-3.71	0.46	-6.80	3.14	0.18	1.85	3.14	3.48
SWED	0.74	0.44	0.76	8.74	3.5	-10.57	-0.29	-1.46	-1.86
SWIT	3.01	1.99	0.08	6.23	-2.10	-7.52	-1.59	-0.75	0.65
UK	7.27	-2.44	1.67	2.97	-0.11	-8.93	0.29	0.23	-0.96
USSR	-1.23	-3.86	-1.05	-1.73	-1.59	9.05	1.82	0.11	-1.53
WGER	2.32	5.20	1.39	2.98	-0.56	-11.21	1.25	-1.34	-0.02
YUGO	-5.74	-3.15	-1.83	-8.15	-3.82	22.63	-1.41	2.53	-1.07

Finland has a positive residual in cell (8,4) and a positive influence on the 'Milk' co-ordinate g_{14} which also lies on the RHS of the plot. A positive influence is a decrease in the co-ordinate (since we are omitting rows) so it moves to the left and so inwards towards the origin.

Portugal is the most influential row on all the column co-ordinates in the second dimension except on 'Eggs' and 'Nuts' where it only has ranked positions 9 and 20 respectively. The most influential rows on g_{23} (corresponding to 'Eggs') and g_{28} (corresponding to 'Nuts') are 'Spain' and 'Bulgaria' respectively. We find the sample and empirical differ on the values of influence for 'Portugal' but both give it as the most influential on the same $\underline{g}_2$ co-ordinates. We have the influences on $\underline{g}_2$ for 'Portugal' are

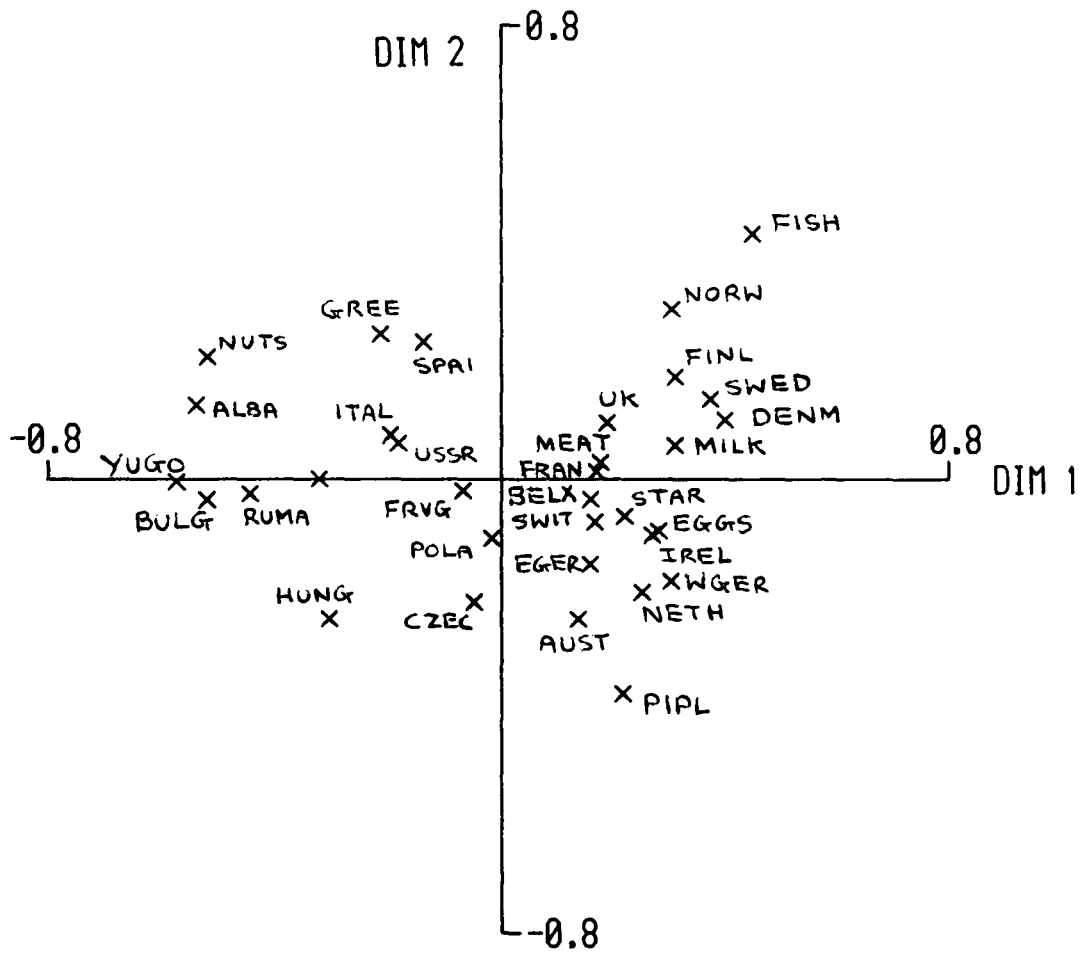
	Sample	Empirical	Estimate
g_{21}	-0.066	-0.025	-0.049
g_{22}	0.148	0.070	0.076
g_{23}	0.009	-0.016	0.008
g_{24}	-0.165	-0.066	-0.130
g_{25}	0.286	0.135	0.290
g_{26}	-0.032	-0.028	-0.031
g_{27}	0.180	0.089	0.146
g_{28}	0.003	-0.026	0.003
g_{29}	0.266	0.109	0.230

The empirical is two to three times smaller than the sample influences and sometimes varies in sign, but this is only for the small influences. This underestimation for the very large influences has been observed in previous sections. We discussed in Chapter 2 how for the correlation coefficients the empirical based on the full dataset tends to underestimate the deleted sample function and over estimate the sample changes when we add in an extra point. The third column above, under 'Estimate', is the average of the empirical influence calculated from the data without 'Portugal' included and our usual empirical based on the full dataset. We see the 'Estimate' tends to give values closer to the sample function than the empirical function recorded in the table, since the two empirical functions tend to sandwich the sample results.

As for the extreme points in the first dimension the sign of the influence of 'Portugal' on the $\underline{g}_2$ co-ordinates is linked to the sign of Portugal's residual on the column categories. Since 'Portugal' is at the top of t he plot i.e. positive on the second axis, the sign of the residual coincides with the sign of the influence, as it did for Finland which was positive on the first axis. This again represents the display moving inwards towards the origin. There were exceptions to the above pattern on the co-ordinates in the second dimension for the 'Pigs and Poultry' and 'Eggs' categories. The affect of 'Portugal' on the 'Eggs' co-ordinate g_{23} was very small compared to some of its other influences, so this lack of pattern maybe less important. The affect on 'Pigs and Poultry' will be discussed further by looking at the change in the correspondence display.

In § 4.9 we examined how the covariance biplot for the above dataset changed when 'Portugal', which was again extreme in the second dimension, was omitted. The markers for 'Fish' and 'Pigs and Poultry' were found to move much, changing some of our initial interpretation of the plot. In the original biplot the 'Pigs and Poultry' marker was positioned close to some of the meat products which it had quite low correlations with, due to its high correlation with 'Eggs' and negative correlation with 'Fish'. In the original correspondence display, in Fig. 6.3.1, 'Pigs and Poultry' is somewhat separated from the other food sources but it is close to 'Eggs' in the first dimension, and we find its positioning is quite steady. Fig 6.3.4 is the correspondence analysis when 'Portugal' is omitted. This is also given by Greenacre (1984, p288) who displays 'Portugal' on the plot as a supplementary point. (The connection between the supplementary point expression and the empirical influence for the included point was discussed in § 5.4.1). Compared to the covariance biplot, the correspondence plot has

Figure 6.3.4 Plot of the First Two Dimensions from the Correspondence Analysis of the Dataset on Protein Consumption in Europe When 'Portugal' is Deleted.



changed little even though the influences for 'Portugal' on five of the $\underline{g}_2$ co-ordinates are much larger than for any other row. Whereas the 'Fish' and 'Pigs and Poultry' markers were unsteady in the covariance biplot we see that they remain in similar positions in the perturbed correspondence analysis display. 'Fish' has become less extreme going from 0.72 to 0.43 in the second dimension but 'Pigs and Poultry' has become more extreme changing from -0.23 to -0.38. As noted above, 'Pigs and Poultry' is thus contrary to the usual pattern of the display moving inwards towards the origin. This may be caused by the low association between 'Pigs and Poultry' and 'Fish', which meant a negative correlation in the covariance biplot. Since 'Fish' has moved downwards 'Pigs and Poultry' moves down also to keep away from it. The only co-ordinate to change its relative position is 'Fruit and Vegetables' which has moved away from 'Spain', even though from Table 6.3.4 it has a large residual on it, and is positioned closer to 'Poland' which it also has a large positive residual for. We will consider the row co-ordinate changes in the next section.

For the Israeli dataset there is one large f_{1i} co-ordinate for 'Personal economics' and we find it is the most influential on all the g_{1j} co-ordinates except for g_{15}, relating to the category 'Israel:father Israel ' (IFI). 'Personal economics' lies on the LHS of the first axis, and like 'Yugoslavia' from the previous dataset it has a negative on a g_{1j} co-ordinate when its residual for that category is positive, see Table 6.3.5. However, from the perturbed correspondence plot for the first two dimensions, see Fig. 6.3.3, we see this pattern may be coincidental since it is caused by the anti-clockwise rotation of the plot. This means that 'Israel:father Europe/America' (IFEA) has moved to the left, but does not represent the display moving inwards since it has become extreme in the perturbed display with a large negative score.

Table 6.3.5

Table of Residuals from the Independence Model (Multiplied by $n_{..}$) for the Contingency Table on Israeli Worries.

Worry	ASAF	EUAM	IFAA	IFEA	IFI
ENR	0.64	2.84	-0.49	-0.91	-2.08
SAB	1.49	2.19	-0.64	-2.00	-1.03
MIL	-14.36	31.36	-3.67	-14.27	0.94
POL	-25.64	21.39	-2.11	6.12	0.24
ECO	-1.73	1.39	0.19	-0.18	0.33
OTH	-5.62	-17.16	1.81	19.13	1.84
MTO	-1.13	6.59	-0.97	-2.02	-2.48
PER	46.36	-48.61	5.89	-5.88	2.24

Table 6.3.6

Table of Residuals form the Independence Model (Multiplied by $n_{..}$) for the Contingency Table on Science Doctorates

Science	1960	1965	1970	1971	1972	1973	1974	1975
ENG	-318.00	212.80	283.85	142.84	112.19	-26.23	-108.02	-299.41
MATH	-114.53	6.61	73.91	13.52	54.63	-4.89	10.04	-39.30
PHYS	0.05	159.48	154.67	142.45	32.37	-13.31	-215.83	-259.88
CHEM	374.17	266.61	241.42	82.29	-117.45	-280.35	-266.33	-300.37
EART	58.51	49.65	-39.62	-36.30	-8.16	-11.41	1.22	-13.90
BIOL	50.84	-34.65	-20.77	33.15	-31.29	23.19	-19.31	-1.17
AGRI	113.60	73.48	-47.46	-5.57	-53.45	-55.83	-48.52	23.75
PSYC	-0.28	-337.90	-298.38	-212.06	-73.46	107.55	328.48	486.05
SOCI	-36.31	-92.74	-57.43	-14.81	38.29	-0.96	65.05	98.91
ECON	48.87	49.32	-1.03	-89.62	-20.42	23.20	-21.32	11.00
ANTH	-26.87	-78.38	-54.43	-49.01	-29.93	33.94	100.62	104.07
OTH	-150.04	-274.27	-234.74	-6.87	96.68	205.09	173.91	190.25

However, the same pattern as observed for 'Yugoslavia' etc. is shown by the other rows in this dataset, such as for 'Political situation' (and is even exhibited in the small artificial dataset on smoking used in § 6.2). We noted above that 'Personal economics' had little sample influence on g_{15}, for the category 'IFI', and this true for g_{25} as well. From the correspondence analysis displays this appears to be due to its central position so it has not moved so much in the rotation. The empirical again underestimates the large sample changes of 'Personal economics' but both agree on it being the most influential on the first four g_{1j} co-ordinates. The comparisons for the other rows are very close even though n (the number of rows) is only 8. The empirical gives 'Personal economics' as the most influential on g_{15}, but on the sample it is ranked seventh. The sample changes in the second dimension are large when 'Personal economics is omitted, but the empirical changes are not large as it ignores the rotation that has occurred by the perturbed, rather than the original eigenvalues, being close. This was seen to occur in principal component analysis, see Section 4.3, and would be similar in any method using eigenvectors.

The pattern of the sign of influence being determined by the sign of the residual and the side of the plot the row co-ordinate lies on, is also observed for the Doctoral dataset. The most influential rows are 'Chemistry' and 'Psychology' in the first dimension, and which comes out as most influential on a given g_{1j} co-ordinate coincides with which has the largest residual of the two for that category. We find that 'Anthropology', which has the largest absolute f_{1i} co-ordinate, has little influence on the g_{1j} co-ordinates. From expression (5.4.16) this again seems to be due to its low mass, since m multiplies a number of terms. The exact part the mass plays is not clear due to the complicated nature of the expression. Generally, if one of the rows

'Chemistry' or 'Psychology' is the most influential on the g_{1j} co-ordinate the other is the most influential on the g_{2j} co-ordinate. There are just two exceptions to this, where 'Engineering' is the most influential on the g_{2j} co-ordinate. Although the sign of influence does not always depend on the sign of the residual for the second dimension, it does on the most influential rows. The largest change in the first two dimensions is 'Chemistry' on '1960' with the latter falling from 0.155 to 0.030, to be placed between '1971' and '1972' on the first axis. 'Chemistry' has the only large positive residual on the '1960' category, see Table 6.3.6, with 'Agricultural sciences' next largest, and this explains why the effect is so large.

6.3.3. Influence on the F Co-ordinates

The comparisons between sample and empirical are similar to those for the G co-ordinates, with them very close on the smaller influences but differing in value but not usually in rank on the very large changes. We will not consider the comparisons further here. For the 'Protein Consumption' dataset the most extreme points in the first dimension are the most influential rows on the $\underline{f}_1$ co-ordinates, as on the $\underline{g}_1$ co-ordinates. Different rows come out as the most influential on different f_{1i} co-ordinates and there seems little pattern to which come out on which row co-ordinates. We will show this in the more simple second dimension which is dominated by 'Portugal'. 'Yugoslavia', 'Albania', 'Hungary' etc. which lie on the LHS of the plot have positive influences on the other row co-ordinates in the first dimension, irrespective of what side of the plot they lie on. Conversely, 'Norway' and 'Denmark' have negative influences on all the row co-ordinates when they are omitted. Since a positive influence represents a decrease in the co-ordinates, this means omitting 'Yugoslavia' etc. causes the origin of the first dimension to move to the right. A similar pattern to this was observed in Section 4.8 on

the covariance principal component scores. Thus, the affect of omitting a row differs on the G and F co-ordinates with the former taking differing signs of influence.

Portugal is the most influential row on all of the $\underline{f}_2$ co-ordinates except for 'Belgium/Luxembourg', 'Denmark', 'France' and the 'Netherlands' (and of course on itself) where it has ranked positions 2, 19, 2 and 18 respectively. The most influential rows on these co-ordinates are 'Norway', 'Ireland', 'Norway' and 'Norway' respectively. There is no obvious reason why these two rows come out as the most influential on the specific row co-ordinates above, since from the correspondence plot in Fig. 6.3.1 it is not just due to how close they are. From the Table of residuals in Table 6.3.4 we see, for example, that although 'Norway' and the 'Netherlands' are separated in the second dimension they have similar sized residuals on two categories. However, looking for patterns using such an approach is likely to be complicated. The largest influences on the f_{2i} co-ordinates in the second dimension are all for 'Portugal' and are,

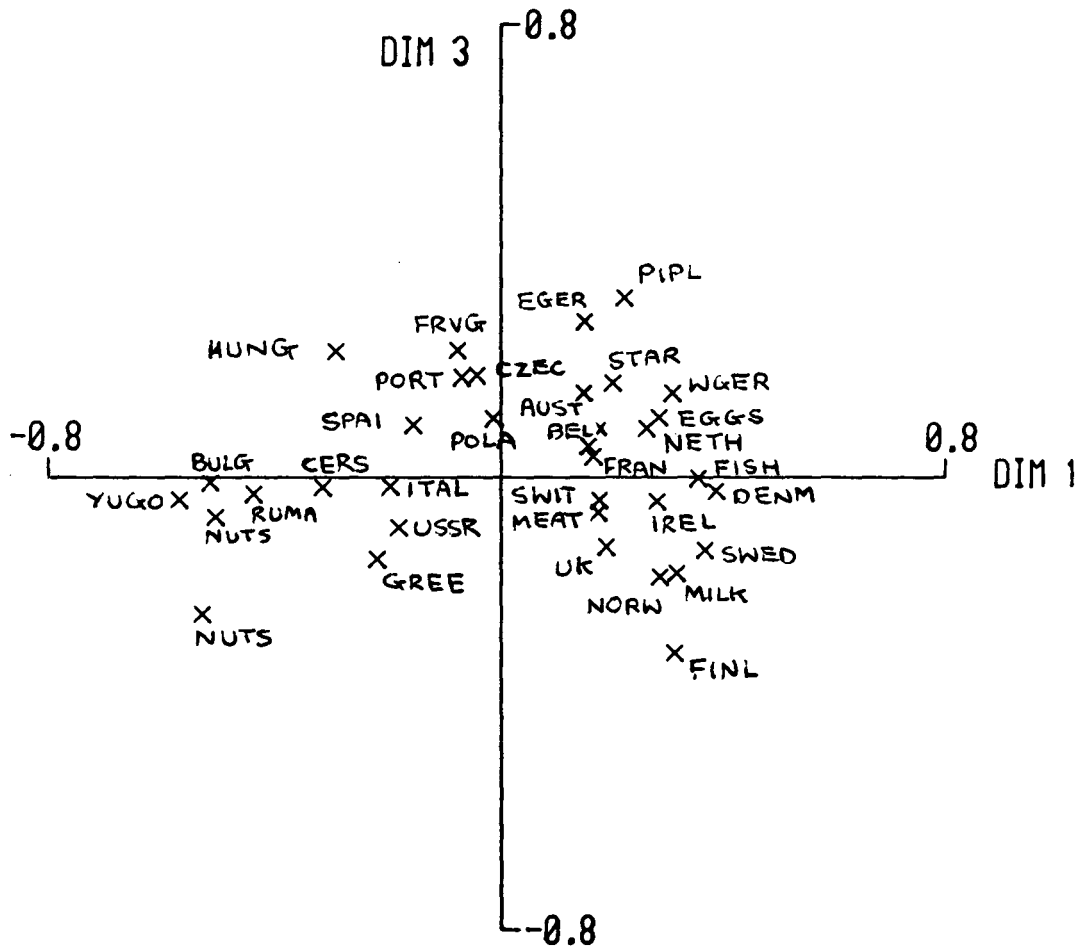
Country	f_{2i}	Sample Influence
Albania	-0.085	-0.216
East Germany	0.030	0.180
Finland	-0.042	-0.224
Spain	0.367	0.125
UK	-0.005	-0.106

Ten of the other influences for 'Portugal' lie between 0.07-0.09. The effect of 'Norway', for example, on 'Belgium/Luxembourg', 'France' and the 'Netherlands' are -0.022, -0.023 and -0.025 respectively, which are much smaller in comparison. The above changes for 'Portugal' are large compared to the sizes of the actual co-ordinates. We see from the above that Portugal's influences vary in sign unlike that noted for first dimension, and the sign of the influence for the above coincide with the sign of the co-ordinate. This is

not generally the case; there was little pattern to the signs of the influences for 'Portugal' or to which co-ordinates it had the largest effects on. From Fig. 6.3.1 and Fig. 6.3.4 we see that even though some of the influences for 'Portugal' are large, the relative positions of most of the rows remain unchanged. 'Finland' has moved noticeably upwards and 'East Germany' downwards towards 'Pigs and Poultry'. Fig. 6.3.5 is the plot of the third dimension against the first for the original dataset and we see here that 'Finland' is extreme in the third dimension at the opposite end to 'Pigs and Poultry' and 'East Germany' is very close to 'Pigs and Poultry'. The perturbed second dimension thus seems to have taken on some of the characteristics of the third dimension. See also the movements of 'UK', 'Starch', and 'Sweden'. We will return to this point in § 6.3.4 and § 6.3.5, where we consider the deletion of the 'Fish' category.

We have noted for the 'Israeli Worries' dataset that when 'Personal economics' is taken out the first two dimensions rotate. However, 'Personal economics' is only the most influential on 3/8 of the row co-ordinates (7/8 is the maximum as it cannot be most influential on itself) and 'Political situation' is the most influential on the other 5/8. For the rows where it is most influential 'Personal economics' is very influential, i.e. 0.277, 0.124 and 0.216 on 'Political situation', 'Economic situation.' and 'Other' respectively, whereas 'Political situation' has influences ranging from -0.045 to -0.061. The signs of influence for 'Political situation' are all negative and represents the origin of the first axis moving to the left. The influences for 'Personal economics' were usually positive but not always, this is caused by the rotation. 'Personal economics' is the most influential on all the other row co-ordinates in the second dimension. The influences are greater for the most extreme points in the first two dimensions, presumably because these will need to move

Figure 6.3.5 Plot of the First and Third Dimensions from the Correspondence Analysis of the Dataset on Protein Consumption in Europe.



further in the rotation than those in the middle, when 'Personal economics' is omitted.

For the 'Doctoral' dataset the most extreme point 'Anthropology', has little influence. An interesting pattern of influence occurred for the f_{1i} co-ordinates, which was that the two most influential rows on 'Chemistry', 'Earth', 'Agricultural', 'Psychology', 'Economics' and 'Anthropology', which all lie on the top of the second dimension were 'Other', with a positive influence, and 'Physics' with a negative influence. On the other row co-ordinates, which all lie beneath the axis, the two most influential rows were 'Psychology' with a positive influence and 'Chemistry' with a negative influence. For 'Other' and 'Psychology', the positive influences represent a decrease in the co-ordinates i.e. origin of the first dimension moving to the right, and vice versa for the negative influences. It is not clear why this occurs, and although this pattern is interesting the actual values of influence are not large. The most influential rows in the second dimension are 'Engineering', 'Chemistry' and 'Psychology' which tend to have influences representing the origin moving in the appropriate direction. Again, the influences are not particularly large. The only extreme co-ordinate in the second dimension is for the first column co-ordinate '1960' and we shall investigate the removal of this in § 6.3.5.

6.3.4. Scalar Measures of Influence

In a practical situation we would not usually wish to examine the influences on the individual coefficients, due to the numbers involved. We thus require some scalar measure of influence for the G and F co-ordinates in a given dimension, and having decided whether some rows are highly influential (using a gap test) we can investigate the individual changes by looking at the perturbed two dimensional correspondence analysis display.

From the previous work we see that different rows do not tend to come out on the G and F co-ordinates. Thus, we can use a scalar measure that combines both the G and F co-ordinates, which is the sum of squares of the changes in all the co-ordinates in a given dimension. As noted in PCA one needs to be careful in the sample case of a change in sign of the perturbed eigenvector. Table 6.3.7 gives the most influential rows in the first two dimensions, for the three datasets, using the sum of squares of the changes in both the G and F co-ordinates. The sums of squares are given in parentheses.

An alternative measure of influence, as in PCA, is the angle between the original and perturbed kth axes. Although the G and F co-ordinates are derived as eigenvectors, as noted in § 5.1.2 they are principal co-ordinates with respect to the principal axes A (in the chi-square metric D_r^{-1}) and B (in the chi-square metric D_c^{-1}) respectively. We can thus write,

$$G = (D_c^{-1}P' - \underline{1}\underline{r}')D_r^{-1}A \text{ where } A = D_r F D_{\lambda^{1/2}}^{-1} \qquad (6.3.3)$$

$$F = (D_r^{-1}P - \underline{1}\underline{c}')D_c^{-1}B \text{ where } B = D_c G D_{\lambda^{1/2}}^{-1} \qquad (6.3.4)$$

Details of this are omitted, see Greenacre (1984, p88-89). When omitting a row we can find the angle between the original $\underline{b}_k$ and the perturbed $\underline{\tilde{b}}_k$ (which has an empirical influence function that is similar to that for $\underline{g}_k$ due to expression (6.3.4)) but we cannot find the angle for $\underline{a}_k$ since from (6.3.3) we will have vectors of different lengths when we omit a row. However, one angle seems sufficient in indicating the influence of a row (this will be seen particularly in the next section). Table 6.3.8 gives the most influential observations in the first two dimensions, for the three datasets, when using the angular measure of change. The angles are given in parentheses. 'Portugal' and 'Personal economics' using either of the scalar measures in Table 6.3.7 or 6.3.8 stand out as highly influential. The influences in the 'Doctoral' dataset appear more important when using the angle rather than the sums of squares.

Table 6.3.7

Most Influential Observations using the Sums of Squares of Changes in the G and F Co-ordinates on the First Two Dimensions for Three Datasets.
SS of Sample Changes Given in Parentheses.

Dataset	1st Dim.		2nd Dim.	
'Protein Consumption'	Yugo	(0.031)	Port	(0.468)
	Bulg	(0.028)	Spain	(0.022)
	Alba	(0.028)	Norw	(0.017)
'Israeli Worries'	P.ecs	(0.397)	P.ecs	(0.202)
	Pol.sit	(0.023)	Mil	(0.044)
	Mil	(0.017)	Other	(0.033)
'Doctoral'	Chem	(0.022)	Chem	(0.016)
	Psych	(0.010)	Psych	(0.011)
	Other	(0.005)	Eng	(0.008)

Table 6.3.8

Most Influential Observations using the Angular Measure of Influence for the First Two dimensions of Three Datasets.
Sample Angular Measure Given in Parentheses.

Dataset	1st Dim.		2nd Dim.	
'Protein Consumption'	Alba	(3.34°)	Port	(56.45°)
	Finl	(3.09°)	Spain	(11.08°)
	Yugo	(2.97°)	Neth	(8.22°)
'Israeli Worries'	P.ecs	(48.51°)	P.ecs	(82.31°)
	Pol	(4.20°)	Pol	(10.65°)
	Mil	(4.06°)	Mil	(8.37°)
'Doctoral'	Chem	(13.21°)	Chem	(22.42°)
	Psych	(11.06°)	Psych	(14.94°)
	Other	(6.65°)	Eng	(12.00°)

The difference in influences for 'Chemistry' and 'Psychology' is greater in the first dimension using the sum of squares measure, and greater in the second dimension using the angle. The angle for 'Portugal' in the third dimension is 34.26° so there could be some rotation between the second and third dimensions, resulting in some of the similar positionings in the original third and perturbed second dimension, which were commented upon earlier (c.f. Fish in § 6.3.5). The row 'personal economics' in the 'Israeli' dataset has a larger angle in the second than the first dimension indicating that it is not just rotation between these two dimensions that has taken place. The angle for the third dimension is also large at 36.72° but not in the fourth dimension, so there could be rotation in the three dimensions. One can use the empirical influence function for $\underline{b}_k$ to obtain an estimate of this angle, which would be given by,

$$1 + \epsilon^2 \left[\frac{1}{2(\underline{b}_k' \underline{b}_k)^2} \left[\underline{b}_k' EIC(\underline{r}^*, \underline{b}_k) \right]^2 - \frac{1}{2(\underline{b}_k' \underline{b}_k)} EIC(\underline{r}^*, \underline{b}_k)' EIC(\underline{r}^*, \underline{b}_k) \right] \qquad (6.3.5)$$

$EIC(\underline{r}^*, \underline{b}_k)$ is obtained in exactly the same way as $EIC(\underline{r}^*, \underline{g}_k)$, see § 5.4. This differs from the estimate of the angle in PCA, see § 4.2, as our vectors do not have the same normalisation. Here, $B' D_c^{-1} B = I$ and not $B'B = I$. Expression (6.3.5) is again found to underestimate the actual angle. For example, it gives the angles 19.12° and 10.55° for the angles in the second and third dimensions when 'Portugal' are omitted. However, these are the largest angles for the empirical. It disagrees with the angle given for the sample on 'Personal economics' as the empirical ignores the rotation due to the close perturbed eigenvalues. Greenacre (1984, p213-214) quotes the form of the upper bounds for rotation of the principal axes, derived by Escofier and Le Roux (1979). These are similar to those given in § 3.8.4 for PCA, with n replaced by a function of the mass and z_{ki} by f_{ki}. These were found to be good in the

first dimension but poor in the later dimensions, especially if a row had affected an earlier dimension. For example, on 'Portugal' in the second and third dimensions it gives upper bounds of 31.23° and 0.97° respectively.

6.3.5. Influence when Deleting a Column

In all three datasets, considered in this Section we have extreme column co-ordinates in at least one of the dimensions, and we may be interested in the effects from deleting these categories. For the empirical algebra in § 5.4 we have considered the addition of rows, assuming $I > J$ since this does not lead to a change in the actual number of (non-trivial) dimensions, $J-1$, that exist. Since adding in a column leads to an extra dimension the normal theoretical influence techniques for an eigenvalue/eigenvector from a symmetric matrix, given in § 3.5, will not hold for the additional dimension since we need to multiply (3.5.2) by the original eigenvector which will not exist. However, we probably will not be actually interested in this extra added in dimension, and for deletion of columns the perturbed problem will have the smaller dimensions. We find that if we transpose our contingency table, so the columns are now the rows, and put this through the programs for omitting a row, we obtain good estimates of the actual sample change when we delete a column. We need to allow for the fact that the results will be less asymptotic than for deleting the rows as $J < I$. The empirical again gives zero influences for the trivial dimension when the column is omitted. Since we transpose the matrix, this means in the program the previous G co-ordinates are the F co-ordinates, and vice versa, and so influence expressions (5.4.16) and (5.4.20) now refer to the row and column influence functions respectively. The summation terms in equations (5.4.16) and (5.4.20) are taken over the original J (including the trivial dimension) dimensions and not the $J-1$ dimensions for the perturbed problem. Since the same empirical expressions

can be used for deleting columns as rows, this means we would expect the same patterns of influence. However, since $I > J$ the affect of deleting a column is likely to be greater since we are omitting a larger proportion of the data. We observed, particularly in the 'Doctoral' dataset, that the mass (or row sum) of the category, and not just the size of the co-ordinate, played an important role in determining the size of its influences. The masses for some, if not most, of the columns will be bigger than the row masses since $I > J$ and both sets of masses sum to one. This indicates that if a row and column co-ordinate were close in the original two dimensional display the effect of omitting the column could be much greater than for deleting the row. We will only examine column effects in the 'Protein Consumption ' dataset and the 'Doctoral' dataset. The three most influential columns in the first three dimensions of the 'Protein Consumption' dataset (we have considered three dimensions due to the rotations and swoppings) using the angular measure (given in parentheses) are given in Table 6.3.9.

Table 6.3.9

Most Influential Columns by the Change in the Principal Axis.

Dim.1		Dim.2		Dim.3	
Cereals	(57.22°)	Fish	(87.73°)	Fish	(83.14°)
Milk	(17.14°)	Cereals	(55.01°)	Pigs and Poultry	(60.66°)
Fish	(10.63°)	Milk	(17.14°)	Milk	(50.24°)

As expected these changes are much larger than for the row influences in the previous section.

The largest column co-ordinate in the first dimension is for 'Nuts' but this has only the seventh largest mass, with value 0.036, but 'Cereals' (which has the second largest co-ordinate) has a mass of 0.376, and 'Milk' has the second largest mass of 0.199. We thus find from the table above that 'Nuts' has little influence compared to 'Cereals' and 'Milk'. From expression (6.3.1) we know that the influence on the eigenvalue is directly affected by the mass.

'Cereals' has a smaller co-ordinate than the row co-ordinate 'Yugoslavia', but it leads to a 25% change in $\hat{\lambda}_1$ compared to the change for 'Yugoslavia' of 11.6%. The individual changes in the column and row co-ordinates are usually much bigger for 'Cereals' than 'Yugoslavia', but from the angles in Table 6.3.9 and the perturbed display in Fig 6.3.6 we see that all that has really occurred is rotation, with the relative positions of most of the categories remaining unchanged. Rotation has occurred to a smaller extent for 'Milk' in the first two dimensions and occurs for 'Pigs and Poultry' in the third and fourth dimensions.

Fish is the most extreme column co-ordinate in the second dimension and its sample influences of the actual change in the column co-ordinates in the second dimensions are,

g_{21}	g_{22}	g_{23}	g_{24}	g_{25}	g_{26}	g_{27}	g_{28}	g_{29}
-0.082	0.066	0.028	-0.278	-	-0.063	0.301	0.170	0.507

Comparing with the sample influences when 'Portugal' is omitted in § 6.3.2 we see both changes tend to have the largest influences on the same co-ordinates, but those for 'Fish' are usually greater. However, from the angles in Table 6.3.9, and by comparing the original correspondence display of the first and third dimensions in Fig. 6.3.5 with the plot of the first two dimensions when 'Fish' is omitted, see Fig. 6.3.7, we see that there has been a swop in the second and third dimensions. Since 'Fish' and 'Portugal' tend to have large influences on the same co-ordinates, but 'Portugals' were not large enough for a swop to occur, this seems to imply there was some rotation between the second and third dimensions when 'Portugal' was omitted.

The most extreme co-ordinate in the first two dimensions of the 'Doctoral' dataset is for the first column category '1960', see Fig. 6.3.2. When it is omitted $\hat{\lambda}_2$ decreases from 0.0033 to 0.0006, but we do not get any

Figure 6.3.6 Plot of the First Two Dimensions from the Correspondence Analysis of the Dataset on Protein Consumption in Europe When 'Cereals' is Omitted.

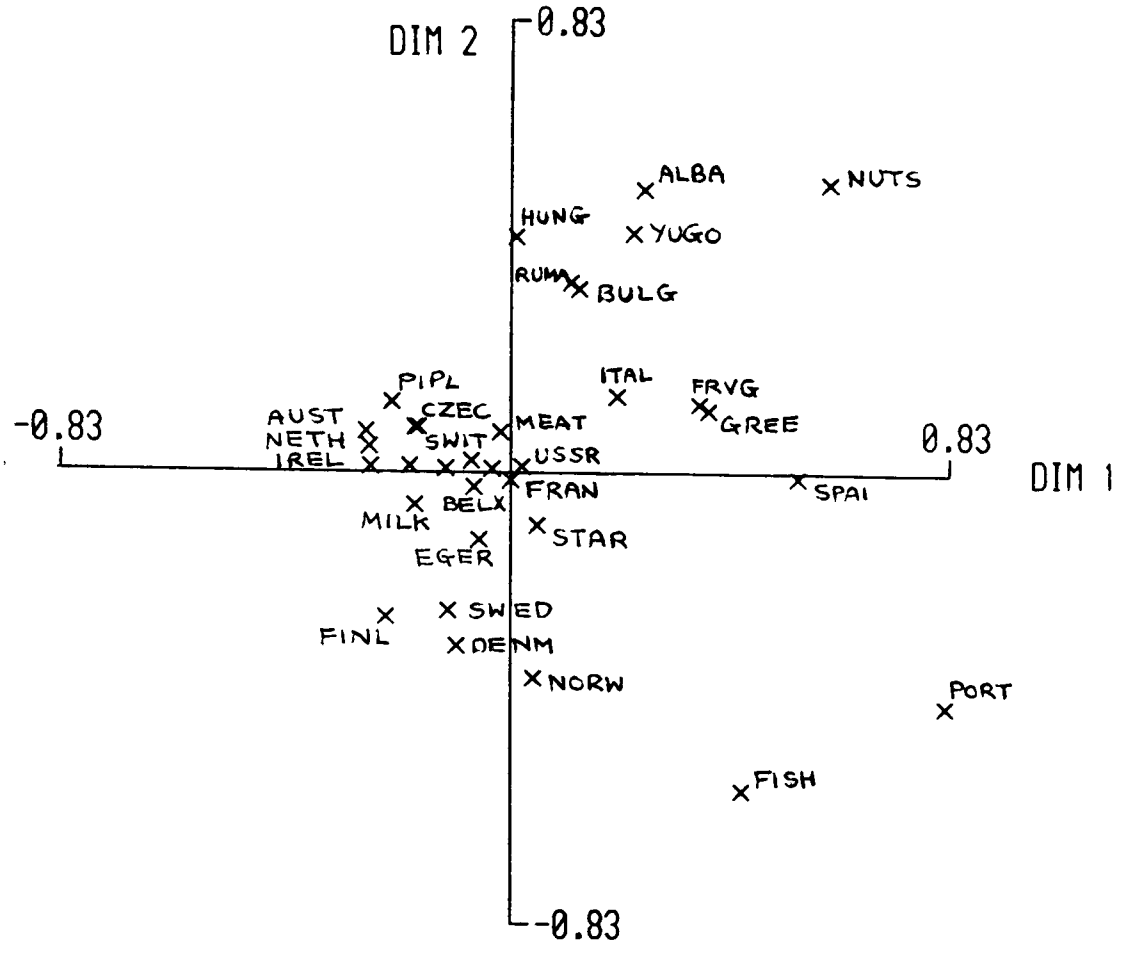
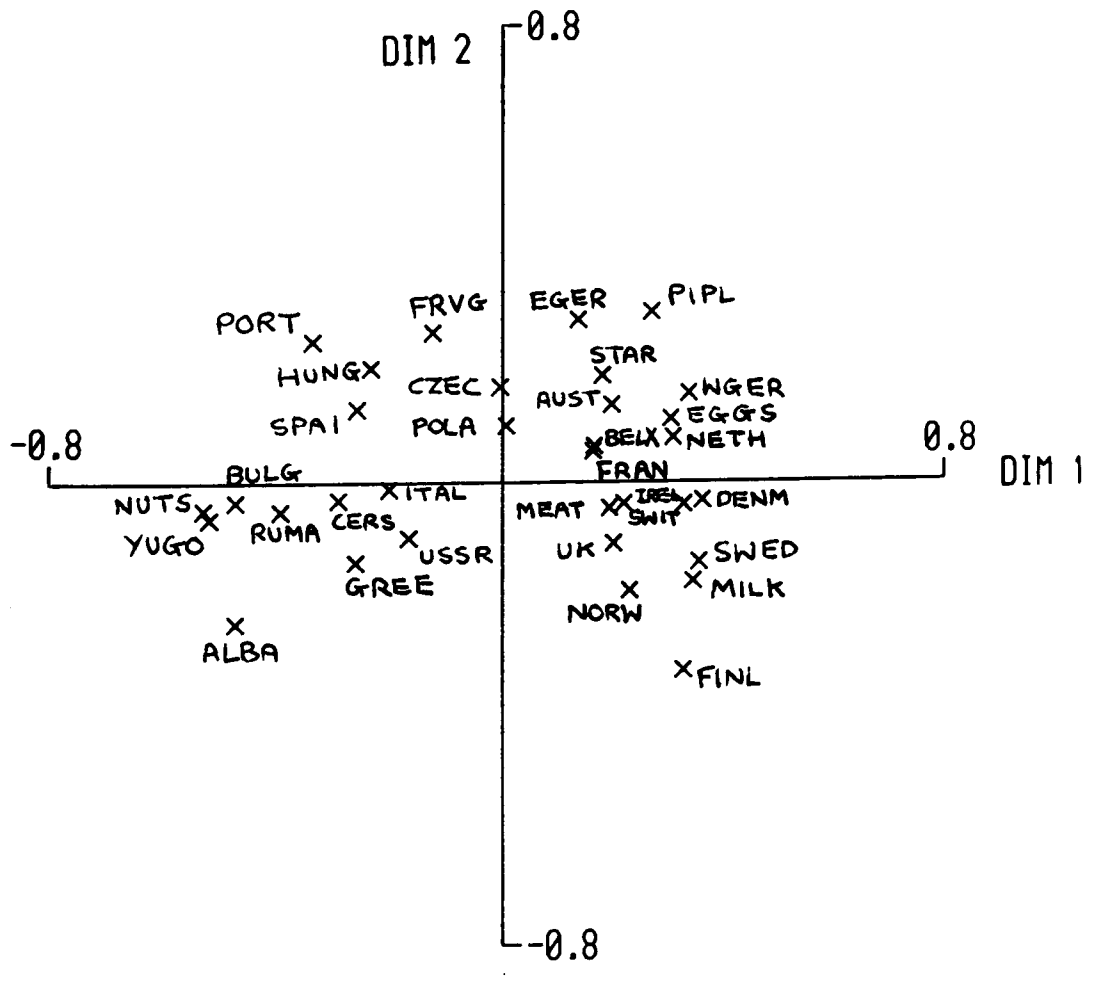


Figure 6.3.7 Plot of the First Two Dimensions from the Correspondence Analysis of the Dataset on Protein Consumption in Europe When 'Fish' is deleted.



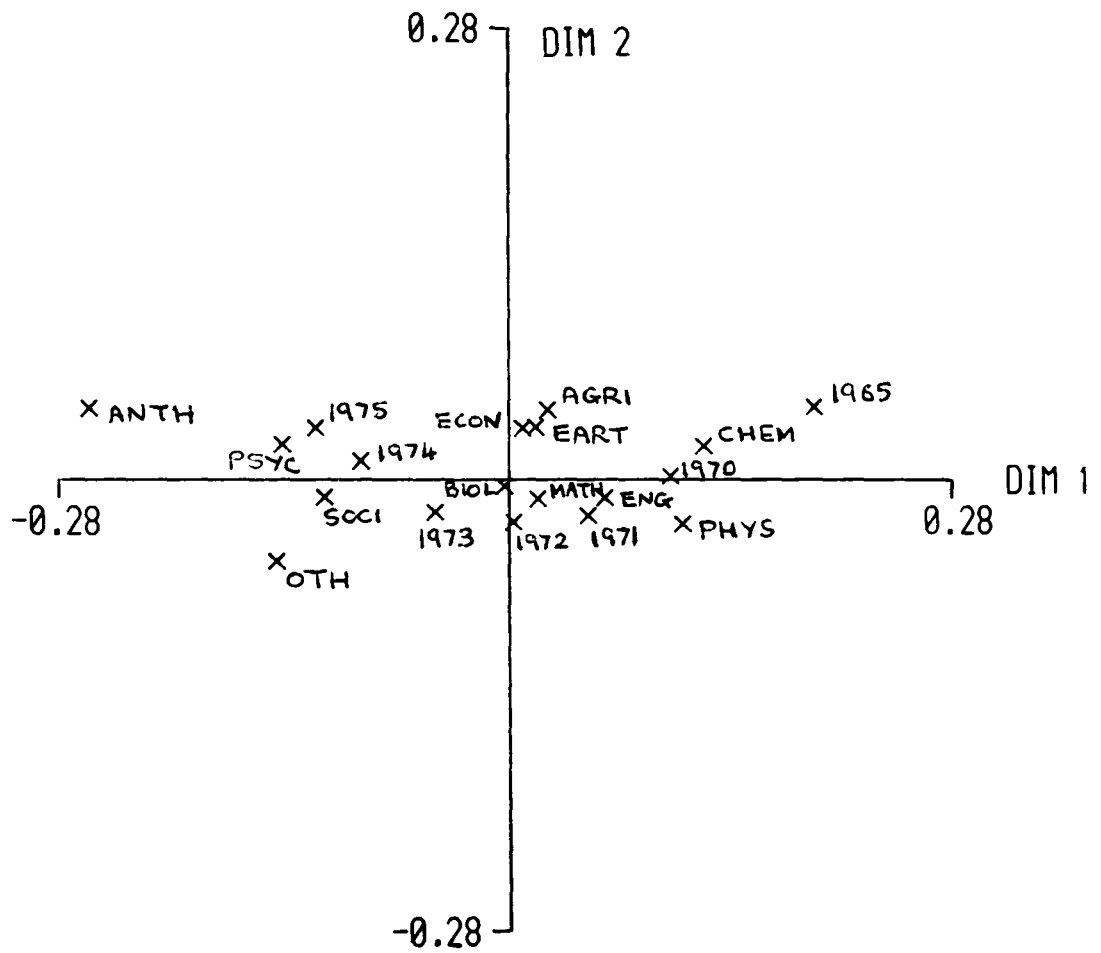
swopping. The category '1960' has a small mass compared to the other columns, and even seven of the rows have larger masses. However, its g_{21} co-ordinate is 0.220 compared to the next largest column co-ordinate of 0.045, and row co-ordinate of 0.083. Since both empirical and sample agree on the large change in $\hat{\lambda}_2$ this means the large co-ordinate has outweighed the low mass. However, we find '1960' is only the most influential column on 3 of the 7 column co-ordinates, and its influences compared to those for the row 'Chemistry' are,

	g_{21}	g_{22}	g_{23}	g_{24}	g_{25}	g_{26}	g_{27}	g_{28}
1960	-	-0.064	-0.040	-0.009	-0.008	0.003	0.001	0.013
Chem	0.004	-0.048	-0.011	-0.005	-0.004	-0.002	0.020	0.022

Presumably, the low mass has had some affect here since the influences for '1960' are not much larger than those for 'Chemistry' and not as large as one may initially think when looking at the original plot. '1960' is the most influential on 9 of the 12 row co-ordinates. Fig. 6.3.8 is the correspondence display when '1960' is omitted and we see how the second dimension has become squashed, and rows 'Engineering, 'Mathematics' and 'Anthropology' have moved up in there relative positions, and 'Other' has moved down. We thus find that the angle for the second dimension (the principal axes being derive in a similar way to F) is 41.84° and subsequent angles are also large.

It is interesting that despite such large angular and sums of squares changes, for omitting the above columns, few interpretations of the plots have changed. We have not observed some of the patterns for the columns as we did for the rows, but this is partly due to the rotations and swoppings. Certainly the greater the influence of a category the less these patterns may hold due to the complicated reappraisal of the relationships between the rows and columns that may occur.

Figure 6.3.8 Plot of the First Two Dimensions from the Correspondence Analysis of the Dataset on Science Doctorates in the USA When '1960' is Omitted.



6.3.6. Summary and Discussion

The eigenvector problems of rotation and swopping, observed in PCA, has also occurred in correspondence analysis, even though we have mostly only considered the first two dimensions. This occurred most when deleting the columns which can represent a large proportion (assuming $I > J$) of the data being removed. Rotations between the first and second dimensions can lead to large influences while the relative positions of the points in the two dimensional display remain virtually unchanged. However, if a swop has occurred between the second and third dimensions, the plot of just the first two dimensions can change dramatically due to the difference in the second and third dimensions. If $\hat{\lambda}_2$ and $\hat{\lambda}_3$ are close it is worth considering plots for the first three dimensions and not just the first two (of course, we may also consider other dimensions if the first two inertias do not account for most of the total inertia $\sum_{k=1}^{J-1} \hat{\lambda}_k$). It must be remembered that we also get rotational problems if the perturbed rather than the original eigenvalues are close. Using the angular measure of change, rather than the sums of squares of individual changes, provides the most useful information on whether rotation has taken place. Again, the swopping is highlighted by the empirical from looking at the changes in the eigenvalues. For example, in the 'Protein Consumption Dataset' $\hat{\lambda}_2 = 0.0390$, and $\hat{\lambda}_3 = 0.0200$, and when 'Fish' is omitted the empirical gives the perturbed eigenvalues as $\hat{\lambda}_2^* = 0.0180$, and $\hat{\lambda}_3^* = 0.0211$, which means that the second eigenvalue has fallen below the third. If rotation occurs because the perturbed rather than the original eigenvalues are close the empirical will not give the large changes of the sample angles.

When we do not have rotation and swopping of the eigenvectors we do observe the same patterns for the columns as for the rows, discussed in § 6.3.2 and § 6.3.3. This was not exhibited for the columns deleted in § 6.3.5, due to

the above problems of rotation, but was observed for the less extreme columns. That these patterns should hold for the columns as well follows from the fact the same empirical expressions have been used to obtain estimates of the sample changes when row and columns are omitted, with equally good results. The patterns for omitting the rows on the column and other row co-ordinates will be the same as the patterns for omitting columns on the row and other column co-ordinates respectively. For the most influential rows, the sign of influence on a column co-ordinate is linked to the sign of the residual, in the matrix $P - \underline{rc}'$, that the row has for that column, and to the sign of its own co-ordinate in the appropriate dimension. A row was usually the most influential on a given column co-ordinate when it was extreme in the same dimension and had a large residual for that column. The sign of influence for a row on the other row co-ordinates tends to be the same for all the co-ordinates and represents a shift in the center of the dimension. The mass of the row or column was also found to be important in determining whether a row/column was influential in the early dimensions. Extreme co-ordinates in a dimension will usually lead to a decrease in the variance (a positive influence) and expression (6.3.1) shows clearly the role played by the mass in determining the size of the positive term. The part played by mass on the G and F co-ordinates is shown to be much more complicated, but from expression (5.4.16) we can see it is important. The original row masses in the 'Protein Consumption' dataset are all roughly equal. An illustration of the importance of the mass of a row is given by dividing through the existing row by some number, L, which maintains the structure of the row but decreases its mass. Taking the sample angular measure of change, Table 6.3.10 gives the influences for 'Portugal', as L is varied, in the second dimension. Portugal did not have any large influences in the other dimensions either when its mass

Table 6.3.10

Angular Measure of Influence for the Second Dimension of the Dataset on Protein Consumption, and the Ranked Position, when Portugal is Omitted with its Mass Divided by L.

L	Angle	Rank
1	56.45°	1
2	39.94°	1
4	21.23°	2
8	9.99°	6
16	4.73°	12

Table 6.3.11

Two Most Influential Observations Using the Empirical and $EMP^{\bullet}$ for the Angular Influence Measure in the First Two Dimensions of the Dataset on Protein Consumption in Europe.

Dime.1				Dime.2			
EMP		$EMP^{\bullet}$		*EMP*		$EMP^{\bullet}$	
Country	Angle	Country	Angle	Country	Angle	Country	Angle
Finl	2.99°	Alba	2.98°	Port	19.12°	Port	19.26°
Alba	2.98°	Finl	2.48°	Spain	9.98°	Spain	7.15°

Table 6.3.12

Two Most Influential Observations Using the Empirical and $EMP^{\bullet}$ for the Angular Influence Measure in the First Two Dimensions for the Dataset on Doctorates in the USA

Dime.1				Dime.2			
EMP		$EMP^{\bullet}$		*EMP*		$EMP^{\bullet}$	
Science	Angle	Science	Angle	Science	Angle	Science	Angle
Chem	8.56°	Chem	8.46°	Chem	11.85°	Chem	11.76°
Psyc	6.91°	Psyc	6.89°	Psyc	10.15°	Psyc	9.95°

was small. Table 6.2.10 clearly shows the importance of mass on the co-ordinates.

The contingency tables examined in this section were not large but the empirical gave good comparisons of the sample changes. The two disagree more on the latter dimensions, largely due to the rotations, etc, that are treated differently by the two functions. We again observed that for the very large sample changes the empirical was smaller, but both usually agree on the same observations being the most influential, except where rotations may have taken place due to close perturbed eigenvalues. As the number of dimensions increase, as for all other eigenvector influence functions considered in this thesis, the empirical expressions for the G and F co-ordinates take longer to compute as expressions (5.4.16) and (5.4.20) involve summations over all dimensions. Thus, the empirical may take longer to compute than the sample function. Also, if the contingency table was not large it would not be time consuming to run through the sample influences. However, if we are only interested in the first two or three dimensions, we find that if we retain only the first four dimensions, then forming the empirical influences in the first two or three dimensions by summing over the four retained dimensions leaves the empirical influences virtually unchanged. This occurs as the summation terms involve $(\lambda_j - \lambda_k)^{-1}$, and as λ_j becomes smaller, with λ_k fixed, $(\lambda_j - \lambda_k)^{-1}$ decreases. The more distinct the first few dimensions are the less important the summations over the minor dimensions will be. Some examples of this will be given using the estimate of the angular change, when a row is omitted, given by (6.3.5) (the influence function for $\underline{b}_k$ involves similar summation terms to that for $\underline{g}_k$). The two most influential rows for the first two dimensions by the original empirical and EIC^*, which uses sums over the first four dimensions only, are given in Tables 6.3.11 and 6.3.12 for the 'Protein

Consumption' and 'Doctoral' datasets respectively. The empirical does underestimate the sample angles in Table 6.3.8 but the ranked order of rows is similar. EIC^* gives similar values to the estimated angle based on the full empirical function. The empirical expression for the eigenvalues is simple to calculate since it only involves terms from the given dimension.

An influence analysis to assess the affects of row and columns on the two dimensional correspondence display provides useful information, and reveals whether we can interpret the display with confidence. For the datasets examined in this section most of the large changes do seem to be caused by the rotation of the eigenvectors or swops.

6.4. Adding in an Extra Observation to Multiple Correspondence Analysis

The patterns of influence for adding in to a cell of a multiway table for a multiple correspondence analysis are similar to those when we add into a cell of a contingency table, discussed in § 6.2. This could occur since like the two way contingency table results the empirical expressions for the eigenvalues and eigenvectors from the Burt matrix just involve the co-ordinates of the categories involved in the cell that has been added to. See expressions (5.5.8) and (5.9.11). Also, since we can do a Burt analysis when we have two variables we would expect the same patterns of influence from adding to a given cell for the two types of correspondence analyses. The Burt matrix when $Q = 2$ has block diagonal matrices of the contingency table's row and column sums on its diagonal, and the contingency table and its transpose on the off diagonals. Adding to the (i,j)th cell of a contingency table results in changes to four entries of the Burt matrix. The contingency table eigenvalues and those from the Burt matrix have the relationship,

$$\lambda_k^C = 4(\sqrt{\lambda_k^B} - \frac{1}{2})^2 \quad , \tag{6.4.1}$$

and the $\underline{f}_k$ and $\underline{g}_k$ co-ordinates are rescaled versions of the Burt co-ordinates $\underline{h}_k$ such that,

$$\underline{f}_k = \left(\frac{\lambda_k^C}{\lambda_k^B}\right)^{1/2} \underline{h}_{k1} \quad \text{where} \quad \underline{h}_k = \begin{pmatrix} \underline{h}_{k1} \\ \underline{h}_{k2} \end{pmatrix} \tag{6.4.2}$$

is partitioned to hold the original order of row and column co-ordinates.

From expression (6.4.1) we find that the order of cells ranked by their influences (sample or theoretical) on the contingency table or Burt matrix eigenvalues must be the same (except if influences are close where rounding errors from the two methods may lead to some changes in order). The ranked absolute influences could only change if λ_k^B was close to 0.25 and adding to one cell lead to a decrease in λ_k^B below 0.25 and another cell lead to an increase in λ_k^B, due to the different gradients either side of a 1/4 for the plot of (6.4.1). However, only inertias above 1/4 from the Burt matrix are of interest, the rest are artifacts of the analysis, so this situation is not likely to occur. See Greenacre (1984, p144) for information on the artificial dimensions. The empirical influence function for $\hat{\lambda}_k^C$ in terms of $\hat{\lambda}_k^B$ is

$$EIC(\underline{x}, \hat{\lambda}_k^C) = 2\left(\frac{\hat{\lambda}_k^C}{\hat{\lambda}^B}\right)^{1/2} EIC(\underline{x}, \hat{\lambda}_k^B) \tag{6.4.3}$$

which shows that we must have the same ordering of cells for the two eigenvalues, when using this influence function. Expression (6.4.3) results in the empirical version of expression (5.3.1) but multiplied up by 4, this accounts for $n_{..}$ from the Burt matrix being 4 times greater than that from the contingency table.

The ranking of cells by their absolute influence on the co-ordinate for a given category can differ in the two analyses. From (6.4.2) the sample influence function for adding in an extra observation is

$$\frac{1}{n_{..}+1} SIC(\underline{x}, \underline{f}_k) = \left(\frac{\hat{\lambda}_k^C}{\hat{\lambda}_k^B}\right)^{1/2} \underline{h}_{k1} - \left(\frac{\hat{\lambda}_k^{C*}}{\hat{\lambda}_k^{B*}}\right)^{1/2} \underline{h}_{k1}^{*} \quad ,$$

where $\underline{h}_{k1}^{*}$ is the perturbed $\underline{h}_{k1}$. Thus, the rankings of cells on $\underline{f}_k$ and $\underline{h}_{k1}$ need not be the same as $\left(\frac{\hat{\lambda}_k^{C*}}{\lambda_k^{B*}}\right)^{1/2}$ will be different for each cell that we add to. However, it was found that the different rankings on the co-ordinates for the same category in the two analyses tended to occur for the smaller influences. Specifically, it occurred on categories with high row totals which we noted changed little when cells were perturbed in the contingency table analysis. The same patterns discussed in § 6.2, where the categories with low masses have the largest influences and the most influential cells involve the number of the row/column co-ordinate, apply to the Burt matrix correspondence analysis as well. Two examples of the changing in ranked influences on the co-ordinates when using the contingency table or Burt matrix analysis are given below, for the two datasets used in § 6.2. Noting that h_{13} refers to the same category as f_{13} we have as our examples,

(i) Israeli dataset

	h_{13}			f_{13}	
Cell	Influence	Rank	Cell	Influence	Rank
8,4	0.090	1	8,4	0.0032	3
3,1	0.080	2	3,1	0.0034	1

The largest changes on a single co-ordinate in the first dimension is cell (5,1) on h_{15} and f_{15} with influences 0.190 and 0.075 respectively, which are much larger than the influences above.

(ii) Smoking dataset

	h_{13}			f_{13}	
Cell	Influences	Rank	Cell	Influence	Rank
5,1	0.030	1	5,1	0.009	6
3,4	0.025	2	3,4	0.020	1

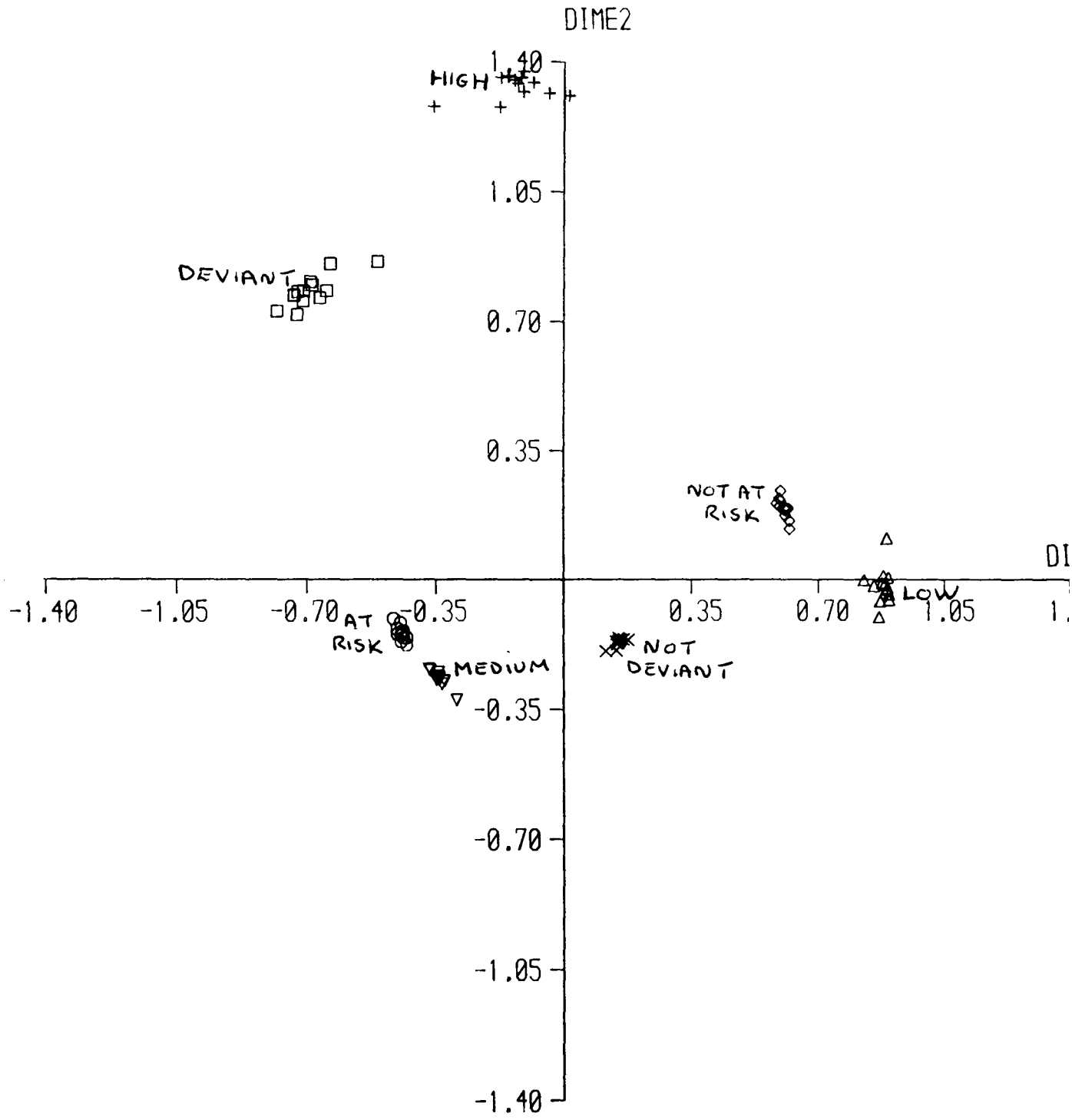
The difference in ranks is greater in this example than that above. However,

these influences are small compared to the affect of cell (1,1) on h_{11} and f_{11} with influences 0.257 and 0.111 respectively.

We will now consider an example where the number of variables, $Q = 3$. The data is taken from Everitt (1977, Chapter 4) and is concerned with the classroom behaviour of ten year old children. The children are classified into non-deviant or deviant and the other two variables account for the 'adversity of their school conditions' and whether they are at 'Risk' from their home conditions. The Burt matrix for this data is given in Table 6.4.1 and the original correspondence analysis plot in Fig. 6.4.1. The most influential cell (Low,Deviant,At Risk) leads to a decrease in $\hat{\lambda}_1$ which coincides with the patterns for $Q = 2$ as 'Low' is at the opposite end of the first dimension to the other two categories, and increasing the cell by one has led these categories to move inwards towards each other. We actually find that 'Low' and 'Deviant' move inwards but 'At Risk' does not. The most influential cell (High,Deviant,Not At Risk) leads to an increase in $\hat{\lambda}_2$ when one is added to it. This again coincides with the patterns observed for the contingency table eigenvalues in § 6.2.1 as these three categories have the three largest positive co-ordinates in the second dimension, and increasing their association makes them more extreme and so the variance increases. In fact, only 'Deviant' and 'Not At Risk' increase but 'High' decreases, perhaps to become closer to the other two co-ordinates as it was originally very extreme in the second dimension. The other co-ordinates move downwards as well, perhaps to keep away from the category 'High'.

We have low row sums for the three categories 'Low', 'High' and 'Deviant', and we find in the first two dimensions that their co-ordinates are the least steady. This shown by Fig. 6.4.2, which is a plot of the original and perturbed co-ordinates for the first two dimensions when we add to each cell

Figure 6.4.2 Plot of the First Two Dimensions from the Multiple Correspondence Analysis of the Dataset on Deviant Children When 1 is Added to Each Cell in Turn.



in turn. For the co-ordinates with low totals we find in the first dimension that it is cells involving their column/row number that give the largest influences. There is little pattern to influence on the categories with the large masses, but it is often cells involving categories extreme in the dimension that come out. However, the influences are small. These are all the same patterns as observed in § 6.2.

We shall not discuss adding in $m > 1$ or the comparisons between sample and empirical functions is this section. The comments on these topics are similar to those investigated in other situations. See, for example, Chapter 7 for a summary of these results.

Chapter 7: Summary of the Use of Influence Functions

The study of influence can be divided into two areas. The first is the derivation of expressions for the sample or theoretical/empirical influence functions based on terms from the original dataset or model. Such expressions will hopefully be interpretable and so reveal information on what type of observation is influential on that part of the analysis. This can also provide invaluable information on how to make our statistical analysis more robust (although this has not been considered in this thesis). The theoretical expressions for the eigenvectors revealed why we can get large changes in the eigenvectors corresponding to close eigenvalues. We also saw from the theoretical influence function that, provided the number of observations (m) deleted is not too large compared to n, then influence is reasonably additive in samples. This prevents the need to consider multiple block procedures, which can be very time consuming. However, this additivity will probably not hold if there are two extreme observations in one direction, so that when both are removed the dimension disappears. We observed, in § 4.8, how the empirical does not deal effectively with the changes when a dimension actually disappears. The second side to investigating influence is the detection of influential observations in datasets. This involves the numerical calculation of the influence functions for the individual observations. This provides detailed information on how the analysis may (or may not) alter due to changes in the structure of the data from the deletion of individual observations. This serves to increase our knowledge of the dataset that we are analysing. The observations found to be 'highly influential' maybe as interesting as the actual analysis, if they are not found to be recording errors etc. Thus, influence is not a way of discarding unusual observations since one should be noting the changes in the analysis caused by these observations

rather than just take the perturbed analysis as the end result. We have seen how the most extreme observations need not be the most influential, see for example § 2.3.3 on the bivariate correlation coefficient, or that observations found to outliers may not be influential on the part of analysis of interest but may be on others, see § 4.7. Influential observations may be easier to find than outliers in multivariate data and they provide the added information on the stability of our analysis.

We shall now consider which type of influence function one should use for the two areas of influence discussed above. It is preferable to obtain an expression for the sample influence function rather than the theoretical where possible, since the former will give the exact change in the analysis when points are omitted from the dataset. However, wé have seen a sample expression is not always possible, for example for the eigenvectors and eigenvectors, or the mathematics of the sample influence function can be intractable when the statistics/parameters of interest have a complicated form. For the theoretical influence function we can expand out square root signs etc, to $o(\epsilon)$ so we can write the perturbed parameter in terms of the original plus some higher order term. The theoretical influence function will help to give some insight into influence when a sample expression is not possible. Throughout this thesis we have seen that the theoretical can be used to describe influence in samples due to the generally good comparisons of the sample and empirical functions when calculated for individual observations. In Chapter 2 we examined how the sample and theoretical expressions compared when both had an algebraic form. The more complicated the initial expressions for the correlation coefficients are the greater the sample and empirical functions may differ, as we would have expanded out more terms to $o(\epsilon)$ for the theoretical expression. For the multiple correlation coefficient in §

2.4 we could see that the two types of functions would differ when the X s were remote in the factor space. We also observed how the empirical would usually underestimate influence when points are deleted and overestimate when observations are added. However, the asymptotic results were found to be reliable indicators of influence in samples, even for quite small sample sizes, for the correlation coefficients. In Chapter 3 we only had expressions for the actual sample change in the eigenvalues and eigenvectors when we add in points along one direction or in a plane. The theoretical/empirical expressions reflected these sample results, as well as giving insight into the more general case. The comparisons of sample and empirical were found to be good, particularly in the early dimensions where there is not any swopping or rotations of the eigenvectors. In Chapter 5 we had no sample results for the eigenvalues and eigenvectors in correspondence analysis. The empirical expressions for the eigenvectors were complicated and not easy to interpret, but those for the eigenvalues were found to be useful. The above shows that the theoretical expressions, when interpretated, can be useful for describing influence in samples, when no sample expressions exist.

In practice the sample or empirical influence function can be calculated. The former may involve the calculation of n extra separate analyses corresponding to the deletion of each observation, but the empirical just involves terms from the full dataset. If an algebraic expression can obtained for the sample influence function involving terms from the full analysis then this is what should be used to detect influential observations. When this is not possible we need to consider whether the recalculation of the n separate analyses will take much longer than using the empirical expressions. When calculating the sample influences for the eigenvalues and eigenvectors from the covariance (correlation) matrix we can use the deletion formula for the

covariance matrix given by (2.3.3) (and calculate the perturbed correlation matrix from this) and then recall the eigenvalue/eigenvector routines. With, for example, NAG (1982) routines this is simple to program. An alternative approach is to use the algorithm of Bunch, Nielson and Sorensen (1978) that updates the eigenvalues and eigenvectors of a symmetric matrix when the perturbed matrix is of the form $\tilde{B} = B + r\underline{bb}'$, where r is some constant. Only the perturbed covariance (sums of squares) matrix falls into this form. The updated eigenvalues are found by iteration and the eigenvectors can then found explicitly. There is some timed saved using this algorithm but it does not seem considerable as noted by Bunch *et al.* (1978) in their final section. This algorithm is also used in another paper by Bunch and Nielsen (1978) to update the singular value decomposition. The empirical influence functions for the eigenvalues and eigenvectors from the covariance matrix were very quick to calculate. No formal investigation of the number of operations needed to calculate the various influence functions has been done. For the eigenvalues and eigenvectors in PCA some investigation of CPU time has been examined. These revealed the empirical expressions to be much faster to calculate the sample influence functions for the covariance eigenvalues and eigenvectors but the times were more comparable for the eigenvalues and eigenvectors from the correlation matrix. The times taken to compute the eigenvalue and eigenvector sample influence functions from the correlation matrix are close to those for the covariance matrix as the computations only differ in the divisions required to obtain the perturbed correlation matrix from the perturbed covariance matrix. However, empirical expressions for the correlation influence functions are more complicated than corresponding ones for the covariance matrix. One advantage of the empirical over the sample is the ease to which we can look at any of the eigenvalues and eigenvectors.

Contingency tables may be small, in which case deleting the rows or columns individually to form the sample influence function may not be very time consuming. If there are a large number of dimensions and one is only interested in the first two then one disadvantage of the empirical expressions is that for the eigenvectors the functions involve summations over all the dimensions. This is a particular problem for eigenvectors from the Burt matrix where we have additional dimensions that are 'artifacts of the analysis', see Greenacre (1984, p144-145). One alternative to calculating the empirical expressions for the eigenvectors in correspondence or principal component analysis is to only consider terms in the summation for the dimensions closest to the one of interest. This was looked at in § 6.3.6 for the deletion of a row in correspondence analysis. These dimensions will be the most important due to the size of the terms $(\lambda_j - \lambda_k)^{-1}$ which become smaller the further λ_j is from λ_k. We can improve the estimate of the actual sample change given by the empirical influence function by considering second order terms. The second order terms will be more complicated than those for the first order and for the eigenvectors in particular would not be worth the computation time. If one wanted extra precision to the sample influence it is best to use the sample influence function itself.

In summary, influence is a valuable technique that can be applied to most statistical analyses. As well as detecting observations that are highly influential on our analysis, and so provide information on the reliability of our conclusions, it also adds to our understanding of the structure of the dataset.

References

Barnett, V. and Lewis, L. (1984). *Outliers in Statistical Data* (2nd Edition). Wiley, New York.

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

Belsley, D.A., Kuh, E. and Welch, R.E. (1980). *Regression Diagnostics.* Wiley, New York.

Benasseni, J. (1985). Influence des poids des unités statistiques sur les valeurs propres en analyse en composantes principales. *Revue de Statistiques Appliquées,* **23**, 41-55.

Benzécri, J.P. (1979). Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire. Addendum et erratum a [BIN.MULT]. *Cahiers de l'Analyse des Données* **4**, 377-378.

Bunch, J.R., Nielsen, C.P. and Sorensen, D.C. (1978). Rank-one modifications of the symmetric eigenproblem. *Numerische Mathematik,* **31**, 31-48.

Bunch, J.R., Nielsen, C.P. (1978). Updating the singular value decomposition. *Numerische Mathematik,* **31**, 111-129.

Calder, P., Jolliffe, I.T. and Morgan, B.J.T. (1986). Influential observations in principal component analysis: a case study. Submitted for publication.

Campbell, N.A. (1978). The influence function as an aid to outlier detection in discriminant analysis. *Applied Statistics,* **27**, 251-258.

Chernick, M.R. (1982). The influence function and its application to data validation. *American Journal of Mathematical and Management Sciences,* **2**, 264-288.

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall: London.

Critchley, F. (1985). Influence in principal component analysis. *Biometrika* , **72**, 627-636.

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

Escofier, B. and LeRoux, B. (1976). Influence d'un élément sur les facteurs en analyse des correspondances. *Cahiers de l'Analyse des Données,* **1**, 297-318.

Everitt, B.S. (1977). *The Analysis of Contingency Tables*. Chapman and Hall, London.

Gabriel, K.R. (1981). Biplot display of multivariate matrices for inspection of data and diagnosis. In *Interpreting Multivariate Data* (Ed. Barnett, V.), 147-174. Wiley, New York.

Gabriel, K.R. and Zamir, S. (1979). Lower rank approximations of matrices by least squares with any choice of weights. *Technometrics*, **21**, 489-498.

Goursat, E. (1933). Cours d'Analyse Mathématique. Tome 11 (fifth edition). Gauthier Villars, Paris.

Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, London.

Hampel, F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, **69**, 383-393.

Hawkins, D.M. and Fatti, L.P. (1984). Exploring multivariate data using the minor principal components. *The Statistician*, **33**, 325-338.

Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.

Jolliffe, I.T. (1972). Discarding variables in a principal component analysis, I: Artificial data. *Applied Statistics*, **21**, 160-173.

Jolliffe, I.T. (1986). *Principal Component Analysis*. Springer-Verlag, New York.

Krzanowski, W.J. (1984). Sensitivity of principal components. *Journal of the Royal Statistical Society, Series B*, **46**, 558-563.

Lin, S.P. and Bendel, R.B. (1985). Generation of population correlation matrices with specified eigenvalues. *Applied Statistics*, **34**, 193-198.

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

NAG (1982). *Numerical Algorithms Group Library Manual*. Numerical Algorithms Group, Oxford.

Radhakrishnan, R. and Kshirsagar, A.M. (1981). Influence functions for certain parameters in multivariate analysis. *Communications in Statistics*, **A10** , 515-529.

Rey, W.J. (1983). *Introduction to Robust and Quasi-Robust Statistical Methods*. Springer-Verlag, New York.

Sibson, R. (1979). Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling. *Journal of the Royal Statistical Society, Series B,* **41**, 217-229.

Snedecor, G.W. and Cochran, W.G. (1967). *Statistical Methods* (6th Edition). Iowa State University Press, Ames, Iowa.

Wilkinson, J.H. (1965). *The Algebraic Eigenvalue Problem.* Clarendon Press, Oxford.