Application of Transformations for Orthogonality

Thesis

Application of Transformations for Orthogonality

Md. Zillur Rahman Shabuz

BSc. (Hons.) & MSc.(University of Dhaka)

A thesis submitted for the Degree of

Doctor of Philosophy

School of Mathematics and Statistics

The Open University, UK

March 2018

Abstract

In the statistical analysis of multivariate data, principal component analysis is widely used to form orthogonal variables. Realizing the difficulties of interpreting the principal components, Garthwaite et al. (2012) proposed two transformations, each of which yield surrogates of the original variables. Recently, Garthwaite and Koch (2016) proposed a transformation that also produces orthogonal components and can be used to partition the contribution of individual variables to a quadratic form. The aim of this thesis is to discover and explore applications of these transformations.

We consider bootstrap methods for forming interval estimates of the contribution of individual variables to a Mahalanobis distance and their percentages. New bootstrap methods are proposed and compared with the percentile, bias-corrected percentile, non-studentized pivotal, and studentized pivotal methods via a large simulation study. The new methods enable use of a broader range of pivotal quantities than with standard pivotal methods, including vector pivotal quantities. Both equal-tailed intervals and shortest intervals are constructed; the latter are particularly attractive when (as here) squared quantities are of interest.

Using a transformation to orthogonality, new measures are constructed for evaluating the contribution of individual variables to a regression sum of squares. The transformation yields an orthogonal approximation of the columns of the predictor scores matrix. The new measures are compared with three previously proposed measures through examples, and the properties of the measures are examined.

We consider one new procedure and two older procedures for identifying collinear sets. The new procedure is based on transformations that partition variance infla-

tion factors into contributions from individual variables, and they provide detailed information about the collinear sets. The procedures are compared using three examples from published studies that addressed issues of multicollinearity.

Declaration

The research reported in this thesis describes original contributions of the author, except where due acknowledgement to other sources is made in the text and in the bibliography. No part of the material in this thesis has been submitted for a degree to this university or any other institution.

.....

Md. Zillur Rahman Shabuz

Acknowledgement

I would like to thank my lead supervisor Prof. Paul Garthwaite for his excellent supervision, continuous encouragement and support over the period of this work. I want to express my thanks to my co-supervisor Dr. Nickolay Trendafilov for his support during my period at the Open University. I am also thankful to Prof. Chris Jones, The Open University, UK for his advice about the sinh–arcsinh transformation.

I am grateful to the Open University for funding me for this research work. I am also grateful to the University of Dhaka, Bangladesh, for approving the study leave for doctoral degree. Many thanks go to The Charles Wallace Bangladesh Trust for small study grant. I am thankful to many of my friends and colleagues specially Asif, Dr. Fadlalla, Lax, Tsegay, Dr. Yonas, Sagor and Maha.

My deepest gratitude goes to my examiners Dr. Alvaro Faria, The Open University, UK and Prof. John Kent, University of Leeds, UK, for their valuable comments and suggestions.

Finally, my appreciation goes to my wife Nurunnahar Akter, my son Nehan Sadit, my parents and other family members for their support and sacrifice during my period in the UK.

Contents

Chapter 1

Introduction

Analysis of data and the interpretation of results is more straightforward if variables are independent or uncorrelated. However, in a real life situation many of the variables of interest are usually correlated. For example, in predicting the blood pressure of a person on the basis of age, weight, body surface area, duration of hypertension and basal pulse, the variables weight and body surface area are highly correlated. If the variables are correlated then it is less easy to analyze the data and implications are less transparent. For example, in multiple regression analysis, some of the parameter estimates will have large variances and covariances if some regressors are correlated (Gujarati, 2003, p.350). Also, determination of the relative importance of individual regressors in a regression analysis has widespread interest in many fields (Kruskal and Majors, 1989), but the assignment of relative importance becomes a challenging task when variables are correlated (Grömping, 2007) and, if there are near collinearities, the contribution of individual variables depends on the other variables of the model.

These problems do not arise with orthogonal variables, so transforming vari-

ables to yield orthogonal variables is attractive. This is one of the benefits of principal component analysis (PCA). PCA converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables (the principal components). The new variables are linear combinations of the original variables and the columns of the transformation matrix are the eigenvectors of the correlation matrix. The eigenvalues of the correlation matrix are the variance of the corresponding principal component. The first principal component has the property of having the largest possible variance of any linear combination of the original variables that is a unit length. The second component is orthogonal to the first component and has a larger variance than any other unit-length linear combination of the original variables, and so on.

Using these new variables (principal components) as regressors provides a model that is free from multicollinearity problems. Also the overall statistics from the new model does not change from the original model as the overall model is not affected by linear transformation (Freund et al., 2006, p.199–200). These advantages hold for any transformation of the original variables to orthogonal variables (components).

Unfortunately, it is often difficult, though not impossible, to discover the true interpretation of principal components since the new variables are linear combinations of the original variables (Freund et al., 2006, p.205), i.e., a component is typically associated with a number of the original variables and an original variable may be associated with more than one principal component. Also, the first principal component has a larger variance than the second component and the second component has a larger variance than the third component, and so forth. That is,

each principal component does not contribute equally to the total variability.

In this thesis we are interested in transformations that move the original variables by only a small amount but give orthogonal variables. The new variables are called surrogates of the original variables. Suppose $X_1, \ldots, X_p$ are the original set of correlated variables and let $\boldsymbol{x}_j$ be an $n \times 1$ vector of observations of X_j for $j = 1, \ldots, p$. Put $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$. A transformation is applied to $\mathbf{X}$ that yields an orthogonal data matrix $\mathbf{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p)$. The transformation is chosen so that each $\boldsymbol{z}_j$ is closely related to $\boldsymbol{x}_j$, where the definition of 'closeness' determines the optimal transformation. The variable Z_j whose data values are $\boldsymbol{z}_j$ is called the surrogate of X_j.

Constructing surrogates of X variables was first suggested in the context of regression by Gibson (1962). He proposed them as a method of evaluating the contribution of individual variables to regression model. Johnson (1966) also addressed the task of evaluating the contribution of individual variables to a regression model and independently introduced the same orthogonalization procedure. Their idea is to minimize the sum of the squared distance between $\boldsymbol{x}_j$ and $\boldsymbol{z}_j$ or, alternatively, to maximize the sum of the correlations between X_j and Z_j. The matrix $\mathbf{Z}$ is related to $\mathbf{X}$ by a linear transformation, $\mathbf{Z} = \mathbf{XA}$, and the square matrix $\mathbf{A}$ is referred to as the transformation matrix. With their transformation, the transformation matrix is closer to the identity matrix than any other transformation matrix used to orthonormalize the columns of $\mathbf{X}$ in the least square sense (Johnson, 1966). The only application considered by Johnson (1966) for the surrogate variables was in assessing the relative importance of individual variables to a regression model.

Recently Garthwaite et al. (2012) independently proposed two transformations, which they called the cos-max and cos-square transformations. The cos-max transformation is the same as the transformation proposed by Gibson (1962) and Johnson (1966), while the cos-square transformation can be shown to be a special case of a transformation of Bolla et al. (1998). [The transformation of Bolla et al. (1998) is described in Subsection 2.2.4.] The cos-max and cos-square transformations give orthogonal components with a one-to-one correspondence between the original vectors and the components, i.e., $\boldsymbol{z}_j$ links strongly to $\boldsymbol{x}_j$ but not to the other X vectors and vice-versa. The idea is to maximize the sum of the scalar products of $\boldsymbol{x}_j$ and $\boldsymbol{z}_j$ or the sum of the squares of the scalar products of $\boldsymbol{x}_j$ and $\boldsymbol{z}_j$. The transformations have different properties but typically give similar components. The cos-square transformation has an attractive *duplicate invariance property*. Suppose the set of vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ is increased by adding the set of vectors $\boldsymbol{x}_{k+1}, \ldots, \boldsymbol{x}_p$ where each of the vectors $\boldsymbol{x}_{k+1}, \ldots, \boldsymbol{x}_p$ is identical to $\boldsymbol{x}_k$. With the cos-square transformation, this duplication of $\boldsymbol{x}_k$ has no effect on the transformed values of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{k-1}$ (i.e $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{k-1}$ are unchanged). Recent work by Garthwaite and Koch (2016) implies that the cos-max transformation has a rotation invariance property. Suppose the first d columns of $\mathbf{X}$, $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)$, are rotated and the remaining $p - d$ columns $\boldsymbol{x}_{d+1}, \ldots, \boldsymbol{x}_p$ are not rotated. With the cos-max transformation, the rotation of the first d columns has no effect on $\boldsymbol{z}_{d+1}, \ldots, \boldsymbol{z}_p$. Rotation will only affect $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_d$ and the remaining $p-d$ columns remain unchanged by the rotation. Garthwaite et al. (2012) showed that the transformations have applications in collinearity diagnosis, setting prior probability in Bayesian model averaging, and in evaluating the upper bound of a multivariate

Chebyshev inequality.

Garthwaite and Koch (2016) adapted the cos-max transformation to yield a transformation that they called the corr-max transformation. The cos-max transformation transforms a data matrix while the corr-max transformation transforms a random vector. The corr-max transformation yields a vector $\boldsymbol{W} = (W_1, \ldots, W_p)^\top$ whose components are uncorrelated and W_j^2 is defined as the contribution of X_j to the quadratic form. Each of the original variables is associated with exactly one component of the transformed vector. Garthwaite and Koch (2016) used the corr-max transformation to partition a quadratic form and quantify the contribution of individual variables to the quadratic form. This decomposition is simple to implement and has a straightforward interpretation. It also has the rotation invariance property.

The three transformations mentioned above (the cos-max, cos-square and corr-max transformations) have uses in a variety of contexts, as illustrated in the applications mentioned above. This thesis stems from these transformations. The purpose of this project is to discover and explore applications of these transformations. An outline of this thesis is the following.

Detailed description of the methods used to transform the correlated variables to orthogonal variables are considered in Chapter 2. We also discuss further their properties and applications.

Rogers (2015) applied the corr-max transformation to identify the key predictor variables in determining the distributions of vector-borne diseases in the present and future. He kindly named the corr-max transformation as the Garthwaite–Koch partition and mentioned this transformation as a novel way of identifying

the most important predictors in predicting the presence or absence of a species' in an area. The partition yields point estimates of individual variables' contribution to a quadratic form but interval estimates of the contributions are also important. Garthwaite and Koch (2016) illustrated that bootstrap percentile intervals for these contributions are easily constructed but considered only one method and did not evaluate its performance. In Chapter 3, we consider four common bootstrap methods and propose two new methods for forming confidence intervals of individual contributions to a Mahalanobis distance and their percentages. We also compare their performances through a simulation study.

In Chapter 4, we consider the task of quantifying the contribution of individual variables to a multiple regression. This task was first addressed in the work reported in Gibson (1962) and Johnson (1966), in which the orthogonal counterparts of correlated regressors were derived and used to measure the relative importance of the regressors. Green et al. (1978) and Johnson (2000) also addressed this task and proposed measures they call relative importance measures. These are also based on orthogonal counterparts — the initial step in constructing relative importance measures is to determine the orthogonal counterparts of Gibson (1962) and Johnson (1966). A criticism of all these methods is that the relationship between the regressors and the criterion is ignored when deriving the orthogonal variables. This seems inappropriate as the purpose of the transformation is to evaluate the contribution of individual variables to the regression model. We propose three new measures of relative importance that are also based on the orthogonal counterparts of the original regressors. However, the new orthogonal counterparts are determined by maximizing the sum of the correlations

between the cross-product of individual regressors and the response variable and the cross-product of the orthogonal variables and the response variable. Hence the transformation is influenced by the relationship between the response and the regressors. The new methods are compared with the orthogonal counterparts measure of Gibson (1962) and Johnson (1966), relative weights measure of Johnson (2000) and a well-respected measure proposed by Budescu (1993), which is called the dominance analysis measure. Comparison is made through a simulation study and the analysis of real data.

In Chapter 5, we address the task of determining the existence of multicollinearity in a multiple regression and identifying the variables that cause each multicollinearity. Variance inflation factors (VIFs) are commonly used to examine whether collinearity is present, but it does not indicate the number of collinearities or which variables form each collinearity. Instead, if a collinearity is detected through VIFs, the most common procedure for determining which variables form the collinearity is to examine the eigenvectors that correspond to small eigenvalues of the correlation matrix. The larger elements of an eigenvector that correspond to a near zero eigenvalue identify those regressors that are most responsible for multicollinearity. However, there is no one-to-one relationship between a VIF and a particular eigenvector, so this method is not a well-integrated approach. An alternative is the regression coefficient variance-decomposition procedure of Belsley et al. (1980), which calculates the proportion of variance of $\widehat{\beta}_j$ (jth estimated regression coefficient) associated with the eigenvalues. However, for more than one collinearity, it is often difficult to identify separate collinear sets. Garthwaite et al. (2012) suggested that the transformation matrix of either the cos-max

transformation or the cos-square transformation could be used to identify them — they noted that rows of both transformation matrices have a one-to-one relationship with VIFs. The transformation matrices provide more information than the eigenvector-eigenvalue method. In Chapter 5, we consider three examples from published studies that identify collinear sets. The published results are compared with the output obtained using the other procedures.

Concluding comments and directions for further research are given in Chapter 6.

Chapter 2

Review of Literature

Orthogonality is important not only in geometry and mathematics, but also in science and engineering in general, and in data processing in particular. The aim of this chapter is to review work on transformations that yield new orthogonal variables that are close approximations of the original variables. The oldest of the transformations that we review was proposed by Johnson (1966). This is first described before describing two transformations proposed by Garthwaite et al. (2012), the cos-max transformation and the cos-square transformation. The cos-max transformation is essentially the same as Johnson's transformation but is derived from weaker assumptions. This chapter also includes a transformation proposed by Bolla et al. (1998) that addresses a more general problem and which relates to the cos-square transformation. Garthwaite and Koch (2016) develop a transformation, called the corr-max transformation, that is similar to the cos-max transformation. However, the corr-max transformation is used to transform a random vector while the cos-max transformation (in common with most of the methods described here) transforms a data matrix. Some properties and appli-

cations of the transformations are also described. The emphasize here is on the transformations that we will used in later chapters. Section 2.1 reviews Johnson's transformation for constructing surrogates of original variables, Section 2.2 reviews the cos-max, cos-square transformations and the transformation of Bolla et al. (1998), and Section 2.3 describes the corr-max transformation. Properties of the transformations are considered in Section 2.4 and some of their applications are described in Section 2.5.

2.1 Johnson's transformation

Suppose $\mathbf{X}$ is an $n \times p$ matrix with $n \geq p$ and is of full rank. Gibson (1962) and Johnson (1966) suggested a method of finding an orthogonal matrix $\mathbf{Z}$ whose columns are as similar as possible to the columns of the original matrix $\mathbf{X}$ in the least square sense. That is, $\mathbf{Z}$ is of order $n \times p$ and the p orthonormal vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$, each of dimension $n \times 1$, must be chosen such that $\sum_{j=1}^{p} (\boldsymbol{x}_j - \boldsymbol{z}_j)^\top (\boldsymbol{x}_j - \boldsymbol{z}_j)$ is minimized.

Johnson defines the problem being addressed more clearly than Gibson and we proceed along his lines. He assumed that $\mathbf{Z}$ is related to $\mathbf{X}$ through a linear transformation. Hence, he address the task of finding the $n \times p$ matrix $\mathbf{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p)$ that minimizes the residual sum of squares

$$\tau = tr\left\{ (\mathbf{X} - \mathbf{Z})^\top (\mathbf{X} - \mathbf{Z}) \right\} \tag{2.1}$$

such that

1. The transformation of $\mathbf{X} \to \mathbf{Z}$ is of the form $\mathbf{Z} = \mathbf{X}\mathbf{A}$, where $\mathbf{A}$ is a $p \times p$ transformation matrix; and

2. The transformed matrix $\mathbf{Z}$ is column orthogonal, i.e., $\mathbf{Z}^\top\mathbf{Z} = \mathbf{I}_p$.

The transformation matrix $\mathbf{A}$ and consequently the transformed orthogonal matrix $\mathbf{Z}$ can be obtained using the singular value decomposition of $\mathbf{X}$. The singular value decomposition of $\mathbf{X}$ is

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^\top \tag{2.2}$$

where the p columns of $\mathbf{U}$ are the first p orthonormal eigenvectors of $\mathbf{X}\mathbf{X}^\top$ (if eigenvalues are arranged in descending order), $\boldsymbol{\Delta}$ is a $p \times p$ diagonal matrix with diagonal elements equal to the square roots of the first p (descending order) common eigenvalues of $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}\mathbf{X}^\top$ (common eigenvalues are basically the eigenvalues of $\mathbf{X}^\top\mathbf{X}$), and the columns of $\mathbf{V}$ are the orthonormal eigenvectors of $\mathbf{X}^\top\mathbf{X}$. The diagonal elements of $\boldsymbol{\Delta}$ are unique and usually considered as positive, called the singular values of $\mathbf{X}$. If $\mathbf{X}$ is of full rank, eigenvalues of $\mathbf{X}^\top\mathbf{X}$ will never be zero. Here $\mathbf{U}$ is column orthogonal and $\mathbf{V}$ is both row and column orthogonal, i.e., $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^\top\mathbf{V} = \mathbf{V}\mathbf{V}^\top = \mathbf{I}$.

The problem is to find the transformation matrix $\mathbf{A}$ by minimizing

$$\tau = tr\left\{(\mathbf{X} - \mathbf{Z})^\top (\mathbf{X} - \mathbf{Z})\right\} \tag{2.3}$$

under the conditions that $\mathbf{Z}^\top\mathbf{Z} = \mathbf{I}_p$, $\mathbf{Z} = \mathbf{X}\mathbf{A}$ and $\boldsymbol{x}_j^\top\boldsymbol{z}_j > 0$ for $j = 1, \ldots, p$. Expanding $\tau = tr\left\{(\mathbf{X} - \mathbf{Z})^\top(\mathbf{X} - \mathbf{Z})\right\}$ and substituting $\mathbf{Z} = \mathbf{X}\mathbf{A}$ gives

$$\tau = tr\left(\mathbf{X}^\top\mathbf{X}\right) - 2tr\left(\mathbf{X}^\top\mathbf{X}\mathbf{A}\right) + tr\left(\mathbf{A}^\top\mathbf{X}^\top\mathbf{X}\mathbf{A}\right). \tag{2.4}$$

The first term of the right hand side of equation (2.4) is independent of $\mathbf{A}$ and the last term is a constant p, since $\mathbf{A}^\top\mathbf{X}^\top\mathbf{X}\mathbf{A} = \mathbf{Z}^\top\mathbf{Z} = \mathbf{I}_p$. Thus for evaluating $\mathbf{A}$, the minimization of τ is equivalent to the maximization of

$$\eta = tr\left(\mathbf{X}^\top\mathbf{X}\mathbf{A}\right) \tag{2.5}$$

subject to the constraint that $\mathbf{Z}^\top\mathbf{Z} = \mathbf{I}_p$, $\mathbf{Z} = \mathbf{XA}$ and $\boldsymbol{x}_j^\top \boldsymbol{z}_j > 0$ for $j = 1, \ldots, p$ (Johnson, 1966). Substituting $\mathbf{X} = \mathbf{U\Delta V}^\top$ into equation (2.5) gives

$$\eta = tr\left(\mathbf{V\Delta U}^\top\mathbf{U\Delta V}^\top\mathbf{A}\right) = tr\left(\mathbf{V\Delta^2 V}^\top\mathbf{A}\right) = tr\left(\mathbf{\Delta^2 V}^\top\mathbf{AV}\right) \tag{2.6}$$

since the trace of a product of matrices is independent of the order of multiplication.

Let $\mathbf{M} = \mathbf{\Delta V}^\top\mathbf{A}$. Then $\mathbf{M}$ is an orthogonal matrix as $\mathbf{M}^\top\mathbf{M} = \mathbf{A}^\top\mathbf{V\Delta\Delta V}^\top\mathbf{A}$ $= \mathbf{A}^\top\mathbf{V\Delta U}^\top\mathbf{U\Delta V}^\top\mathbf{A} = \left(\mathbf{U\Delta V}^\top\mathbf{A}\right)^\top\left(\mathbf{U\Delta V}^\top\mathbf{A}\right) = (\mathbf{XA})^\top(\mathbf{XA}) = \mathbf{Z}^\top\mathbf{Z} = \mathbf{I}_p$. So equation (2.6) becomes

$$\eta = tr\left(\mathbf{\Delta MV}\right). \tag{2.7}$$

Let $\mathbf{T} = \mathbf{MV}$. Then $\mathbf{T}$ is an orthogonal matrix as the product of two orthogonal matrices is orthogonal matrix. Thus equation (2.7) becomes

$$\eta = tr\left(\mathbf{\Delta T}\right). \tag{2.8}$$

The diagonal elements of $\mathbf{\Delta}$ are positive. So $tr\left(\mathbf{\Delta T}\right)$ can be maximize by choosing $\mathbf{T}$ to have maximum diagonal elements with positive sign. The elements of an orthogonal matrix cannot be greater than one. Hence one can choose $\mathbf{T} = \mathbf{I}$ to maximize $tr\left(\mathbf{\Delta T}\right)$.

Now $\mathbf{T} = \mathbf{MV} = \mathbf{\Delta V}^\top\mathbf{AV} = \mathbf{I}$ implies $\mathbf{A} = \mathbf{V\Delta}^{-1}\mathbf{V}^\top$. Replacing $\mathbf{X}$ by $\mathbf{U\Delta V}^\top$ and $\mathbf{A}$ by $\mathbf{V\Delta}^{-1}\mathbf{V}^\top$ gives

$$\mathbf{Z} = \mathbf{XA} = \mathbf{U\Delta V}^\top\mathbf{V\Delta}^{-1}\mathbf{V}^\top = \mathbf{UV}^\top. \tag{2.9}$$

Now $\mathbf{Z}^\top\mathbf{Z} = \mathbf{VU}^\top\mathbf{UV}^\top = \mathbf{VV}^\top = \mathbf{I}$, verifying the constraint of orthonormal columns. Finally, $\mathbf{Z} = \mathbf{UV}^\top$ is the least-square orthonormal approximation of $\mathbf{X}$. Johnson (1966) has shown that the transformation matrix $\mathbf{A} = \mathbf{V\Delta}^{-1}\mathbf{V}^\top$ is closer

to the identity matrix (in the least-square sense) than any other transformation matrix that orthonormalizes the columns of $\mathbf{X}$. Hence the columns of $\mathbf{Z}$ are a close approximation to the columns of $\mathbf{X}$, i.e., column vectors of $\mathbf{Z}$ are surrogates of the column vectors of $\mathbf{X}$.

When the columns of $\mathbf{X}$ have been standardized to have means of 0 and unit lengths, then Gibson's method and Johnson's (1966) method yield the same orthogonal components. Gibson (1962) describes the set of orthogonal factors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ as having the highest degree of one-to-one correspondence with $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$.

The purpose of finding the orthogonal matrix $\mathbf{Z}$ in both Gibson and Johnson's transformations was to find the contribution of regressors to a multiple regression model. In the remainder of this section, suppose $\boldsymbol{y}$ (response) and the columns of $\mathbf{X}$ (regressors) have been centered and scaled to have sample means of zero and unit lengths. Define $\widehat{\boldsymbol{\beta}}_Z$ as the vector of regression coefficients when regressing $\boldsymbol{y}$ on $\mathbf{Z}$, so

$$\widehat{\boldsymbol{\beta}}_Z = \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \boldsymbol{y} = \mathbf{Z}^\top \boldsymbol{y}. \tag{2.10}$$

The jth element of $\widehat{\boldsymbol{\beta}}_Z$, $\widehat{\beta}_{Z_j}$, is the beta weight of Z_j, and it is also the correlation coefficient between Y and Z_j. Gibson (1962) and Johnson (1966) suggested $\widehat{\beta}_{Z_j}$ as the *importance* weight of X_j, as each Z variable is paired with an X variable, and $\widehat{\beta}^2_{Z_j}$ is the proportion of the variation in Y that is explained by Z_j.

Instead of using the singular value decomposition of $\mathbf{X}$ to obtain the orthogonal matrix $\mathbf{Z}$, Gibson (1962) and Garthwaite et al. (2012) independently used the symmetric square-root matrix of the correlation matrix $\mathbf{R}_{xx}$. The orthogonal matrix $\mathbf{Z}$ is obtained from

$$\mathbf{Z} = \mathbf{X}\mathbf{R}_{xx}^{-1/2}, \tag{2.11}$$

where $\mathbf{R}_{xx} = \mathbf{X}^\top\mathbf{X} = \mathbf{V}\boldsymbol{\Delta}\mathbf{U}^\top\mathbf{U}\boldsymbol{\Delta}\mathbf{V}^\top = \mathbf{V}\boldsymbol{\Delta}^2\mathbf{V}^\top$ and hence the square-root and inverse square-root matrices of $\mathbf{R}_{xx}$ are $\mathbf{R}_{xx}^{1/2} = \mathbf{V}\boldsymbol{\Delta}\mathbf{V}^\top$ and $\mathbf{R}_{xx}^{-1/2} = \mathbf{V}\boldsymbol{\Delta}^{-1}\mathbf{V}^\top$. Finally, $\widehat{\boldsymbol{\beta}}_Z$ can be expressed as

$$\widehat{\boldsymbol{\beta}}_Z = \mathbf{Z}^\top\boldsymbol{y} = \mathbf{V}\mathbf{U}^\top\boldsymbol{y} = \mathbf{V}\boldsymbol{\Delta}^{-1}\mathbf{V}^\top\mathbf{V}\boldsymbol{\Delta}\mathbf{U}^\top\boldsymbol{y} = \mathbf{R}_{xx}^{-1/2}\mathbf{X}^\top\boldsymbol{y} = \mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy}. \quad (2.12)$$

The sum of squared beta weights of the Z variables is equal to the proportion of variation in Y that is explained by $\boldsymbol{X}$ (or $\boldsymbol{Z}$). This is because the proportion of variation in Y explained by a multiple regression model is

$$R^2 = \boldsymbol{y}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{y} = \mathbf{R}_{xy}^\top\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}, \quad (2.13)$$

and

$$\widehat{\boldsymbol{\beta}}_Z^\top\widehat{\boldsymbol{\beta}}_Z = \mathbf{R}_{xy}^\top\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xx}^{-1/2}\mathbf{R}_{xy} = \mathbf{R}_{xy}^\top\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy} = R^2. \quad (2.14)$$

Unexpectedly, the correlation matrix between the X variables and the Z variables, $\mathbf{R}_{xz}$, is symmetric and is, in fact, the symmetric square-root matrix of $\mathbf{R}_{xx}$ (Johnson, 1966). The symmetry of $\mathbf{R}_{xz}$ is verified by

$$\mathbf{R}_{xz} = \mathbf{X}^\top\mathbf{Z} = \mathbf{V}\boldsymbol{\Delta}\mathbf{U}^\top\mathbf{U}\mathbf{V}^\top = \mathbf{V}\boldsymbol{\Delta}\mathbf{V}^\top = \mathbf{R}_{xx}^{1/2}. \quad (2.15)$$

Using an example, J. W. Johnson (2000) argued that, when the original X variables are highly correlated, the measure proposed by Gibson (1962) and Johnson (1966) does not adequately measure the relative importances of the X variables.

Let λ_{jk} denote the correlation between X_j and Z_k. Then the variation in X_j that is explained by Z_k is λ_{jk}^2. As the Z variables are orthogonal, $\sum_{k=1}^{p}\lambda_{jk}^2$ is the variation in X_j that is explained in a multiple linear regression of X_j on $Z_1, \ldots, Z_p$. But $\mathbf{Z}$ is a linear transformation of $\mathbf{X}$, and vice-versa, so $Z_1, \ldots, Z_p$ explain all the variance in each X variable, including X_j. And consequently, it follows that

$\sum_{k=1}^{p} \lambda_{jk}^2 = 1$. Hence, the symmetry given by equation (2.15) implies that

$$\sum_{j=1}^{p} \lambda_{jk}^2 = 1. \tag{2.16}$$

The proportion of the variation in Y that is explained by Z_k is $\widehat{\beta}_{Z_k}^2$. J. W. Johnson (2000) suggested that $\widehat{\beta}_{Z_k}^2$ should not be allocated solely to X_k, but should be divided between the X variables to reflect (the square of) their correlation with Z_k. That is, he suggested the proportion of $\widehat{\beta}_{Z_k}^2$ that is allocated to X_j should be $\lambda_{jk}^2 \widehat{\beta}_{Z_k}^2$. In total, X_j is allocated $\sum_{k=1}^{p} \lambda_{jk}^2 \widehat{\beta}_{Z_k}^2$ from $\widehat{\beta}_{Z_1}^2, \ldots, \widehat{\beta}_{Z_p}^2$, and he defined the relative importance of X_j as:

$$\text{Relative importance of } X_j : \sum_{k=1}^{p} \lambda_{jk}^2 \widehat{\beta}_{Z_k}^2. \tag{2.17}$$

From equation (2.16), $\sum_{j=1}^{p} \lambda_{jk}^2 = 1$, so $\sum_{j=1}^{p} \sum_{k=1}^{p} \lambda_{jk}^2 \widehat{\beta}_{Z_k}^2 = \sum_{k=1}^{p} \left[\sum_{j=1}^{p} \lambda_{jk}^2\right] \widehat{\beta}_{Z_k}^2$ $= \sum_{k=1}^{p} \widehat{\beta}_{Z_k}^2$. Thus equation (2.17) assigns relative importances to $X_1, \ldots, X_p$ that sum to the variance in Y that is explained by $X_1, \ldots, X_p$ in a multiple regression.

2.2 Cos-max and cos-square transformations

Suppose $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ is a set of $n \times 1$ observed vectors of the variables $X_1, \ldots, X_p$ and let $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$. Garthwaite et al. (2012) suggested two methods for obtaining orthonormal components $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ that have a one-to-one correspondence with the original vectors and are close to them, i.e., each component is closely related to a single X variable and each X variable is related to a single component. That is the vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ are surrogates of the original data vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$.

The $n \times p$ matrix $\mathbf{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p)$ is chosen to maximize either

$$\psi = \sum_{j=1}^{p} \boldsymbol{x}_j^\top \boldsymbol{z}_j \tag{2.18}$$

or

$$\phi = \sum_{j=1}^{p} \left(\boldsymbol{x}_j^\top \boldsymbol{z}_j\right)^2 \tag{2.19}$$

under the following conditions:

Condition 1. *The transformed matrix* $\mathbf{Z}$ *is column orthogonal, i.e.,* $\mathbf{Z}^\top\mathbf{Z} = \mathbf{I}_p$.

Condition 2. *The relationship between* $\boldsymbol{x}_j$ *and* $\boldsymbol{z}_j$ *is positive, i.e.,* $\boldsymbol{x}_j^\top \boldsymbol{z}_j > 0$ *for* $j = 1, \ldots, p$.

Unlike Gibson (1962) and Johnson (1966), they do not assume that $\mathbf{Z}$ is a linear transformation of $\mathbf{X}$, but show this must be the case. That is, $\mathbf{Z} = \mathbf{XA}$, where $\mathbf{A}$ is the transformation matrix. The transformation of $\mathbf{X} \to \mathbf{Z}$ is called the cos-max transformation when ψ is maximized and when ϕ is maximized it is called the cos-square transformation. If the columns of $\mathbf{X}$ are normal vectors, i.e., $\boldsymbol{x}_j^\top \boldsymbol{x}_j = 1$, then $\boldsymbol{x}_j^\top \boldsymbol{z}_j$ is the cosine of the angle between $\boldsymbol{x}_j$ and $\boldsymbol{z}_j$ for each $j = 1, \ldots, p$. The higher the closeness between $\boldsymbol{x}_j$ and $\boldsymbol{z}_j$ the higher the magnitude of the product $\boldsymbol{x}_j^\top \boldsymbol{z}_j$. Hence the objective is to determine orthonormal vectors by maximizing the sum of the cosines or their squares. That justifies the names of the transformations. In addition, if the columns of $\mathbf{X}$ and $\mathbf{Z}$ are standardized to have zero means and unit lengths, then $\boldsymbol{x}_j^\top \boldsymbol{z}_j$ is the correlation between $\boldsymbol{x}_j$ and $\boldsymbol{z}_j$. So maximization of ψ or ϕ means maximization of the sum of the correlations between the original vectors and the orthonormal vectors or the sum of the squared correlations.

For both transformations, the transformation matrix is determined by $\mathbf{X}^\top\mathbf{X}$, rather than $\mathbf{X}$ itself. $\mathbf{X}^\top\mathbf{X}$ can be replace by the covariance matrix ($\mathbf{X}$ is centred so that each of its columns has a mean of zero) or by the correlation matrix ($\mathbf{X}$

is standardized to have unit length of each column) or by some known positive-definite matrix.

2.2.1 Cos-max transformation

The transformed orthogonal matrix $\mathbf{Z}$ for the cos-max transformation can be obtained using either the singular value decomposition, spectral decomposition or the symmetric square-root of a positive-definite matrix.

The spectral decomposition or eigen decomposition of a $p \times p$ symmetric matrix $\mathbf{B}$ is $\mathbf{B} = \mathbf{L}\mathbf{D}\mathbf{L}^\top$, where $\mathbf{L}$ is a $p \times p$ orthogonal matrix with eigenvectors of $\mathbf{B}$ as the columns and $\mathbf{D}$ is a $p \times p$ diagonal matrix with diagonal elements equal to the eigenvalues of $\mathbf{B}$. And the square-root matrix of $\mathbf{B}$ is $\mathbf{B}^{1/2} = \mathbf{L}\mathbf{D}^{1/2}\mathbf{L}$, where $\mathbf{D}^{1/2}$ is a $p \times p$ diagonal matrix having square-roots of the eigenvalues of $\mathbf{B}$ as the diagonal elements.

Suppose $\mathbf{X} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^\top$ is the singular value decomposition of $\mathbf{X}$, where $\mathbf{U}$, $\mathbf{V}$ and $\boldsymbol{\Delta}$ are the same as in equation (2.2). Then $\mathbf{X}^\top\mathbf{X} = \mathbf{V}\boldsymbol{\Delta}^2\mathbf{V}^\top$ and $\{\mathbf{X}^\top\mathbf{X}\}^{1/2} = \mathbf{V}\boldsymbol{\Delta}\mathbf{V}^\top$ are respectively the spectral decomposition and square-root matrix of $\mathbf{X}^\top\mathbf{X}$. Since $\mathbf{X} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^\top$ this implies $\mathbf{U} = \mathbf{X}\mathbf{V}\boldsymbol{\Delta}^{-1}$ as $\mathbf{V}$ is an orthogonal matrix and hence $\left(\mathbf{V}^\top\right)^{-1} = \mathbf{V}$. Also $\{\mathbf{X}^\top\mathbf{X}\}^{1/2} = \mathbf{V}\boldsymbol{\Delta}\mathbf{V}^\top$ implies $\{\mathbf{X}^\top\mathbf{X}\}^{-1/2} = \mathbf{V}\boldsymbol{\Delta}^{-1}\mathbf{V}^\top$ (according to the definition of the inverse of a square-root matrix). The problem of maximization in equation (2.18) becomes

$$\psi = \sum_{j=1}^{p} \boldsymbol{x}_j^\top \boldsymbol{z}_j = tr\left(\mathbf{X}^\top\mathbf{Z}\right) = tr\left(\mathbf{V}\boldsymbol{\Delta}\mathbf{U}^\top\mathbf{Z}\right) = tr\left(\boldsymbol{\Delta}\mathbf{U}^\top\mathbf{Z}\mathbf{V}\right) \tag{2.20}$$

subject to the condition that $\mathbf{Z}$ is an orthogonal matrix.

As $\mathbf{Z}$ and $\mathbf{V}$ are orthogonal matrices their product $\mathbf{Z}\mathbf{V}$ is orthogonal and, for the same reason, $\mathbf{U}^\top\mathbf{Z}\mathbf{V}$ is also orthogonal. Now since $\boldsymbol{\Delta}$ is a diagonal matrix with

positive diagonal elements and $\mathbf{U}^\top\mathbf{ZV}$ is an orthogonal matrix, $tr\left(\mathbf{\Delta U}^\top\mathbf{ZV}\right)$ is maximized when $\mathbf{U}^\top\mathbf{ZV}$ is an identity matrix i.e., $\mathbf{U}^\top\mathbf{ZV} = \mathbf{I}_p$. Now since $\mathbf{U}$ is column orthogonal and $\mathbf{V}$ is orthogonal $\mathbf{U}^\top\mathbf{ZV} = \mathbf{I}_p$ implies

$$\mathbf{Z} = \mathbf{UV}^\top = \mathbf{XV\Delta}^{-1}\mathbf{V}^\top = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}. \tag{2.21}$$

Now $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}$ is unique as it is the symmetric square-root of $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$. Hence equation (2.21) determines $\mathbf{Z}$ uniquely.

Equation (2.9) and (2.21) implies that Johnson's (1966) method and the cos-max transformation of Garthwaite et al. (2012) produce the same orthogonal vectors. Both methods yield a unique orthogonal matrix $\mathbf{Z}$ whose columns are highly correlated with the corresponding columns of $\mathbf{X}$.

2.2.2 Cos-square transformation

The following theorem underpins the cos-square transformation and also gives the transformation matrix $\mathbf{A}$ for the cos-max transformation. Proof of the theorem is given in Garthwaite et al. (2012, p.789).

Theorem 1. *Suppose* $\mathbf{C} = diag(c_1, \ldots, c_p)$ *where the* c_j*'s are known positive constants. Then the unique maximum of* $tr\left(\mathbf{CX}^\top\mathbf{Z}\right) = \sum_{j=1}^{p} c_j\boldsymbol{x}_j^\top\boldsymbol{z}_j$*, under conditions 1 and 2, occurs when* $\mathbf{Z} = \mathbf{XC}\left(\mathbf{CX}^\top\mathbf{XC}\right)^{-1/2}$.

Putting $\mathbf{C} = \mathbf{I}_p$ in Theorem 1 gives the unique cos-max transformed matrix $\mathbf{Z} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}$ which maximizes $\psi = \sum_{j=1}^{p}\boldsymbol{x}_j^\top\boldsymbol{z}_j = tr\left(\mathbf{X}^\top\mathbf{Z}\right)$ under conditions 1 and 2.

With the cos-square transformation, the c_j's cannot be obtained from a simple formula. Rather, they are calculated from an iterative algorithm (Garthwaite

et al., 2012), reproduced below. Theorem 1 yields $\mathbf{Z}$ by maximizing $\sum_{j=1}^{p} c_j \boldsymbol{x}_j^\top \boldsymbol{z}_j$, which is the weighted version of the cos-max transformation. Each product $\boldsymbol{x}_j^\top \boldsymbol{z}_j$ is weighted by c_j. The algorithm repeatedly maximizes $\sum_{j=1}^{p} c_j \boldsymbol{x}_j^\top \boldsymbol{z}_j$ until convergence. After each iteration, c_j is set equal to the most recent estimate of $\boldsymbol{x}_j^\top \boldsymbol{z}_j$ $(j = 1, \ldots, p)$.

Algorithm 1. Algorithm for the cos-square transformation (Garthwaite et al., 2012).

Step 1. Set $\mathbf{C}_1$ equal to the $p \times p$ identity matrix and put $v = 1$.

Step 2. At the vth iteration, determine the square-root matrix of $\mathbf{C}_v \mathbf{X}^\top \mathbf{X} \mathbf{C}_v$, i.e., $\left(\mathbf{C}_v \mathbf{X}^\top \mathbf{X} \mathbf{C}_v\right)^{1/2}$.

Step 3. Set $\mathbf{C}_{v+1}$ equal to a diagonal matrix, with diagonal elements equal to the diagonal elements of $\mathbf{C}_v^{-1} \left(\mathbf{C}_v \mathbf{X}^\top \mathbf{X} \mathbf{C}_v\right)^{1/2}$.

Step 4. Repeat Steps 2 and 3 until convergence, i.e., until $\mathbf{C}_{v+1} \approx \mathbf{C}_v$.

Step 5. Set $\mathbf{Z}$ equal to $\mathbf{X}\mathbf{C}_v \left(\mathbf{C}_v \mathbf{X}^\top \mathbf{X} \mathbf{C}_v\right)^{-1/2}$.

Garthwaite et al. (2012) discussed the rationale behind the algorithm and proved that the algorithm converges to a unique maximum.

Step 3 of the algorithm sets the diagonal elements of $\mathbf{C}_{v+1}$ equal to the diagonal elements of $\mathbf{C}_v^{-1} \left(\mathbf{C}_v \mathbf{X}^\top \mathbf{X} \mathbf{C}_v\right)^{1/2}$. Let $\mathbf{C}$ denote the value of $\mathbf{C}_v$ when the algorithm converges. Step 4 implies that at convergence the diagonal elements of $\mathbf{C}$ and $\mathbf{C}^{-1} \left(\mathbf{C}\mathbf{X}^\top \mathbf{X}\mathbf{C}\right)^{1/2}$ are equal. Olkin and Pratt (1958) proved that a positive definite matrix $\mathbf{B}$ can be uniquely decomposed as $\mathbf{B} = \mathbf{RDR}$, where $\mathbf{R}$ is a correlation matrix and $\mathbf{D}$ is a diagonal matrix. Now $\mathbf{C}^{-1} \left(\mathbf{C}\mathbf{X}^\top \mathbf{X}\mathbf{C}\right)^{1/2} \mathbf{C}^{-1}$ is positive

definite and has diagonal elements that each equal 1, so it is a correlation matrix. Also $\mathbf{X}^\top\mathbf{X}$ is a positive definite matrix. If we put $\mathbf{R} = \mathbf{C}^{-1}(\mathbf{C}\mathbf{X}^\top\mathbf{X}\mathbf{C})^{1/2}\mathbf{C}^{-1}$ and $\mathbf{X}^\top\mathbf{X} = \mathbf{RDR}$ then $\mathbf{D} = \mathbf{C}^2$ as $\mathbf{RDR} = \mathbf{C}^{-1}(\mathbf{C}\mathbf{X}^\top\mathbf{X}\mathbf{C})^{1/2}\mathbf{C}^{-1}\mathbf{D}\mathbf{C}^{-1}(\mathbf{C}\mathbf{X}^\top\mathbf{X}\mathbf{C})^{1/2}\mathbf{C}^{-1}$ $= \mathbf{C}^{-1}(\mathbf{C}\mathbf{X}^\top\mathbf{X}\mathbf{C})^{1/2}\mathbf{C}^{-1}\mathbf{C}^2\mathbf{C}^{-1}(\mathbf{C}\mathbf{X}^\top\mathbf{X}\mathbf{C})^{1/2}\mathbf{C}^{-1} = \mathbf{C}^{-1}(\mathbf{C}\mathbf{X}^\top\mathbf{X}\mathbf{C})\mathbf{C}^{-1} = \mathbf{X}^\top\mathbf{X}$. Since $\mathbf{D}$ is unique (Olkin and Pratt, 1958) and the elements of $\mathbf{C}$ are positive, so $\mathbf{C}$ is also unique. Also the symmetric square-root matrix $\left(\mathbf{C}\mathbf{X}^\top\mathbf{X}\mathbf{C}\right)^{1/2}$ is unique. Hence at convergence the transformation matrix $\mathbf{A} = \mathbf{C}\left(\mathbf{C}\mathbf{X}^\top\mathbf{X}\mathbf{C}\right)^{-1/2}$ is unique. Also the algorithm converges to a global maximum as there is only one local maximum (see Lemma 1 & Theorem 2 of Garthwaite et al. (2012)).

2.2.3 Relationship between the cos-max and cos-square transformations for two variables

Consider two $n \times 1$ normalized vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, i.e., $\boldsymbol{x}_1^\top\boldsymbol{x}_1 = \boldsymbol{x}_2^\top\boldsymbol{x}_2 = 1$. We want to determine two orthonormal vectors $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ such that $\boldsymbol{z}_1$ is close to $\boldsymbol{x}_1$ and $\boldsymbol{z}_2$ is close to $\boldsymbol{x}_1$. These four vectors are displayed in Figure 2.1 in two-dimensional space. This two-dimensional space is a plane from n-dimensional space. Suppose $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are each of length 1. Denote the angle between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ by α. Define β by letting the angle between $\boldsymbol{x}_1$ and $\boldsymbol{z}_1$ be $\frac{\pi}{4} - \frac{\alpha}{2} + \beta = \gamma + \beta$. As $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are orthogonal, the angle between $\boldsymbol{x}_2$ and $\boldsymbol{z}_2$ is $\frac{\pi}{4} - \frac{\alpha}{2} - \beta = \gamma - \beta$. Since $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are normalized vectors, $\boldsymbol{x}_1^\top\boldsymbol{z}_1 = cos(\gamma + \beta)$ and $\boldsymbol{x}_2^\top\boldsymbol{z}_2 = cos(\gamma - \beta)$. Also, $(\boldsymbol{x}_1^\top\boldsymbol{z}_1)^2 = cos^2(\gamma + \beta)$ and $(\boldsymbol{x}_2^\top\boldsymbol{z}_2)^2 = cos^2(\gamma - \beta)$. Hence the maximization problems in Equations (2.18) and (2.19) can be written as:

Choose β_1 to maximize

$$\psi = cos(\gamma + \beta_1) + cos(\gamma - \beta_1) = 2cos\gamma\, cos\beta_1 \tag{2.22}$$

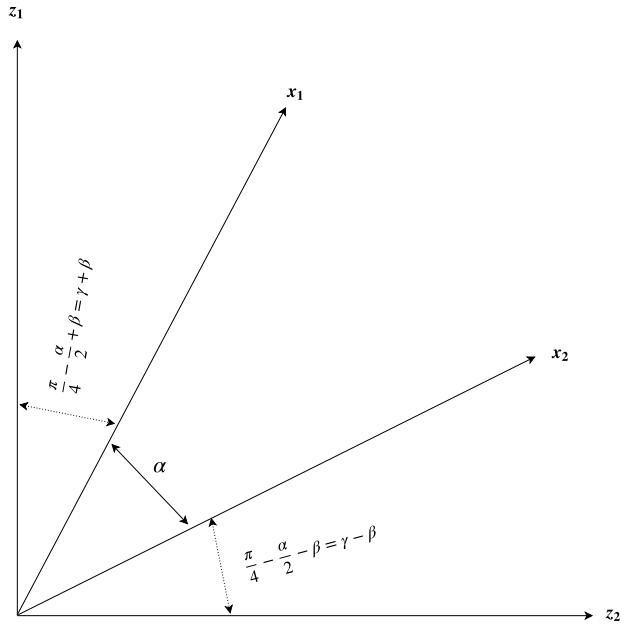

Figure 2.1: Geometric presentation of original vectors and transformed orthonormal vectors

for the cos-max transformation and choose β_2 to maximize

$$\phi = cos^2(\gamma + \beta_2) + cos^2(\gamma - \beta_2). \tag{2.23}$$

for the cos-square transformation.

Equation (2.22) is maximized when $\beta_1 = 0°$. That is the angle between $\boldsymbol{x}_1$ and $\boldsymbol{z}_1$ and between $\boldsymbol{x}_2$ and $\boldsymbol{z}_2$ is $\gamma = \frac{\pi}{4} - \frac{\alpha}{2}$. Now equation (2.23) can be written as

$$\begin{aligned} \phi &= 2(cos^2\gamma\, cos^2\beta_2 + sin^2\gamma\, sin^2\beta_2) \\ &= 2(cos^2\beta_2\, cos(2\gamma) + sin^2\gamma) \end{aligned} \tag{2.24}$$

which is maximized when $\beta_2 = 0°$ as $\gamma < 45°$. It is noted that ϕ has another candidate point at $\beta_2 = \frac{\pi}{2}$. However, from Figure 2.1, $\beta_2 = \frac{\pi}{2}$ is not feasible.

When $\alpha \neq 0°$ the cos-max and cos-square solutions are identical as $\beta_1 = \beta_2$. However, when $\alpha = 0°$, then $\mathbf{X}^\top\mathbf{X}$ is singular and hence both solutions are undetermined.

The identical solutions for the cos-max and cos-square transformations holds

for the normalized vectors. However, for the unscaled vectors Equations (2.22) and (2.23) will have a length term and the cos-max and cos-square transformation will not be identical.

2.2.4 Related optimization problem

Bolla et al. (1998) considered choosing orthonormal vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ such that

$$\zeta = \sum_{j=1}^{p} \boldsymbol{z}_j^\top \mathbf{A}_j \boldsymbol{z}_j \tag{2.25}$$

is maximized, where $\mathbf{A}_1, \ldots, \mathbf{A}_p$ are $n \times n$ symmetric positive definite matrices ($p \leq n$). Bolla (2001) relaxed the requirement that the $\mathbf{A}_j$ are positive definite matrices and allowed them to be positive semi-definite. The optimization problem given by (2.19), which leads to the cos-square transformation, is a special case of (2.25) that is obtained by setting $\mathbf{A}_j$ equal to $\boldsymbol{x}_j \boldsymbol{x}_j^\top$ for $j = 1, \ldots, p$.

Bolla et al. (1998) form an $np \times np$ block-diagonal matrix $\mathbf{A}$ with $\mathbf{A}_1, \ldots, \mathbf{A}_p$ in the diagonal blocks and $\mathbf{A}$ is zero elsewhere. Then the orthonormal vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ which maximize ζ must satisfy the matrix equation

$$\mathbf{A}(\mathbf{Z}) = \mathbf{Z}\mathbf{S} \tag{2.26}$$

where the $n \times p$ matrices $\mathbf{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p)$ and $\mathbf{A}(\mathbf{Z}) = (\mathbf{A}_1 \boldsymbol{z}_1, \ldots, \mathbf{A}_p \boldsymbol{z}_p)$ are formed by considering enumerated vectors as their columns and $\mathbf{S}$ is a $p \times p$ symmetric matrix.

The matrix equation (2.26) is linear in $\mathbf{Z}$ and has a non-trivial (non-zero) solution for the appropriate $\mathbf{S}$ if and only if

$$|\mathbf{A} - \mathbf{I}_n \otimes \mathbf{S}| = 0 \tag{2.27}$$

where $\otimes$ denotes the Kronecker-product of matrices. The Kronecker product of an $m \times n$ matrix $\mathbf{F}$ and a $p \times q$ matrix $\mathbf{L}$ is a $mp \times nq$ matrix defined as

$$\mathbf{F} \otimes \mathbf{L} = \begin{pmatrix} f_{11}\mathbf{L} & \dots & f_{1n}\mathbf{L} \\ \vdots & \ddots & \vdots \\ f_{m1}\mathbf{L} & \dots & f_{mn}\mathbf{L} \end{pmatrix}. \tag{2.28}$$

The task is to choose $\mathbf{S}$ so that the columns of $\mathbf{Z}$ are orthonormal. Bolla et al. (1998) proposed an iterative algorithm to find $\mathbf{S}$ and the corresponding orthonormal vectors $\boldsymbol{z}_1, \dots, \boldsymbol{z}_p$. The algorithm uses polar decomposition. The polar decomposition of an $m \times n$ matrix $\mathbf{M}$ is a factorization of the form $\mathbf{M} = \mathbf{RH}$, where $\mathbf{R}$ is an $m \times n$ column orthogonal matrix ($\mathbf{R}^\top \mathbf{R} = \mathbf{I}_p$) and $\mathbf{H}$ is an $n \times n$ positive semi-definite matrix. $\mathbf{R}$ and $\mathbf{H}$ can be obtained using the singular value decomposition of $\mathbf{M}$. If the singular value decomposition of $\mathbf{M}$ is $\mathbf{M} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^\top$ then $\mathbf{R} = \mathbf{U}\mathbf{V}^\top$ and $\mathbf{H} = \mathbf{V}\boldsymbol{\Delta}\mathbf{V}^\top$.

Algorithm 2. Algorithm for optimizing sums of heterogeneous quadratic forms.

Step 1. Choose an initial (arbitrary) orthonormal set of vectors $\boldsymbol{z}_1^{(0)}, \dots, \boldsymbol{z}_p^{(0)}$ and form $\left(\mathbf{A}_1\boldsymbol{z}_1^{(0)}, \dots, \mathbf{A}_p\boldsymbol{z}_p^{(0)}\right)$.

Step 2. Perform a polar decomposition of $\left(\mathbf{A}_1\boldsymbol{z}_1^{(0)}, \dots, \mathbf{A}_p\boldsymbol{z}_p^{(0)}\right)$ that yields a set of orthonormal vectors $\boldsymbol{z}_1^{(1)}, \dots, \boldsymbol{z}_p^{(1)}$ and symmetric positive definite matrix $\mathbf{S}^{(1)}$ such that $\left(\mathbf{A}_1\boldsymbol{z}_1^{(0)}, \dots, \mathbf{A}_p\boldsymbol{z}_p^{(0)}\right) = \mathbf{Z}^{(1)}\mathbf{S}^{(1)}$, where $\mathbf{X}^{(1)} = \left(\boldsymbol{z}_1^{(1)}, , \dots, \boldsymbol{z}_1^{(p)}\right)$.

Step 3. Repeat Step 2 until convergence, when

$$\sum_{j=1}^{p} \left(\boldsymbol{z}_j^{(m)}\right)^\top \mathbf{A}_j \boldsymbol{z}_j^{(m)} \approx \sum_{j=1}^{p} \left(\boldsymbol{z}_j^{(m-1)}\right)^\top \mathbf{A}_j \boldsymbol{z}_j^{(m-1)}.$$

The unique polar decomposition determines $\mathbf{Z}^{(m)}$. Bolla et al. (1998) showed that their algorithm converges to a local maximum of the objective function, and they conjectured that it converges to a global maximum. As noted earlier, the cos-square transformation of Garthwaite et al. (2012) is a special case of Bolla et al. (1998) optimization problem when each $\mathbf{A}_j = \boldsymbol{x}_j \boldsymbol{x}_j^\top$. Garthwaite et al. (2012) proved the conjecture of Bolla et al. (1998) for this important special case.

2.3 Corr-max transformation

The cos-max and the cos-square transformations were designed to transform a data matrix and yield a matrix with orthonormal columns, where each transformed variable is closely associated with only one of the original variables. The corr-max transformation of Garthwaite and Koch (2016) transforms a random vector (instead of a data matrix) and has been used to partition the contribution of individual variables to a quadratic form.

Suppose $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ is a $p \times 1$ random vector with $\text{var}(\boldsymbol{X}) \propto \boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ is an arbitrary $p \times 1$ vector. Then the quadratic form Q is defined as

$$Q = (\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \tag{2.29}$$

where $\boldsymbol{\mu}$ is not necessarily the mean of $\boldsymbol{X}$. Such types of quadratic form arise in various situations, such as in the probability density function of a multivariate normal distribution, Hotelling's T^2-statistic, Mahalanobis distance and discriminant analysis.

If $\boldsymbol{\Sigma}$ is an identity matrix then the contribution of each variable is clearly the square of the corresponding component of $\boldsymbol{X} - \boldsymbol{\mu}$. This partitioning can be

extended for a diagonal $\boldsymbol{\Sigma}$. However, variables are usually correlated and then $\boldsymbol{\Sigma}$ is not diagonal. Garthwaite and Koch (2016) proposed a method for evaluating the contribution of individual variables to a quadratic form for correlated variables. They consider a transformation of $\boldsymbol{X}$ to $\boldsymbol{W}$ of the form

$$\boldsymbol{W} = \mathbf{A}\left(\boldsymbol{X} - \boldsymbol{\mu}\right), \tag{2.30}$$

where $\mathbf{A}$ is a $p \times p$ transformation matrix such that $\boldsymbol{W}$ is a $p \times 1$ vector and

$$\boldsymbol{W}^\top \boldsymbol{W} = \left(\boldsymbol{X} - \boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{X} - \boldsymbol{\mu}\right). \tag{2.31}$$

Equation (2.31) implies that

$$Q = \boldsymbol{W}^\top \boldsymbol{W} = \sum_{j=1}^{p} W_j^2, \tag{2.32}$$

where $\boldsymbol{W} = (W_1, \ldots, W_p)^\top$. And since the quadratic form can be expressed as the sum of the squared elements of $\boldsymbol{W}$, so $\boldsymbol{W}$ forms a partition of Q. They argue that the partition will be useful and meaningful if

(a) the components of $\boldsymbol{W}$ are uncorrelated with variance proportional to the identity matrix and

(b) X_j is associated with only W_j, so that the contribution of X_j to Q can sensibly be defined as W_j^2.

Since $\boldsymbol{W} = \mathbf{A}\left(\boldsymbol{X} - \boldsymbol{\mu}\right)$ and $\boldsymbol{W}^\top \boldsymbol{W} = \left(\boldsymbol{X} - \boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{X} - \boldsymbol{\mu}\right)$, it follows that $\mathbf{A}^\top \mathbf{A} = \boldsymbol{\Sigma}^{-1}$. And hence $\text{var}\left(\boldsymbol{W}\right) \propto \mathbf{A}\text{var}\left(\boldsymbol{X} - \boldsymbol{\mu}\right)\mathbf{A}^\top = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top = \mathbf{A}\left(\mathbf{A}^\top \mathbf{A}\right)^{-1}\mathbf{A}$ $= \mathbf{I}_p$, which indicates that the components of $\boldsymbol{W}$ are uncorrelated. The degree to which W_j approximates X_j can be determined by the correlation between X_j and W_j. The task is to find $\mathbf{A}$ such that $\sum_{j=1}^{p} \text{corr}\left(X_j, W_j\right)$ is maximized subject

to the condition $\mathbf{A}^\top\mathbf{A} = \mathbf{\Sigma}^{-1}$. The following theorem from Garthwaite and Koch (2016) solves this to give the corr-max transformation.

Theorem 2. *Suppose var* $(\boldsymbol{X}) \propto \mathbf{\Sigma}$ *and* $\mathbf{D\Sigma D}$ *has diagonal elements of 1 with* $\mathbf{D}$ *a diagonal matrix. The maximum of* $\sum_{j=1}^{p} corr\,(X_j, W_j)$ *subject to* $\boldsymbol{W} = \mathbf{A}\,(\boldsymbol{X} - \boldsymbol{\mu})$*, where the square matrix* $\mathbf{A}$ *is such that* $\mathbf{A}^\top\mathbf{A} = \mathbf{\Sigma}^{-1}$ *and* $\boldsymbol{\mu}$ *is any given vector, occurs when* $\mathbf{A} = (\mathbf{D\Sigma D})^{-1/2}\,\mathbf{D}$ *and then* $\boldsymbol{W} = (\mathbf{D\Sigma D})^{-1/2}\,\mathbf{D}\,(\boldsymbol{X} - \boldsymbol{\mu})$.

Proof of Theorem 2 is given in the appendix of Garthwaite and Koch (2016). The $p \times p$ diagonal matrix $\mathbf{D}$ has diagonal elements that are the reciprocals of the square root of the corresponding diagonal elements of $\mathbf{\Sigma}$. $\mathbf{D\Sigma D}$ is the correlation matrix of $\boldsymbol{X}$ and (j,k)th element of $(\mathbf{D\Sigma D})^{1/2}$ is the correlation between X_j and W_k. Hence the diagonal elements of $(\mathbf{D\Sigma D})^{1/2}$ measure the degree of relationship between the components of $\boldsymbol{X}$ and the corresponding components of $\boldsymbol{W}$.

2.4 Duplicate invariance and rotation invariance properties

2.4.1 Duplicate invariance property

Garthwaite et al. (2012) showed that the cos-square transformation has a duplicate invariance property. Suppose we have two sets of vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{x}_{(p+1)}, \ldots, \boldsymbol{x}_{(p+k)}$ where the first set is a subset of the second set. If both sets are transformed separately then the transformed values of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ in the two sets are likely to differ. This is also true for the case where the original sets are transformed to orthogonal sets, the only exception is when the vectors

$\boldsymbol{x}_{(p+1)}, \ldots, \boldsymbol{x}_{(p+k)}$ are orthogonal to $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$. However, suppose each of the vectors $\boldsymbol{x}_{(p+1)}, \ldots, \boldsymbol{x}_{(p+k)}$ is an exact duplicate of $\boldsymbol{x}_p$. Then the transformed orthogonal vectors associated with $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{(p-1)}$ obtained under the cos-square transformation will be the same in the two transformed sets. That is, duplication of $\boldsymbol{x}_p$ has no effect on the transformed values of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{(p-1)}$. Also the transformed values of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{(p-1)}$ are virtually the same if each $\boldsymbol{x}_{(p+1)}, \ldots, \boldsymbol{x}_{(p+k)}$ is virtually a duplicate of $\boldsymbol{x}_p$, i.e., if $x_l = x_p + \alpha\zeta_l$ for $l = p+1, \ldots, p+k$, where $\alpha \approx 0$ and ζ_l can be any value. Garthwaite et al. (2012) called this property of the cos-square transformation the 'duplicate invariance property'. For more details of the duplicate invariance property, see Garthwaite et al. (2012).

2.4.2 Rotation invariance property

When there is a strong collinearity between some of the X variables, then some columns of the transformed orthogonal data matrix are not close representation of the corresponding columns of $\mathbf{X}$. Orthogonal rotation of collinear variables can remove collinearities but, with most transformations, rotation of some variables will typically affect the transformed values of all variables. However, as noted in the introduction, the cos-max and the corr-max transformations have a rotation invariance property. In this subsection we first show that rotation can reduce collinearity and then describe the rotation invariance property.

An orthogonal rotation of axes X_1, X_2 to axes X_1^*, X_2^* is illustrated in Figures 2.1(a) and 2.1(b). In Figure 2.1(a), the positions of 10 points (x_1, x_2) are plotted and new axes X_1^* and X_2^* are shown. The new axes are obtained by rotating the original axes X_1 and X_2 (by 45^o in this case). Figure 2.1(b) shows the same 10

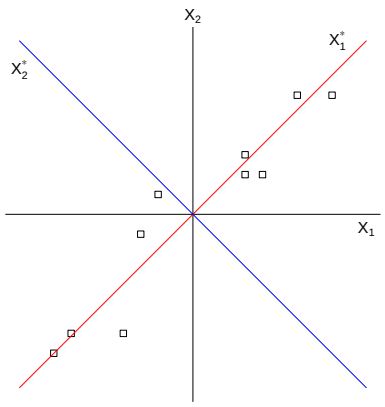

Figure 2.1(a) Points before rotation with X_1, X_2 as axes.

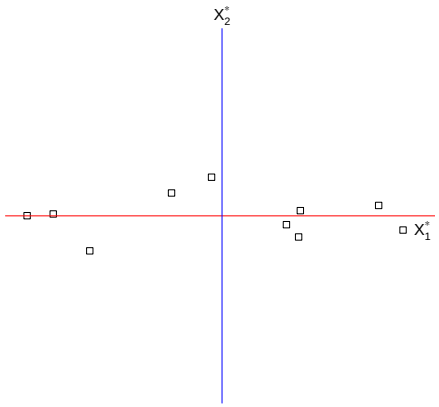

Figure 2.1(b) Points after rotation with X_1^*, X_2^* as axes.

points, but drawn with X_1^* and X_2^* as the horizontal and vertical axes. It can be seen that rotation of axes changes the correlation between variables: Figure 2.1(a) shows that the points are highly correlated when expressed in terms of X_1 and X_2, while Figure 2.1(b) shows that the correlation is low when the points are expressed in terms of X_1^* and X_2^*. Consequently, orthogonal rotation can be used to remove or reduce collinearity between variables.

We only need to rotate those variables that are involved in collinearities. For example, suppose there is just one collinearity and it involves only the first d variables $X_1, \ldots, X_d$. Then axes are rotated using a rotation matrix, $\mathbf{\Gamma}$ say, that has the following block-diagonal form:

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{\Gamma}_d & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-d} \end{pmatrix}, \tag{2.33}$$

where Γ_d is an orthogonal matrix of order d and $\mathbf{I}_{p-d}$ is a $(p-d)$ order identity matrix.

Rotation produces new variables that are linear combinations of the original variables. The rotation matrix should be chosen in such a way that the variables

that are created have meaningful interpretation. For example, if only the first two variables X_1 and X_2 are responsible for one collinearity then $\mathbf{\Gamma}_d$ can be set as:

$$\mathbf{\Gamma}_d = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \tag{2.34}$$

This rotation creates two meaningful variables, the first one is proportional to $X_1 + X_2$ and the second one is proportional to $X_2 - X_1$.

In terms of the original variables, X_1 and X_2, the ten points in Figure 2.1 form the data matrix:

$$\begin{pmatrix} -0.48 & -0.42 & -0.24 & -0.18 & -0.12 & 0.18 & 0.18 & 0.24 & 0.36 & 0.48 \\ -0.48 & -0.41 & -0.41 & -0.07 & 0.07 & 0.14 & 0.21 & 0.14 & 0.41 & 0.41 \end{pmatrix}$$

Post-multiplying this data matrix by $\mathbf{\Gamma}_d$ in equation (2.34) gives the points in terms of the new variables X_1^* and X_2^*:

$$\begin{pmatrix} -0.68 & -0.59 & -0.46 & -0.18 & -0.04 & 0.22 & 0.27 & 0.27 & 0.55 & 0.63 \\ 0.00 & 0.01 & -0.12 & 0.08 & 0.13 & -0.03 & 0.02 & -0.07 & 0.04 & -0.05 \end{pmatrix}$$

The sample correlation between X_1 and X_2 is 0.951, while the sample correlation between X_1^* and X_2^* is 0. (The correlation between the sum and difference of two variables that have been standardised to have equal variances is always 0.)

To illustrate the rotation invariance property of the cos-max and the corr-max transformations, suppose the rotation matrix is of the form in equation (2.33). Let the data matrix after rotation be $\mathbf{X}^*$, i.e., $\mathbf{X}^* = \mathbf{X}\mathbf{\Gamma}$. This rotation rotates only the first d columns of $\mathbf{X}$, while the last $p-d$ columns of $\mathbf{X}$ and $\mathbf{X}^*$ are the same. If the cos-max transformation is applied separately to $\mathbf{X}$ and $\mathbf{X}^*$ then the last $p-d$ columns of $\mathbf{Z}$ and $\mathbf{Z}^*$ are same, where $\mathbf{Z}$ and $\mathbf{Z}^*$ are, respectively, the transformed matrices of $\mathbf{X}$ and $\mathbf{X}^*$. That is, with the cos-max transformation, the rotation

of variables has no effect on the transformed values of the unrotated variables. This is also true for the corr-max transformations. Rotation of first d columns changes only the first d components of $\boldsymbol{W}$, the remaining $p-d$ components remain unchanged. Consequently the contribution of last $p-d$ variables are unchanged by the rotation. Rotation can be performed before or after transformation and in both cases yields the same result. For a detailed description of the rotation invariance property of the corr-max transformation see Garthwaite and Koch (2016).

2.5 Applications of the transformations

Gibson (1962), Johnson (1966) and Johnson (2000) consider the task of determining the relative importance of regressor in a multiple regression. Applications of the cos-max and cos-square transformations can also yield new statistical methods (Garthwaite et al., 2012), such as (i) a unified approach to the identification and diagnosis of collinearities, (ii) a method of setting prior weights for Bayesian model averaging, and (iii) calculating an upper bound for a multivariate Chebyshev inequality. The corr-max transformation has applications in determining the contribution of individual variables to a quadratic form (Garthwaite and Koch, 2016). These applications are all briefly described in this section.

2.5.1 Relative importance of variables in multiple regression

For this subsection, we assume that the columns of $\mathbf{X}$ have been standardized to have means of 0 and unit lengths. Both Gibson (1962) and Johnson (1966) used

the square of the jth element of $\widehat{\boldsymbol{\beta}}_Z$ obtained in equation (2.10), $\widehat{\beta}^2_{Z_j}$, as the relative importance measure of X_j. J. W. Johnson (2000) argued with an example that when the original X variables are highly collinear, the correlation between X_j and Z_j can be small for some j. Since $\mathbf{Z}$ is a linear transformation of $\mathbf{X}$, X_j might not be only highly correlated with Z_j for correlated X variables. Hence, some X variables that are highly correlated with the response variable can be assigned very small relative importance. J. W. Johnson (2000) considered regressing $\mathbf{X}$ on $\mathbf{Z}$. Denote the matrix of regression coefficients by $\boldsymbol{\Lambda}$, where

$$\boldsymbol{\Lambda} = \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \mathbf{X} = \mathbf{Z}^\top \mathbf{X}. \tag{2.35}$$

The matrix $\boldsymbol{\Lambda} = \mathbf{Z}^\top \mathbf{X}$ is the correlation matrix between the X and Z variables. If λ_{jk} is the correlation between X_j and Z_k, then λ^2_{jk} is the proportion of variance of Z_k accounted by X_j and vice-versa. The proportion of variance of Y accounted by X_j derives from Z_k is $\lambda^2_{jk}\widehat{\beta}^2_{Z_k}$. Finally, the vector of relative importance weights $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)^\top$ is given by

$$\boldsymbol{\epsilon} = \boldsymbol{\Lambda}^{[2]} \widehat{\boldsymbol{\beta}}^{[2]}. \tag{2.36}$$

Again, the sum of the relative importance weights is equal to model R^2.

For further details see Gibson (1962), Johnson (1966) and J. W. Johnson (2000), also see Section 4.4 in Chapter 4.

2.5.2 Detection and identification of collinearities

Variance inflation factors are commonly used to determine whether a particular variable is responsible for collinearity among a set of variables. If collinearities are detected, then eigenvectors corresponding to small eigenvalues are generally examined in order to identify which variables cause them (form a collinear set).

This gives no direct relationship between the quantity used to identify a collinearity and the quantity used to determine its cause, as an eigenvector is not explicitly linked to any particular variance inflation factor. The transformation matrix $\mathbf{A}$ of both the cos-max and the cos-square transformations may simultaneously be used as a diagnostic for determining the number of collinearities and as a means of identifying the variables that contribute to each collinearity (Garthwaite et al., 2012). With either transformation, VIF_j is expressed as a sum of the squared elements of the jth row of the transformation matrix $\mathbf{A}$. That is, there is a one-to-one relationship between VIF_j and the jth row of $\mathbf{A}$. Detailed description of collinearity diagnostics using the cos-max or the cos-square transformation matrix is given in Garthwaite et al. (2012) and it is the focus of Chapter 5 of this thesis, where further detail is given.

2.5.3 Prior weights for Bayesian model averaging

In Bayesian model averaging, a prior weight must be given to each of the models under consideration. The simplest and most common choice is a discrete uniform prior in which each model is given the same prior weight. However, selection of a uniform prior ignores the similarities among the models (George et al., 2010). It can be argued that models that are very similar to each other should be given smaller weights (George, 1999), which is referred as the dilution of prior weights.

The most common situation where model averaging is used is in regression problems that involve variable selection. In that context, each model would be the regression model with one particular set of regressors. Hence model 1 might contain variables X_1, X_3 and X_5; model 2 might contain variables X_1, X_6, X_7

and X_9; and so on. Then, several models might contain similar sets of variables, while other models contain a set of variables that has few similarities to the sets of variables found in other models. If each of the models is given the same prior weight, George et al. (2010) argue that the similar models receive too much weight as they approximately duplicate each other, while the unusual models receive too little weight. The procedure for assigning prior weights should give reduced weights to similar models.

Suppose $p_1, \ldots, p_p$ are the prior probabilities or prior weight assigned to the models $M_1, \ldots, M_p$. One way of choosing the prior probabilities is by using the cos-square transformation (Garthwaite and Mubwandarikwa, 2010). They determine a $p \times p$ matrix $\mathbf{R}$, where $\mathbf{R}$ is the correlation matrix with r_{jk} as the correlation between the predictions of the models M_j and M_k. The cos-square transformation is applied with $\mathbf{X}^\top\mathbf{X}$ set equal to the correlation matrix $\mathbf{R}$ and Algorithm 1 yields the diagonal matrix $\mathbf{C}$. If $\mathbf{C}$ has diagonal elements $c_1, \ldots, c_p$, then the model M_j is assigned the prior probability

$$p_j = \frac{c_j^2}{\sum_{i=1}^{p} c_i^2}. \tag{2.37}$$

This weighting scheme assigns smaller weights to more similar models, i.e., the models that are more highly correlated with other models (Garthwaite et al., 2012). Detailed description of the cos-square weighting scheme and dilution of prior probabilities are given in Garthwaite et al. (2012) and Garthwaite and Mubwandarikwa (2010).

2.5.4 Multivariate Chebychev inequality

Suppose the random vector $(X_1, \ldots, X_p)$ has mean vector $\mathbf{0}$ and non-singular covariance matrix $\mathbf{\Sigma} = (\sigma_{jk})$. Form the positive definite matrix $\mathbf{\Omega}$ with (j,k)th element equal to $\sigma_{jk}/(\sigma_j \sigma_k l_j l_k)$, where $l_1, \ldots, l_p$ are positive constants and $\sigma_{jj} = \sigma_j^2$ for $j = 1, \ldots, p$. Then the multivariate Chebychev inequality suggested by Olkin and Pratt (1958) is

$$pr\left(|X_j| \geq l_j \sigma_j, \text{ for some } j\right) \leq tr\left(\mathbf{T}^{-1}\mathbf{\Omega}\mathbf{T}^{-1}\right) \tag{2.38}$$

where $\mathbf{T}$ is the unique positive definite correlation matrix such that $\mathbf{T}\mathbf{\Omega}^{-1}\mathbf{T}$ is a diagonal matrix. For $p = 1$, inequality (2.38) reduces to the univariate Chebychev inequality, $pr\left(|X_1| \geq l_1\sigma_1\right) \leq 1/l_1^2$.

According to Olkin and Pratt (1958, p.233), $\mathbf{T}$ cannot be obtained from $\mathbf{\Omega}$ by standard matrix operations except in special cases. Garthwaite et al. (2012) suggested using the cos-square transformation to determine $\mathbf{T}$. Suppose $\mathbf{X}^\top\mathbf{X}$ is set equal to $\mathbf{\Omega}$, and the cos-square transformation is applied to $\mathbf{X}^\top\mathbf{X}$, giving the diagonal matrix $\mathbf{C}$ (c.f. Algorithm 1). Garthwaite et al. (2012) show that the upper bound of the inequality (2.38) is $tr(\mathbf{C}^2)$. For a more detailed description of the Multivariate Chebychev inequality and determining the upper bound of the inequality, see Olkin and Pratt (1958) and Garthwaite et al. (2012).

2.5.5 Partition of Hotelling's T^2, Mahalanobis distance and discriminant function

The corr-max transformation can be used to partition the contribution of individual variables to a quadratic form such as Hotelling's one-sample T^2, Hotelling's

two-sample T^2, Mahalanobis distance and a discriminant function. Suppose the statistic of interest is

$$\Theta = \delta \left(\boldsymbol{X} - \boldsymbol{\mu} \right)^{\top} \widehat{\boldsymbol{\Sigma}}^{-1} \left(\boldsymbol{X} - \boldsymbol{\mu} \right), \tag{2.39}$$

where δ is a positive scalar, $\widehat{\boldsymbol{\Sigma}}$ is an estimate of $\boldsymbol{\Sigma}$ with $\text{var}\left(\boldsymbol{X}\right) \propto \boldsymbol{\Sigma}$.

When $\boldsymbol{\Sigma}$ in Theorem 2 is unknown, we replace it by a sample estimate $\widehat{\boldsymbol{\Sigma}}$. If the sample variance of $\boldsymbol{X}$ is proportional to $\widehat{\boldsymbol{\Sigma}}$, then the sample estimate of $\boldsymbol{W}$, denoted by $\widehat{\boldsymbol{W}} = \left(\widehat{W}_1, \ldots, \widehat{W}_p \right)^{\top}$, is obtained by (Garthwaite and Koch, 2016)

$$\widehat{\boldsymbol{W}} = \left(\widehat{\mathbf{D}} \widehat{\boldsymbol{\Sigma}} \widehat{\mathbf{D}} \right)^{-1/2} \widehat{\mathbf{D}} \left(\boldsymbol{X} - \boldsymbol{\mu} \right), \tag{2.40}$$

where $\widehat{\mathbf{D}}$ is a diagonal matrix obtained from $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\mathbf{D}} \widehat{\boldsymbol{\Sigma}} \widehat{\mathbf{D}}$ has diagonal elements of 1. Then the contribution of the jth X variable to Θ is evaluated as $\delta \widehat{W}_j^2$. Partitioning of Hotelling's one and two-sample T^2 statistics and Mahalanobis distance is straightforward as they have precisely the same form as in equation (2.39), while the discriminant function is closely related.

(a) *Hotelling's one-sample T^2 statistic.* Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be a random sample of size n from the multivariate normal population $N_p \left(\boldsymbol{\mu}, \boldsymbol{\Sigma} \right)$. Suppose the sample mean vector is $\bar{\boldsymbol{X}}$ and $\widehat{\boldsymbol{\Sigma}}_1$ is the sample covariance matrix. Then Hotelling's one-sample T^2 statistic for testing $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ is

$$T_1^2 = n \left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0 \right)^{\top} \widehat{\boldsymbol{\Sigma}}_1^{-1} \left(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0 \right). \tag{2.41}$$

Let $\boldsymbol{X} = \bar{\boldsymbol{X}}$, which justifies $\text{var}\left(\boldsymbol{X}\right) \propto \boldsymbol{\Sigma}$ as $\text{var}\left(\boldsymbol{X}\right) = \boldsymbol{\Sigma}/n$. Hence, putting $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}_1$, $\delta = n$ and $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ gives the contribution of individual variables to T_1^2.

(b) *Hotelling's two-sample T^2 statistic.* Suppose $\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{1n_1}$ and $\boldsymbol{X}_{21}, \ldots, \boldsymbol{X}_{2n_2}$ are two random samples of sizes n_1 and n_2 from two multivariate normal populations $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, respectively, having a common covariance matrix $\boldsymbol{\Sigma}$. Let $\bar{\boldsymbol{X}}_1$ and $\bar{\boldsymbol{X}}_2$ be the sample mean vectors and $\mathbf{S}_1$ and $\mathbf{S}_2$ be the sample covariance matrices. Then Hotelling's two-sample T^2 statistic for testing $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ is

$$T_2^2 = \{n_1 n_2/(n_1+n_2)\} \left(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2\right)^\top \widehat{\boldsymbol{\Sigma}}_p^{-1} \left(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2\right), \tag{2.42}$$

where $\widehat{\boldsymbol{\Sigma}}_p = \{(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2\}/(n_1+n_2-2)$ is the pooled estimate of $\boldsymbol{\Sigma}$. Let $\boldsymbol{X} = \bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2$, so $\text{var}(\boldsymbol{X}) = \boldsymbol{\Sigma}/n_1 + \boldsymbol{\Sigma}/n_2 = (1/n_1 + 1/n_2)\boldsymbol{\Sigma} \propto \boldsymbol{\Sigma}$. The contribution of individual variables to T_2^2 is obtained by putting $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}_p$, $\delta = n_1 n_2/(n_1+n_2)$ and $\boldsymbol{\mu} = \mathbf{0}$.

(c) *Mahalanobis distance.* The Mahalanobis distance between two random vectors $\boldsymbol{X}_{[1]}$ and $\boldsymbol{X}_{[2]}$ is

$$\left(\boldsymbol{X}_{[1]} - \boldsymbol{X}_{[2]}\right)^\top \widehat{\Sigma}_M^{-1} \left(\boldsymbol{X}_{[1]} - \boldsymbol{X}_{[2]}\right). \tag{2.43}$$

Let $\boldsymbol{X} = \boldsymbol{X}_{[1]} - \boldsymbol{X}_{[2]}$. Then $\text{var}(\boldsymbol{X}) = k_1\boldsymbol{\Sigma} + k_2\boldsymbol{\Sigma} = (k_1+k_2)\boldsymbol{\Sigma}$, where k_1 and k_2 are the proportionality constants of $\text{var}\left(\boldsymbol{X}_{[1]}\right)$ and $\text{var}\left(\boldsymbol{X}_{[2]}\right)$, respectively. The partition of the Mahalanobis distance is obtained by putting $\delta = 1$, $\boldsymbol{\mu} = \mathbf{0}$ and $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}_M$, where $\widehat{\boldsymbol{\Sigma}}_M$ is an unbiased estimate of $\boldsymbol{\Sigma}$.

(d) *Fisher's linear discriminant function.* Suppose $\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{1n_1}$ and $\boldsymbol{X}_{21}, \ldots, \boldsymbol{X}_{2n_2}$ are two random samples of sizes n_1 and n_2 from two multivariate normal populations π_1 and π_2 having the common covariance matrix $\boldsymbol{\Sigma}$. Let $\bar{\boldsymbol{X}}_1$ and $\bar{\boldsymbol{X}}_2$ be the sample mean vectors and $\mathbf{S}_1$ and $\mathbf{S}_2$ be the sample covariance matrices. Suppose $\widehat{\boldsymbol{\Sigma}}_p$ is the pooled estimate of the covariance matrix from

two samples. A new observation $\boldsymbol{X}_0$ will be allocate to π_1 if

$$\tau\left(\boldsymbol{X}_0\right) = \left(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2\right)^\top \widehat{\boldsymbol{\Sigma}}_p^{-1} \left[\boldsymbol{X}_0 - \frac{1}{2}\left(\bar{\boldsymbol{X}}_1 + \bar{\boldsymbol{X}}_2\right)\right] > 0. \tag{2.44}$$

Now $\operatorname{var}\left(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2\right) \propto \boldsymbol{\Sigma}$ and $\operatorname{var}\left[\boldsymbol{X}_0 - \frac{1}{2}\left(\bar{\boldsymbol{X}}_1 + \bar{\boldsymbol{X}}_2\right)\right] \propto \boldsymbol{\Sigma}$, hence the Garthwaite-Koch partition is valid. The transformations are of the form

$$\widehat{\boldsymbol{W}}^0 = \left(\widehat{\mathbf{D}}\widehat{\boldsymbol{\Sigma}}_p\widehat{\mathbf{D}}\right)^{-1/2} \widehat{\mathbf{D}}\left(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2\right) \tag{2.45}$$

and

$$\widehat{\boldsymbol{W}}^* = \left(\widehat{\mathbf{D}}\widehat{\boldsymbol{\Sigma}}_p\widehat{\mathbf{D}}\right)^{-1/2} \widehat{\mathbf{D}}\left[\boldsymbol{X}_0 - \frac{1}{2}\left(\bar{\boldsymbol{X}}_1 + \bar{\boldsymbol{X}}_2\right)\right]. \tag{2.46}$$

Let $\widehat{W}_j{}^0$ and $\widehat{W}_j{}^*$ denote the jth components of $\widehat{\boldsymbol{W}}^0$ and $\widehat{\boldsymbol{W}}^*$, respectively. Then as $\tau\left(\boldsymbol{X}_0\right) = \sum_{j=1}^{p} \widehat{W}_j{}^0 \widehat{W}_j{}^*$, the contribution of X_j to $\tau\left(\boldsymbol{X}_0\right)$ is $\widehat{W}_j{}^0 \widehat{W}_j{}^*$.

2.6 Concluding comments

In this chapter, we have reviewed some common transformations that transform the correlated variables to produce their orthogonal counterparts. We have also discussed their applications and properties along with some connections between the methods. These transformations will help in finding new statistical applications and deriving new methods. We have discussed only the literature that relates to more than one chapter. Literature relevant to only a particular chapter will be discussed in that chapter.

Chapter 3

Bootstrap confidence interval for the contribution of individual variables to a Mahalanobis distance

3.1 Introduction

In multivariate analysis, Mahalanobis distance (MD) is the most commonly used measure of distance between two vectors. It was proposed by Mahalanobis (1930) as a measure of the distance between groups that takes account of multiple characteristics and the correlations between these characteristics. The initial motivation was to analyze and classify human skulls into groups, based on multiple characteristics and the MD continues to be widely used in classification problems. Mahalanobis distance also underlies Hotelling's one-sample and two-sample T^2 tests:

it forms the test statistic when multiplied by appropriate constants determined from sample size(s).

To give some specific applications of MD, in environmental and health science it has been used to identify and map suitable habitats for a species. For instance, Liu and Weng (2012) calculated MD between a vector of environmental variables and the mean vector of environmental factors at the closest locations to mosquito infections. Small MD values indicated a more favourable habitat for mosquitoes. In multivariate calibration, MD is used to determine multivariate outliers (Martens and Naes, 1992) and evaluate the representativity between two multivariate data sets (Jouan-Rimbaud et al., 1998). In analytical chemistry, Shah and Gemperline (1990) used MD in pattern recognition to classify new samples by comparing them to a set of measurements of predetermined classes. In process control, MD is used in Hotelling's T^2-test to build multivariate control charts using the original or latent variables (Hotelling, 1933). In the field of wildlife biology, MD can be used to find the ideal landscape of some wildlife species. Clark et al. (1993) developed a multivariate model based on MD in a Geographic Information System (GIS) to identify areas of high habitat of female black bears.

When the value of an MD or Hotelling's T^2 is large, then an obvious question is *Which variables cause it to be large?* One approach to answering this question is to form a partition of the MD, where each element of the partition is associated with one variable and an element's size measures the contribution of the variable. Garthwaite and Koch (2016) recently proposed a partition of this form that has attractive properties. They partition the (squared) MD, $\Delta^2 = (\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu})$, by a linear transformation from the random vector

$\boldsymbol{X}$ to $\boldsymbol{W}$ such that $\boldsymbol{W}^\top\boldsymbol{W} = \Delta^2$ and the components of $\boldsymbol{W}$ are uncorrelated, with the transformation chosen to maximize the sum of correlations between corresponding elements of $\boldsymbol{X}$ and $\boldsymbol{W}$. Rogers (2015) developed global predictive risk maps for an important tropical disease, dengue, and used the partition to identify the most important predictors in determining the presence or absence of dengue in an area. Following Rogers, we refer to the partition as the Garthwaite-Koch partition.

The partition gives point estimates of the contribution of individual variables and scientists would often want interval estimates of these contributions. The task of forming confidence intervals for (un-partitioned) MDs has, of course, attracted attention. Madansky and Olkin (1969) provide an approximate confidence interval based on the asymptotic distribution of the likelihood ratio statistic. More recently, Reiser (2001) gave a method for constructing exact confidence intervals using the non-central F-distribution. A Bayesian approach is also possible (Radhakrishnan, 1984). In contrast, little work has been conducted on forming confidence intervals for the contribution to an MD given by the Garthwaite-Koch partition, although Garthwaite and Koch (2016) illustrated that bootstrap confidence intervals are readily constructed. Here we consider common non-parametric bootstrap methods for forming confidence intervals and propose new methods. We then use simulation to compare their performance for confidence intervals of individual contributions of variables to MD.

If we let $\boldsymbol{W} = (W_1, \ldots, W_p)^\top$, then the contribution of the jth variable is either expressed as an absolute value, W_j^2, or as a proportion $W_j^2 / \sum_{i=1}^{p} W_i^2$. The standard bootstrap pivotal methods apply a one-to-one transformation to W_j^2 or

$W_j^2 / \sum_{i=1}^p W_i^2$ and assume the transformed quantity is pivotal. Our new methods broaden the range of pivotal quantities that can be used. For W_j^2, a one-to-one functions of W_j is treated as a pivotal quantity. (Standard bootstrap methods cannot use a one-to-one function of W_j as a pivotal quantity because the function would not have a one-to-one mapping to W_j^2.) For $W_j^2 / \sum_{i=1}^p W_i^2$, a multivariate function of $(W_1, \ldots, W_p)$ is taken as a pivotal quantity.

In the simulation study, both equal-tailed and shortest intervals are constructed. An attraction of the shortest interval for W_j^2 is that its lower limit will be 0 if the equal-tailed interval for W_j contains 0. This is intuitively desirable, as only an upper bound for W_j^2 seems of interest when it is unclear whether W_j is positive or negative. This is also the case for $W_j^2 / \sum_{i=1}^p W_i^2$ if it is unclear whether $W_j / \{\sum_{i=1}^p W_i^2\}^{1/2}$ is positive or negative.

In Section 3.2, we briefly discuss the Mahalanobis distance and the Garthwaite-Koch partition. Brief description of bootstrapping is given in Section 3.3. In Section 3.4, we describe the methods used to construct bootstrap confidence intervals, including the new methods. An extensive simulation study is reported in Section 3.5 that examines the performance of methods when population distributions are multivariate normal. In Section 3.6, further simulations are reported where population distributions are skew. Concluding comments are given in Section 3.7.

3.2 Mahalanobis distance and the Garthwaite-Koch partition

Suppose we have two distinct groups (populations) which we shall label as π_1 and π_2. For example, π_1 and π_2 might represent genuine bank notes and fake bank notes, or, in a medical diagnosis situation, those with an illness and those without it. Each individual in these groups has a number (say, p) of variables or characteristics. These characteristics may include, for example, physical measurements such as length or height, and medical features, such as body temperature or blood pressure. Let $\boldsymbol{X}$ denote a (random) vector that contains the values of these variables on an item, individual or experimental unit.

Suppose the two populations have means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and share a common covariance matrix $\boldsymbol{\Sigma}$. Then the Mahalanobis distance between the two means is the nonnegative square root of

$$\Delta_1^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) . \tag{3.1}$$

Of course, the population parameters are rarely known and it is usual for them to be estimated by the corresponding sample values. Suppose we have two independent random samples of sizes n_1 and n_2 $(n_1 + n_2 = n)$ from populations π_1 and π_2, yielding sample means $\bar{\boldsymbol{X}}_1$ and $\bar{\boldsymbol{X}}_2$ and sample covariance matrices $\mathbf{S}_1$ and $\mathbf{S}_2$. If the populations have the same covariance $\boldsymbol{\Sigma}$, the sample Mahalanobis distance, D_1, can be similarly defined by

$$D_1^2 = (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)^\top \mathbf{S}^{-1} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2), \tag{3.2}$$

where $\mathbf{S} = \{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2\}/(n - 2)$ is an unbiased estimate of $\boldsymbol{\Sigma}$.

Hotelling's two-sample T^2 statistic is $\{n_1 n_2/(n_1+n_2)\}D_1^2$ and is used to test the hypothesis that $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are equal.

Other forms of Mahalanobis distance are also commonly of interest. The Mahalanobis distance between a vector $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ and the mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$ of a population with covariance matrix $\boldsymbol{\Sigma}$ is the nonnegative square root of

$$\Delta_2^2 = (\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}), \tag{3.3}$$

while a useful dissimilarity measure between two random vectors $\boldsymbol{X}_{[1]}$ and $\boldsymbol{X}_{[2]}$ drawn from a distribution with the common covariance matrix $\boldsymbol{\Sigma}$ is given by:

$$\Delta_3 = \left\{ \left(\boldsymbol{X}_{[1]} - \boldsymbol{X}_{[2]}\right)^\top \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{X}_{[1]} - \boldsymbol{X}_{[2]}\right) \right\}^{1/2}. \tag{3.4}$$

Also, Hotelling's one-sample T^2 statistic is

$$T_1^2 = n \left(\bar{\boldsymbol{X}} - \boldsymbol{\mu_0}\right)^\top \mathbf{S}^{-1} \left(\bar{\boldsymbol{X}} - \boldsymbol{\mu_0}\right) \tag{3.5}$$

when n observations are taken from a multivariate normal (MVN) distribution whose hypothesized mean is $\boldsymbol{\mu}_0$ and $\bar{\boldsymbol{X}}$ and $\mathbf{S}$ are the sample mean and covariance matrix.

The square of the Mahalanobis distance is often referred to as the *Mahalanobis Index* (MI) and we shall do so here. Garthwaite and Koch (2016) consider the MI given by equation (3.3) for the case where $\boldsymbol{X}$ is a $p \times 1$ random vector whose covariance matrix is proportional to $\boldsymbol{\Sigma}$, and $\boldsymbol{\mu}$ is a $p \times 1$ vector that is not necessarily the mean of $\boldsymbol{X}$. They address the task of partitioning Δ_2^2 into the contribution of individual variables, thus giving an evaluation of each variable's contribution to the MI.

Brief description of Garthwaite-Koch partition is given in Section 2.3. For more

detail see Garthwaite and Koch (2016). To obtain the partition they maximize $\sum_{j=1}^{p} \text{corr}\,(X_j, W_j)$ under the condition that

$$\boldsymbol{W}^\top \boldsymbol{W} = (\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \tag{3.6}$$

holds for all $\boldsymbol{X}$. This yields

$$\boldsymbol{W} = (\mathbf{D}\boldsymbol{\Sigma}\mathbf{D})^{-1/2} \mathbf{D} (\boldsymbol{X} - \boldsymbol{\mu}) \tag{3.7}$$

where $\mathbf{D}$ is a positive-definite diagonal matrix and $\mathbf{D}\boldsymbol{\Sigma}\mathbf{D}$ has diagonal elements of 1. That is, $\mathbf{D}$ has diagonal elements equal to the reciprocal of the square-root of the corresponding diagonal elements of $\boldsymbol{\Sigma}$. As $\sum_{j=1}^{p} \text{corr}(X_j, W_j)$ is maximized, each component X_j is identified with the corresponding component W_j in a one-to-one relationship. The sample estimate of $\boldsymbol{W}$, denoted by $\widehat{\boldsymbol{W}} = (\widehat{W}_1, \ldots, \widehat{W}_p)^\top$, is obtained by replacing $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with (unbiased) sample estimates.

In examples given in Garthwaite and Koch (2016), the partition always gives a sensible evaluation of the contributions of individual X variables. Rogers (2015) uses the partition for disease mapping and notes that "Identifying the key model variables in predicting the changing spatial pattern of vector-borne diseases over time is now made possible by the Garthwaite-Koch technique".

In general, the relative importance of variables when evaluated by the corr-max transformation may not match the relative importance given by variable selection methods. For example, Mardia et al. (1979, p.322–323) give an F- test for discarding variables from a discriminant function and the first variable to be discarded is not necessarily the one that is evaluated as least important by the corr-max transformation.

3.3 Bootstrapping

Bootstrapping is a resampling technique for estimating the sampling distribution of estimators and making inference about the corresponding parameters when there are no theoretical results on which to base inferences.

We assume that sample observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independent realizations of a random variable $\boldsymbol{X}$ that has probability density function $(p.d.f.)$, $f(.;\theta)$, and cumulative distribution function $(c.d.f.)$, $F(.;\theta)$. Suppose the parameter of interest is θ and $\hat{\theta}$ is its sample estimate. Since the sample estimate $\hat{\theta}$ is a function of sample observations it has a probability distribution, called sampling distribution of $\hat{\theta}$.

The bootstrap method assumes that the sampling distribution can be estimated from the large number of repeated samples. Two broad areas of approximating the sampling distribution of $\hat{\theta}$ are the parametric and non-parametric bootstrap.

In parametric problems, we assume that the data $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ comes from a known distribution such as Normal, Gamma, i.e., the functional form of the $p.d.f.$ is known, but the parameter values are unknown. The bootstrap resamples are obtained from $F(.;\hat{\theta})$, where $\hat{\theta}$ is the estimate of θ from the data. The estimate $\hat{\theta}$ is typically the maximum likelihood estimate of θ. The basic steps for the parametric bootstrap are as follows. [See, for example, Carpenter and Bithell (2000)].

(a) Sample n observations $\boldsymbol{x}^* = (\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_n^*)$ from $F(.;\hat{\theta})$, i.e., from the $c.d.f.$ with parameter values replaced by the sample statistic.

(b) Calculate $\hat{\theta}^*$ from $\boldsymbol{x}^*$ in the same way that $\hat{\theta}$ is estimated from $\boldsymbol{x}$.

(c) Repeat steps (a) and (b) N times, where N is large. Use the empirical dis-

tribution of the N bootstrap estimates $\hat{\theta}_1^*, \ldots, \hat{\theta}_N^*$ as an approximation of the sampling distribution of $\hat{\theta}$.

The basic idea of a non-parametric bootstrap is to generate samples of sizes n from an empirical distribution function $\hat{F}$. This empirical distribution function is a discrete probability distribution that assigns probability $1/n$ to each sample points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Thus a non-parametric bootstrap resample $\boldsymbol{x}^* = (\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_n^*)$ of size n is a with replacement sample from the data $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Steps in the algorithm for the non-parametric bootstrap are as follows. [See, for example, Ukoumunne et al. (2003)].

Suppose the data set $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is available from a target population with unknown probability distribution.

(a) Sample n observations randomly with equal probability and replacement from $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ to create a bootstrap data set of the same size as the original study data.

(b) Calculate the bootstrap version (replication) of the statistic of interest ($\hat{\theta}$) in the same way as for the study data set. Denote the estimate by $\hat{\theta}^*$.

(c) Repeat stages (a) and (b) N times, where N is large. The empirical distribution of the N bootstrap estimates is taken as an approximation of the sampling distribution of $\hat{\theta}$.

The mean and standard error of $\hat{\theta}$ are estimated by

$$\bar{\hat{\theta}}^* = \frac{1}{N} \sum_{k=1}^{N} \hat{\theta}_k^* \tag{3.8}$$

and

$$\hat{\sigma}(\hat{\theta}^*) = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(\hat{\theta}_k^* - \bar{\hat{\theta}}^*\right)^2}. \tag{3.9}$$

This $\hat{\sigma}(\hat{\theta}^*)$ is used as an approximation of the standard error of $\hat{\theta}$.

3.4 Bootstrap confidence intervals

Confidence intervals are a familiar data analysis tool and in this section we will discuss several methods for the construction of bootstrap confidence intervals.

The aim is to construct a confidence interval for some population characteristic, which we denote by θ. While $\boldsymbol{X}$ may be a vector, θ is a scalar. By a $(1-2\alpha)$ confidence interval, we mean selecting two scalar functions $L = L(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $U = U(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, of a sample such that $pr(L \leq \theta \leq U) = 1 - 2\alpha$, where $pr(.)$ denotes probability under the true distribution F. One general approach for constructing the confidence interval is to make it surround a point estimate of the parameter (Davison and Hinkley, 1997). Suppose that $\hat{\theta}$ estimates θ and that we want an equal-tailed interval with errors in both tails equal to α. We assume that $\hat{\theta}$ is continuous, for convenience. If the pth quantile of $\hat{\theta} - \theta$ is denoted by a_p, then

$$pr\left(\hat{\theta} - \theta \leq a_\alpha\right) = \alpha = pr\left(\hat{\theta} - \theta \geq a_{1-\alpha}\right). \tag{3.10}$$

Rearranging the event $\hat{\theta} - \theta \leq a_\alpha$ and $\hat{\theta} - \theta \geq a_{1-\alpha}$ as $\theta \geq \hat{\theta} - a_\alpha$ and $\theta \leq \hat{\theta} - a_{1-\alpha}$, respectively, we can find the $(1-2\alpha)$ equal-tailed interval as

$$\left(\hat{\theta} - a_{1-\alpha}, \hat{\theta} - a_\alpha\right). \tag{3.11}$$

We let $\hat{\theta}$ denote the estimate of θ given by the original data and $\hat{\theta}_1^\star, \ldots, \hat{\theta}_N^\star$ denote the estimates given by the N bootstrap replications.

Various methods have been proposed for constructing a confidence interval for θ from $\hat{\theta}^\star_1, \ldots, \hat{\theta}^\star_N$. For good reviews of a number of methods, see Efron and Tibshirani (1993), Davison and Hinkley (1997), and Carpenter and Bithell (2000). We will consider four commonly used methods: the percentile method, the bias-corrected percentile method, the non-studentized pivotal method, and the studentized pivotal method. We also propose new methods that deviate from the non-parametric algorithm for obtaining bootstrap replications of θ. The new methods introduce a parameter, γ say, that determines θ, but while θ must be a function of γ, the function need not be one-to-one. It is an estimate of γ that is determined from each bootstrap set and the estimates are manipulated to form a confidence interval for θ, using a method that has similarities to a bootstrap pivotal method (see Subsection 3.4.3).

In the present chapter the characteristic of interest, θ, reflects the contribution of the jth component of $\boldsymbol{X}$ to the Mahalanobis index. This contribution is defined to be W_j^2 $(j = 1, \ldots, p)$, where W_j is the jth component of $(\mathbf{D\Sigma D})^{-1/2}\,\mathbf{D}(\boldsymbol{X} - \boldsymbol{\mu})$. We examine bootstrap methods for forming confidence intervals for (i) W_j^2 and (ii) $W_j^2 / \sum_{i=1}^p W_i^2$. The latter quantity is the *proportion* of the MI that is attributable to the jth X variable and is a readily interpretable measure of the jth variable's importance.

We generally discuss central intervals, i.e. intervals (L, U) such that $pr(\theta \leq L) = pr(\theta \geq U) = \alpha$. The concept of shortest confidence intervals has attracted some attention [see, for example, Tate and Klett (1959) and Guenther (1969)]. Shortest confidence intervals can be preferred, because the length of the interval is the smallest possible that gives the required coverage probability. For symmetric

distributions an equal-tailed interval is also the shortest interval. The following elegant theorem (Casella and Berger, 2002) is applicable in some generality and gives a result for finding a shortest interval in the case of a unimodal distribution (symmetric or asymmetric).

Theorem 3. *Let $f(\hat{\theta})$ be a unimodal p.d.f. If the interval $[a, b]$ satisfies*

1. $\int_a^b f(\hat{\theta})d\hat{\theta} = 1 - 2\alpha$

2. $f(a) = f(b) > 0$ *and*

3. $a \leq \theta^\star \leq b$, *where $\theta^\star$ is the mode of $f(\hat{\theta})$,*

then $[a, b]$ is the shortest intervals among all intervals that satisfy $\int_a^b f(\hat{\theta})d\hat{\theta} = 1 - 2\alpha$.

As W_j^2 is non-negative, the distribution of its sample estimate will be markedly skew when the point estimate of W_j is near 0. Thus an equal-tailed interval will sometimes be markedly longer than the shortest interval that has the same level of confidence. Partly for this reason, we consider shortest confidence intervals as well as equal-tailed confidence intervals. The other reason is that we believe there should be some coherence between a confidence interval for W_j^2 and a confidence interval for W_j (and similarly with $W_j^2/\sum_{i=1}^p W_i^2$ and $W_j/\{\sum_{i=1}^p W_i^2\}^{1/2}$). So alternatively, we also consider forming bootstrap estimates for the un-squared quantity, W_j, and form a confidence interval for W_j^2 from the squares of these estimates. With regard to this, consider the question: *What is a sensible confidence interval for W_j^2 if the confidence interval for W_j includes 0?* When W_j has the sampling distribution given in Figure 3.1(a), then W_j^2 has the sampling distribution in Figure 3.1(b). The equal-tailed confidence intervals for W_j is indicated in

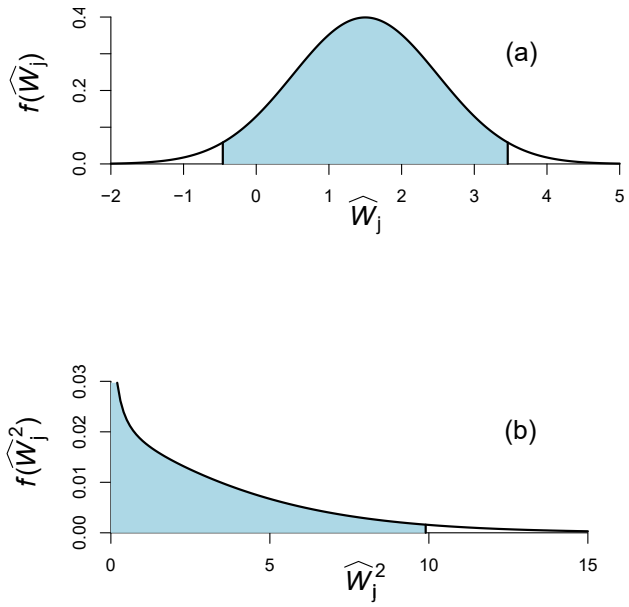

Figure 3.1: (a) Probability distribution for $\widehat{W}_j$ and equal-tailed confidence interval for W_j, and (b) corresponding probability distribution for $\widehat{W}_j^2$ and shortest confidence interval for W_j^2.

Figure 3.1(a) and the shortest interval for W_j^2 is marked in Figure 3.1(b). The latter interval not only includes all the plausible for W_j^2, but also the square of the most plausible values for W_j. This is not true of an equal-tailed confidence intervals for W_j^2, as the interval would not contain 0. (Obviously the shortest confidence intervals for W_j^2 will not have 0 as its lower endpoint when the sign of W_j is clear.)

As Figures 3.1(a) and 3.1(b) illustrate, a shortest confidence interval for W_j^2 should sometimes have 0 as its lower endpoint. However, an empirical bootstrap distribution is discrete and the smallest bootstrap estimate of W_j^2 is unlikely to equal 0 precisely. It follows that the lower endpoint of a confidence interval will not equal 0, as the lower endpoint cannot be less than the smallest bootstrap estimate. As a 'continuity correction', for a 95% confidence interval (for both W_j^2 and $W_j^2/\sum_{i=1}^{p} W_i^2$) we form intervals for the following combinations of (α_1, α_2): (0.000, 0.050), (0.005, 0.045), ... , (0.045, 0.005) and (0.050,0.000), where the lower

interval endpoint is taken as 0 when $\alpha_1 = 0$. We observe which of the intervals is shortest and take that as the 'shortest' confidence interval. This often gives 0 as the lower limit. A similar approach can be used for other levels of confidence, though here we consider the 95% level.

We next describe the bootstrap methods of interest in this chapter. We suppose we wish to form a confidence interval for which the lower and upper tail areas are α_1 and α_2, respectively. For an equal-tailed $100(1-2\alpha)\%$ confidence interval, $\alpha_1 = \alpha_2 = \alpha$. For a shortest interval, α_1 and α_2 are varied, subject to $\alpha_1 + \alpha_2 = 2\alpha$.

3.4.1 Percentile methods

Percentile method

The simplest method of forming a bootstrap confidence interval is the percentile method suggested by Efron (1981), which simply equates quantiles of the distribution of $\hat{\theta}$ to the equivalent quantiles of the bootstrap distribution of $\hat{\theta}^*$. This gives $\left(\hat{\theta}^*(\alpha_1), \hat{\theta}^*(1-\alpha_2)\right)$ as a $100(1-2\alpha)\%$ confidence interval for θ, where $\hat{\theta}^*(q)$ denotes the qth quantile of the bootstrap distribution. Hall (1992, p.19) refers to Efron's (1981) percentile method as the "other percentile method". However, for our discussion we will use 'percentile method' to indicate Efron's percentile method.

This method has simplicity, can be applied to any statistic, and no invalid parameter values will be included in the confidence interval, as the method is range-preserving. Also, the method is transformation respecting, implying that if (θ_L, θ_U) is a $100(1-2\alpha)\%$ confidence interval for θ and g is a monotonic increasing transformation of θ, then $(g(\theta_L), g(\theta_U))$ is a $100(1-2\alpha)\%$ confidence interval for

$g(\theta)$. Largely for these reasons, the method is widely used. However, if the distribution of $\hat{\theta}$ is markedly skew, the coverage error of equal-tailed intervals is often substantial (Efron and Tibshirani, 1993).

Bias-corrected percentile method

Efron (1981) introduced the bias-corrected percentile (BC method) method. The BC method is a modification of the percentile method that aims to improve coverage for non-symmetric distributions. Its steps are as follows:

1. Let $\hat{\theta}_k^\star$ denote the estimate of θ given by the kth bootstrap resample. Count the number of members of $\hat{\theta}_1^\star, \hat{\theta}_2^\star, \ldots, \hat{\theta}_N^\star$ that are less than $\hat{\theta}$ (calculated from the original data). Call this number p and set $p^\star = p/N$. Set $z_0 = \Phi^{-1}(p^\star)$, where Φ denotes the *c.d.f.* of the standard normal distribution.

2. Define $\hat{\alpha}_1$ and $\hat{\alpha}_2$ as $\hat{\alpha}_1 = \Phi(2z_0 + z_1)$ and $\hat{\alpha}_2 = 1 - \Phi(2z_0 + z_2)$, where $z_1 = \Phi^{-1}(\alpha_1)$ and $z_2 = \Phi^{-1}(1 - \alpha_2)$.

3. Take $\hat{\theta}^\star(\hat{\alpha_1})$ and $\hat{\theta}^\star(1 - \hat{\alpha_2})$ as the endpoints of the confidence interval.

This method is as easily implemented as the percentile method and is transformation respecting. If the distribution of $\hat{\theta}^\star$ is symmetric about $\hat{\theta}$, that is when $z_0 = 0$, the bias-corrected percentile interval and percentile interval are the same. Hence the method may be thought of as a "fine-tuning" of the percentile method. However even a small departure between $pr(\hat{\theta}^*_{Median} < \hat{\theta})$ and 0.50 can lead to a substantial difference between the endpoint of the BC method and that of the percentile method (Efron, 1982, p,75-90).

3.4.2 Pivotal methods

Pivotal methods form a function of θ and $\hat{\theta}$ that is treated as pivotal: it is assumed that the sampling distribution of the function does not depend upon any unknown quantities. The most commonly used functions for the non-studentized and studentized pivotal methods are $\hat{\theta} - \theta$ and $(\hat{\theta} - \theta)/\sigma(\hat{\theta})$, respectively, where $\sigma(\hat{\theta})$ is the standard error of $\hat{\theta}$.

The methods can yield confidence intervals that include invalid parameter values if the range of θ is bounded. Moreover, neither $\hat{\theta} - \theta$ nor $(\hat{\theta} - \theta)/\sigma(\hat{\theta})$ could be a pivotal function if the range of θ is bounded. Transformations are the usual approach to counter this problem. If the parameter of real interest has a bounded range, then θ is equated to some monotonic increasing function of the parameter. A confidence interval for θ is constructed and its endpoints transformed back, giving a confidence interval for the true parameter of interest. However, pivotal methods are not transformation respecting, so the choice of transformation will affect the endpoints of confidence intervals. Here we put $\theta = \log W_j^2$ or $\theta = \text{logit}[W_j^2 / \sum_{i=1}^{p} W_i^2]$ when seeking a confidence interval for W_j^2 or $W_j^2 / \sum_{i=1}^{p} W_i^2$, respectively. In both cases the resulting θ has a range of $(-\infty, \infty)$.

Non-studentized pivotal (basic) method

The non-studentized pivotal method makes the assumption that the distribution of $\psi = \hat{\theta} - \theta$ is similar to the distribution of $\hat{\psi}^\star = \hat{\theta}^\star - \hat{\theta}$. Quantiles of the bootstrap distribution of $\hat{\psi}^\star$ are used to form confidence intervals for θ. The steps are as follows:

(a) Set $\hat{\psi}_k^\star = \hat{\theta}_k^\star - \hat{\theta}$, for $k = 1, 2, \ldots, N$.

(b) Determine $\hat{\psi}^\star(\alpha_2)$ and $\hat{\psi}^\star(1-\alpha_1)$, where $\hat{\psi}^\star(q)$ denotes the qth quantile of the bootstrap distribution of $\hat{\psi}^\star$.

(c) Then a $100(1-2\alpha)\,\%$ confidence interval for θ is given by $(\hat{\theta} - \hat{\psi}^\star(1-\alpha_1),\ \hat{\theta} - \hat{\psi}^\star(\alpha_2))$, which can equivalently be written as $(2\hat{\theta} - \hat{\theta}^\star(1-\alpha_1),\ 2\hat{\theta} - \hat{\theta}^\star(\alpha_2))$.

This method is simple to use, but the coverage error can be substantial if the distributions of ψ and $\hat{\psi}^\star$ differ in a clearly noticeably manner.

Studentized pivotal method (bootstrap t method)

The bootstrap t method aims to improve on the basic bootstrap method by treating $(\hat{\theta} - \theta)/\hat{\sigma}(\hat{\theta})$ as a pivotal quantity, where $\hat{\sigma}(\hat{\theta})$ denotes the estimated standard error of $\hat{\theta}$. It derives its name from the fact that when $\hat{\theta} \sim N(\theta, \sigma^2)$, then $(\hat{\theta} - \theta)/\hat{\sigma}(\hat{\theta})$ is a pivotal quantity that has a t-distribution. The method assumes that $(\hat{\theta} - \theta)/\hat{\sigma}(\hat{\theta})$ and $(\hat{\theta}^\star - \hat{\theta})/\hat{\sigma}(\hat{\theta}^\star)$ have similar distributions. The following are its primary steps.

(a) Set $\hat{\xi}_k^\star = (\hat{\theta}_k^\star - \hat{\theta})/\hat{\sigma}_k(\hat{\theta}^\star)$, for $k = 1, 2, \ldots, N$, where $\hat{\sigma}_k(\hat{\theta}^\star)$ is an estimate of the standard error of $\hat{\theta}_k^\star$ (see below).

(b) Determine $\hat{\xi}^\star(\alpha_2)$ and $\hat{\xi}^\star(1-\alpha_1)$, where $\hat{\xi}^\star(q)$ denotes the qth quantile of the bootstrap distribution of $\hat{\xi}^\star$.

(c) Then a $100(1-2\alpha)\,\%$ confidence interval for θ is given as

$$(\hat{\theta} - \hat{\sigma}(\hat{\theta})\hat{\xi}^\star(1-\alpha_1),\ \hat{\theta} - \hat{\sigma}(\hat{\theta})\hat{\xi}^\star(\alpha_2)).$$

The method requires estimates $\hat{\sigma}(\hat{\theta})$ and $\hat{\sigma}_k(\hat{\theta}^\star)$. For some statistics, the standard error can be estimated from known formula. More complicated statistics have no formula for estimating standard error. However, a variety of analytical approximations exists in the literature (see, for example Davison and Hinkley (1997), Chapters 2 and 3). An estimate of $\hat{\sigma}(\hat{\theta})$ is obtained from

$$\hat{\sigma}^2(\hat{\theta}^*) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\hat{\theta}_k^\star - \bar{\hat{\theta}}^\star\right)^2, \tag{3.12}$$

where $\bar{\hat{\theta}}^\star$ is the mean of $\hat{\theta}_1^\star, \ldots, \hat{\theta}_N^\star$. An estimate of $\hat{\sigma}_k(\hat{\theta}^*)$ might be estimated using the jackknife. The alternative, that we use, is to carry out a computationally intensive, but routine, 'second-level bootstrap' to estimate $\hat{\sigma}_k(\hat{\theta}^\star)$, as follows.

Let $\boldsymbol{x}_{k1}^\star, \ldots, \boldsymbol{x}_{kn}^\star$ be the kth ($k = 1, 2, \ldots, N$) bootstrap sample and let $\hat{\theta}_k^\star$ denote the estimate of θ it gives. Obtain a second-level bootstrap sample $\boldsymbol{x}_{k1}^{\star\star}, \ldots, \boldsymbol{x}_{kn}^{\star\star}$ by sampling with replacement from $\boldsymbol{x}_{k1}^\star, \ldots, \boldsymbol{x}_{kn}^\star$ and evaluate the estimate of θ. Repeat this B times and let $\hat{\theta}_{k\ell}^{\star\star}$ denote the estimate given by the ℓth second-level sample ($\ell = 1, 2, \ldots, B$). Then the estimate of the variance of $\hat{\theta}_k^\star$ is

$$\hat{\sigma}_k^2(\hat{\theta}^\star) = \frac{1}{B-1}\sum_{\ell=1}^{B}(\hat{\theta}_{k\ell}^{\star\star} - \bar{\hat{\theta}}_k^\star)^2, \tag{3.13}$$

where $\bar{\hat{\theta}}_k^\star$ is the mean of $\hat{\theta}_{k1}^{\star\star}, \hat{\theta}_{k2}^{\star\star}, \ldots, \hat{\theta}_{kB}^{\star\star}$.

From each bootstrap resample, at least 25 second-level bootstrap samples should be taken (Carpenter and Bithell, 2000). The obvious drawback of the studentized pivotal method is that the process is computationally intensive — to generate a total of N values of $\hat{\theta}^\star$, a total of BN bootstrap samples are required. The method can perform very poorly if $\hat{\sigma}(\hat{\theta})$ is not independent of θ, but simulation results reported in the literature (Carpenter and Bithell, 2000) suggest that the method often gives more accurate coverage than other bootstrap methods.

3.4.3 New methods

The new methods broaden the range of pivotal quantities that can be used to form bootstrap confidence intervals for θ. The methods are partly Bayesian, in that parameters are treated as variables that have probability distributions — but no prior distributions are specified.

Let $\theta = h(\gamma)$, where γ may be a vector and h is not necessarily a monotonic function, nor necessarily a one-to-one function. The sample data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ yield an estimate $\widehat{\gamma}$ of γ. From a bootstrap resample, we determine an estimate $\widehat{\gamma}^*$ of $\widehat{\gamma}$ in the same way as $\widehat{\gamma}$ was determined from the original sample. Let $\widehat{\gamma}^*_1, \ldots, \widehat{\gamma}^*_N$ denote the estimates given by the N resamples.

When seeking a confidence interval for W_j^2, we set $\theta = W_j^2$ and $\gamma = W_j$. For an interval for $W_j^2 / \sum_{i=1}^p W_i^2$, we put $\theta = W_j^2 / \sum_{i=1}^p W_i^2$ and $\gamma = (W_1, \ldots, W_p)^\top$.

Method A

The first of our new methods, Method A, treats $\widehat{\gamma} - \gamma$ as a pivotal quantity and makes the following assumption.

Assumption A. Given any γ, the statistics $\widehat{\gamma} - \gamma$ and $\widehat{\gamma}^* - \widehat{\gamma}$ are from the same distribution.

Let $\widehat{P}_{\widehat{\gamma}^*|\widehat{\gamma}}$ denote bootstrap probabilities when $\widehat{\gamma}^*$ is considered a random variable and $\widehat{\gamma}$ is non-random. Similarly, let $\widehat{P}_{\widehat{\gamma}|\gamma}$ denote bootstrap probabilities when $\widehat{\gamma}$ is random while γ is non-random. We have,

$$\widehat{P}_{\widehat{\gamma}^*|\widehat{\gamma}}(\widehat{\gamma}^* - \widehat{\gamma} = \nu) \;=\; \begin{cases} 1/N & \nu = \widehat{\gamma}^*_k - \widehat{\gamma};\; k = 1, \ldots, N \\ 0 & \text{otherwise,} \end{cases} \tag{3.14}$$

so, from Assumption A,

$$\widehat{P}_{\widehat{\gamma}|\gamma}(\widehat{\gamma} - \gamma = \nu) = \begin{cases} 1/N & \nu = \widehat{\gamma}_k^* - \widehat{\gamma};\ k = 1, \ldots, N \\ 0 & \text{otherwise.} \end{cases} \tag{3.15}$$

To add flexibility, we wish to allow γ to have a probability distribution, so we adopt a Bayesian approach and let $\widehat{P}_{\gamma|\widehat{\gamma}}$ denote bootstrap posterior probabilities, where γ is now random while $\widehat{\gamma}$ is non-random. So that Bayesian credible intervals match frequentist confidence intervals, we assume that γ has a probability matching prior distribution. (See, for example, Datta and Mukerjee (2004) for details of probability matching priors.) Then, from equation (3.15),

$$\widehat{P}_{\gamma|\widehat{\gamma}}(\widehat{\gamma} - \gamma = \nu) = \begin{cases} 1/N & \nu = \widehat{\gamma}_k^* - \widehat{\gamma};\ k = 1, \ldots, N \\ 0 & \text{otherwise,} \end{cases} \tag{3.16}$$

so

$$\widehat{P}_{\gamma|\widehat{\gamma}}(\gamma = \eta) = \begin{cases} 1/N & \eta = 2\widehat{\gamma} - \widehat{\gamma}_k^*;\ k = 1, \ldots, N \\ 0 & \text{otherwise.} \end{cases} \tag{3.17}$$

Let $\widehat{P}_{\theta|\widehat{\gamma}}$ denote the bootstrap posterior probability distribution of θ. As $\theta = h(\gamma)$, equation (3.17) gives

$$\widehat{P}_{\theta|\widehat{\gamma}}[\theta = h(\eta)] = \begin{cases} 1/N & \eta = 2\widehat{\gamma} - \widehat{\gamma}_k^*;\ k = 1, \ldots, N \\ 0 & \text{otherwise.} \end{cases} \tag{3.18}$$

Quantiles from the distribution in equation (3.18) yield a Bayesian credible interval for θ and, because a probability matching prior has been used, we take this as the confidence interval. That is, if $\theta_A^{\#}(q)$ denotes the qth quantile of $\widehat{P}_{\theta|\widehat{\gamma}}(\theta)$ given by (3.18), then $(\theta_A^{\#}(\alpha_1), \theta_A^{\#}(1-\alpha_2))$ is the $100(1-2\alpha)\%$ confidence interval for θ given by Method A.

The following summarises the steps for Method A.

1. Determine $\widehat{\gamma}_1^*, \ldots, \widehat{\gamma}_N^*$ from the N bootstrap resamples.

2. Put $\widehat{\vartheta}_k^* = h(2\widehat{\gamma} - \widehat{\gamma}_k^*)$ for $k = 1, \ldots, N$.

3. Then a $100(1-2\alpha)\%$ confidence interval for θ is given as $(\widehat{\vartheta}^*(\alpha_1), \widehat{\vartheta}^*(1-\alpha_2))$, where $\widehat{\vartheta}^*(q)$ is the qth sample quantile of the sample $\widehat{\vartheta}_1^*, \ldots, \widehat{\vartheta}_N^*$.

Method B

Suppose γ is an r-dimensional vector. Let $\widehat{\sigma}(\widehat{\gamma}_{(j)})$ and $\widehat{\sigma}(\widehat{\gamma}_{(j)}^*)$ denote the estimated standard errors of the jth components of $\widehat{\gamma}$ and $\widehat{\gamma}^*$, respectively. Also, let $\widehat{\tau}(\widehat{\gamma})$ and $\widehat{\tau}(\widehat{\gamma}^*)$ denote $r \times r$ diagonal matrices whose jth diagonal elements are $[\widehat{\sigma}(\widehat{\gamma}_{(j)})]^{-1}$ and $[\widehat{\sigma}(\widehat{\gamma}_{(j)}^*)]^{-1}$, respectively. Method B, makes the following assumption.

Assumption B. Given any γ, the statistics $\widehat{\tau}(\widehat{\gamma})(\widehat{\gamma} - \gamma)$ and $\widehat{\tau}(\widehat{\gamma}^*)(\widehat{\gamma}^* - \widehat{\gamma})$ are from the same distribution.

Consequently, the difference between Method A and Method B is similar to the difference between the non-studentized and studentized pivotal methods. Corresponding to equations (3.14) and (3.16), the bootstrap probabilities are

$$\widehat{P}_{\widehat{\gamma}^*|\widehat{\gamma}}\left[\widehat{\tau}\{\widehat{\gamma}^*)(\widehat{\gamma}^* - \widehat{\gamma}\} = \nu\right] = \begin{cases} 1/N & \nu = \widehat{\tau}_k(\widehat{\gamma}^*)\{\widehat{\gamma}_k^* - \widehat{\gamma}\};\ k = 1, \ldots, N \\ 0 & \text{otherwise,} \end{cases} \tag{3.19}$$

and the resulting bootstrap posterior probabilities are:

$$\widehat{P}_{\gamma|\widehat{\gamma}}\left[\widehat{\tau}(\widehat{\gamma})\{\widehat{\gamma} - \gamma\}\} = \nu\right] = \begin{cases} 1/N & \nu = \widehat{\tau}_k(\widehat{\gamma}^*)\{\widehat{\gamma}_k^* - \widehat{\gamma}\};\ k = 1, \ldots, N \\ 0 & \text{otherwise.} \end{cases} \tag{3.20}$$

This yields the bootstrap posterior distribution of θ:

$$\widehat{P}_{\theta|\widehat{\gamma}}[\theta = h(\eta)] = \begin{cases} 1/N & \eta = \widehat{\gamma} - \{\widehat{\tau}(\widehat{\gamma})\}^{-1}\widehat{\tau}_k(\widehat{\gamma}^*)\{\widehat{\gamma}_k^* - \widehat{\gamma}\};\ k = 1, \ldots, N \\ 0 & \text{otherwise.} \end{cases} \tag{3.21}$$

Let $\theta_B^{\#}(q)$ denote the qth quantile of $\widehat{P}_{\theta|\widehat{\gamma}}(\theta)$ given by (3.21). Then $(\theta_B^{\#}(\alpha_1), \theta_B^{\#}(1-\alpha_2))$ is the $100(1-2\alpha)\%$ confidence interval for θ given by Method B.

Method B requires estimates of the standard errors, $\widehat{\sigma}(\widehat{\gamma}_{(j)})$ and $\widehat{\sigma}_k(\widehat{\gamma}^*_{(j)})$ for $k = 1, \ldots, N$; $j = 1, \ldots, r$. These are obtained in a way analogous to the procedure for obtaining $\widehat{\sigma}(\widehat{\theta})$ and $\widehat{\sigma}_k(\widehat{\theta}^*)$ in the studentized pivotal method. The following summarises the steps in Method B.

1. Generate N bootstrap resamples to obtain estimates $\widehat{\gamma}^*_1, \ldots, \widehat{\gamma}^*_N$. The sample standard deviation of the jth components of the $\widehat{\gamma}^*_k$ is taken as $\widehat{\sigma}(\widehat{\gamma}_{(j)})$. The diagonal elements of $\widehat{\tau}(\widehat{\gamma})$ are set equal to $\{\widehat{\sigma}(\widehat{\gamma}_{(1)})\}^{-1}, \ldots, \{\widehat{\sigma}(\widehat{\gamma}_{(r)})\}^{-1}$.

2. From the kth bootstrap resample ($k = 1, \ldots, N$), generate B second-level bootstrap samples and estimate γ in each. Let $\widehat{\gamma}^{**}_{kl}$ denote the estimate of γ given by the lth second-level sample ($l = 1, \ldots, B$). The sample standard deviation of the jth components of the $\widehat{\gamma}^{**}_{kl}$ is taken as $\widehat{\sigma}_k(\widehat{\gamma}^*_{(j)})$. For $k = 1, \ldots, N$, the diagonal elements of $\widehat{\tau}_k(\widehat{\gamma^*})$ are set equal to $\{\widehat{\sigma}_k(\widehat{\gamma}^*_{(1)})\}^{-1}, \ldots, \{\widehat{\sigma}_k(\widehat{\gamma}^*_{(r)})\}^{-1}$.

3. Put $\widehat{\lambda}^*_k = h[\widehat{\gamma} - \{\widehat{\tau}(\widehat{\gamma})\}^{-1}\widehat{\tau}_k(\widehat{\gamma}^*)\{\widehat{\gamma}^*_k - \widehat{\gamma}\}]$ for $k = 1, \ldots, N$.

4. Then a $100(1-2\alpha)\%$ confidence interval for θ is given as $(\widehat{\lambda}^*(\alpha_1), \widehat{\lambda}^*(1-\alpha_2))$, where $\widehat{\lambda}^*(q)$ is the qth sample quantile of the sample $\widehat{\lambda}^*_1, \ldots, \widehat{\lambda}^*_N$.

When γ is a scalar and h is a monotonic transformation, standard bootstrap methods can be used to first form a bootstrap confidence interval for γ and then the endpoints of the interval can be back-transformed to obtain a bootstrap confidence interval for θ. If the non-studentized pivotal method is used to form the bootstrap confidence for γ, then the resulting confidence interval for θ is identical to the

interval given by Method A. If the studentized pivotal method is used, the resulting confidence interval is identical to the interval given by Method B. The advantages of Methods A and B are that they can be used when γ is not a scalar and h is not a monotonic transformation.

The new methods (Methods A and B) construct pivotal quantities before squaring the bootstrap estimates. With standard pivotal methods, only two bootstrap estimates are pivoted: the αth smallest and the αth biggest. With methods A and B, all the bootstrap estimates are pivoted as ordering is not maintained when a set of positive and negative estimates are squared. For an interval estimate of W_j^2, we set $\gamma = W_j$. Let $\widehat{\gamma}_{jk}^*$ denote the bootstrap estimate of W_j given by the kth sample ($j = 1, \ldots, p$; $k = 1, \ldots, N$) and let $\widehat{\gamma}_j$ denote the estimate given by the original data. We then consider $\widehat{\gamma}_j - \gamma_j$ as pivotal quantity for Method A and $(\widehat{\gamma}_j - \gamma_j)/\hat{\sigma}(\widehat{\gamma}_j)$ as pivotal quantity for Method B, where $\hat{\sigma}(\widehat{\gamma}_j)$ is an estimate of the standard error of $\widehat{\gamma}_j$. For Method A put

$$\widetilde{W}_{jk}^* = 2\widehat{\gamma}_j - \widehat{\gamma}_{jk}^* \tag{3.22}$$

and for Method B put

$$\widetilde{W}_{jk}^* = \widehat{\gamma}_j - \hat{\sigma}(\widehat{\gamma}_j)\{\widehat{\gamma}_{jk}^* - \widehat{\gamma}_j\}/\hat{\sigma}_k(\widehat{\gamma}_j^*), \tag{3.23}$$

where $\hat{\sigma}(\widehat{\gamma}_j)$ is an estimate of the standard error of $\widehat{\gamma}_j$ and $\hat{\sigma}_k(\widehat{\gamma}_j^*)$ is an estimate of the standard error of $\widehat{\gamma}_{jk}^*$. These are obtained in the similar way of obtaining $\hat{\sigma}(\hat{\theta})$ and $\hat{\sigma}_k(\hat{\theta}^*)$ in the studentized pivotal methods. The interval estimates of W_j^2 and $W_j^2/\sum_{i=1}^p W_i^2$ are obtained from the bootstrap distributions of $(\widetilde{W}_{jk}^*)^2$ and its percentage respectively. Let $\widetilde{\theta}_{jk}^* = (\widetilde{W}_{jk}^*)^2$ and $\breve{\theta}_{jk}^* = (\widetilde{W}_{jk}^*)^2/\sum_{i=1}^p (\widetilde{W}_{ik}^*)^2$. Then $(\widetilde{\theta}_j^*(\alpha_1), \widetilde{\theta}_j^*(1-\alpha_2))$ and $(\breve{\theta}_j^*(\alpha_1), \breve{\theta}_j^*(1-\alpha_2))$ are taken as the bootstrap confidence

intervals for W_j^2 and $W_j^2 / \sum_{i=1}^{p} W_i^2$, respectively, where $\widetilde{\theta}_j^*(q)$ and $\breve{\theta}_j^*(q)$ are the qth quantiles of the empirical distributions given by the $\widetilde{\theta}_{jk}^*$ and $\breve{\theta}_{jk}^*$. Since the new methods consider quantiles of the squared quantities the lower endpoints of the confidence intervals of individual contributions and their percentages cannot be negative.

3.5 Simulation Study: Multivariate Normal Distribution

A large simulation study was conducted to evaluate the coverage probabilities of the six methods. In this section we use a multivariate normal distribution to describe each population because this is consistent with the assumptions underlying Hotelling's T^2 hypothesis test and the test of whether a Mahalanobis distance is unusually large. In Subsections 3.5.1 – 3.5.3 the mechanics of the simulations are described and results are presented in Subsection 3.5.4.

3.5.1 Population distributions

We require a number of known population distributions. To mimic reality, we set the mean and variance of each population distribution equal to the sample mean and variance of a real dataset, using the following five datasets.

1. *Swiss bank notes:* The dataset is given in Flury and Riedwyl (1988). It contains six measurements that were made on 100 genuine bank notes and 100 forged bank notes. The measurements were: *length* (length of bank note), *left* (width of note, measured on its left side), *right* (width of note,

measured on the right), *bottom* (width of margin at the bottom), *top* (width of margin at the top) and *diagonal* (length of the image diagonal). All variables were measured in millimetres.

2. *Male and female athletes:* Data on 102 male and 100 female athletes were collected at the Australian Institute of Sport (Cook and Weisberg, 1994). For our study we considered the following nine measurements on each athlete: *Wt* (weight in kg), *Ht* (height in cm), *RCC* (red cell count), *Hg* (haemoglobin), *Hc* (hematocrit), *WCC* (white cell count), *Ferr* (plasma ferritin concentration), *Bfat* (% body fat) and *SSF* (sum of skin fold thickness).

3. *Tibetan skulls:* Data reported in Morant (1923) were collected from south-western and eastern districts of Tibet. Five measurements (all in millimetres) were made on each of 32 skulls: *Length* (greatest length of skull), *Breadth* (greatest horizontal breadth of skull), *Height* (height of skull), *Fheight* (upper face length) and *Fbreadth* (face breadth between outermost points of cheekbones). The first 17 skulls come from graves in Sikkim and the neighbouring area of Tibet, and the remaining 15 were from a battlefield in the Lhasa district.

4. *Psychological measurements:* Beall (1945) gives data on 32 men and 32 women. Four psychological measurements were made on each person: *Pi* (pictorial inconsistencies), *Tr* (tool recognition), *Pb* (paper from board) and *Vo* (vocabulary).

5. *Flea Beetles:* Lubischew (1962) gives data on two species of flea beetles

Table 3.1: Features of the data sets

Dataset	Sample sizes	No. of variables	Range of absolute correlations
Bank notes	100 & 100	6	0.000 to 0.664
Athletes	102 & 100	9	0.017 to 0.967
Skulls	17 & 15	5	0.011 to 0.718
Psychological tests	32 & 32	4	0.322 to 0.628
Flea beetles	19 & 20	4	0.074 to 0.727

(*Haltica Oleracea* and *Haltica Carduorum*). Four measurements were made on each flea: *Dt* (distance of transverse groove from posterior border of prothorax (μm)), *Le* (length of elytra (0.01 mm)), *Ls* (length of second antenatal joint (μm)) and *Lt* (length of third antenatal joint (μm)).

Table 3.1 summarizes some key features of the datasets: their sizes, the number of variables considered here, and the range of correlations between variables. It can be seen that the datasets vary in size from quite small (15) to moderately large (102) and each set contains between 4 and 9 variables. The correlations between variables vary from small to moderate in most datasets, with the largest correlation in a dataset generally lying between 0.62 to 0.72. The Athletes dataset is an exception with some correlations above 0.95.

3.5.2 Simulation procedure for the bank note dataset

To simplify explanation we first focus on the bank notes dataset. We calculated the mean and covariance of the 100 genuine bank note measurements and took these as the mean ($\boldsymbol{\mu}$) and variance ($\boldsymbol{\Sigma}$) of an MVN population distribution describing genuine bank notes. We also took one of the fake bank notes and examined how its measurements ($\boldsymbol{x}$) distinguish it from the genuine bank notes. To this end, we calculated $(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$, the MI (squared MD) between $\boldsymbol{x}$ and $\boldsymbol{\mu}$,

and then applied the Garthwaite-Koch partition to evaluate the contributions of individual variables. Through simulation we investigated different ways of forming confidence intervals for these contributions (W_j^2) and their percentages ($100\% \times W_j^2 / \sum_{i=1}^{p} W_i^2$). We examined various sample sizes (20, 50, 80, 100 and 200) and ten fake bank notes (the first ten in the dataset). For simplicity we describe the simulation procedure for samples of size twenty.

We generated one data sample of size 20 from an MVN($\boldsymbol{\mu}, \boldsymbol{\Sigma}$) distribution and then generated 1000 bootstrap resamples from this data sample. Each resample was a random sample of size 20 drawn with replacement from the data sample. We took one of the first ten fake bank notes and calculated the MI between that bank note and the mean of the resample, using the estimated covariance matrix of the resample, $\widehat{\boldsymbol{\Sigma}}$. Then estimates of $W_1, \ldots, W_6$ from this resample were calculated using the Garthwaite-Koch partition. This gives 1000 estimates of each W_j. From each bootstrap sample, 25 second-level bootstrap samples were generated so as to determine (approximate) standard errors of the estimates. The standard errors were needed for the studentized pivotal method and Method B. After generating the bootstrap samples and second-level samples, 95% confidence intervals for individual contributions and their percentages were calculated using the methods discussed in Section 3.4.

As noted above, the Garthwaite-Koch partition was also applied to $(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$. This yielded 'true values' for individual contributions and their percentages, and we determined which confidence intervals covered their target values. The procedure was repeated 1000 times by generating 1000 data samples, from which we estimated coverage probabilities of the confidence intervals for each

variable's contributions and their percentage contributions.

3.5.3 Simulation procedures for other datasets

Simulation procedures for the Athletes dataset and the Tibetan skulls dataset were the same as for the bank note dataset. For the Athletes dataset, the male athletes took the role of the genuine bank notes so their sample mean and covariance matrix became the mean and variance of the population distribution. The first ten female athletes took the role of the first ten fake bank notes, so the MIs from each of these female athletes to the mean of the men were the quantity of interest. For the Tibetan skulls dataset, the 17 skulls from the Sikkim area took the role of the genuine bank notes while the first ten skulls from the Lhasa district took the role of the fake bank notes.

These simulations all concern the MI between an individual and a mean. However, the MI between two means is also of importance, so with the last two datasets we examined the MI that underlies Hotelling's two-sample T^2 test. From each dataset, two population MVN distributions were constructed that had different means, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, but the same covariance matrix, $\boldsymbol{\Sigma}$. For the psychological test dataset, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ were set equal to the sample means for men and women, and $\boldsymbol{\Sigma}$ was equated to their pooled sample covariance matrix. Two sample groups, each of size n, were generated from the population distributions and resamples of size n were generated by sampling with replacement from each group separately. The MI between the means of the resamples for men and women was calculated using the pooled covariance matrix of the resamples as $\widehat{\boldsymbol{\Sigma}}$. Estimates of the W_j were evaluated using the Garthwaite-Koch partition and confidence inter-

vals were constructed in the same way as with the bank notes data. Sample sizes of $n = 20, 50, 80, 100$ and 200 were examined. This simulation procedure was also followed with the Flea beetles dataset; $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ were set equal to the sample means of *Haltica Oleracea* and *Haltica Carduorum*, respectively, and $\boldsymbol{\Sigma}$ to their pooled sample covariance matrix.

3.5.4 Results: Multivariate Normal distribution

Table 3.2 illustrates the output that was obtained for a single simulation. It is from the simulations of contributions of individual variables for the bank note data and gives results for the studentized pivotal method for the first fake bank note with samples of size 20. Various 95% confidence intervals were constructed, with different nominal coverages in each tail that add to 5%. Each row of the table gives a different confidence interval, with the nominal coverages in each tail shown in the first column. The actual coverages of confidence intervals are given in the columns headed CCI (Coverage of Confidence Interval), while the coverages in tails are given in CLT (left-tail) and CRT (right-tail). The table also gives the median width of intervals for each combination of tail probabilities. The shortest 95% confidence interval was identified for each variable in each of the 1000 samples and results for these shortest intervals are recorded in a separate row. The bank note dataset has six variables and results are presented separately for each of these.

The coverages of the confidence intervals is a little low for the variable,'left', and a little high with some intervals for the variable 'diagonal', but otherwise the coverages are reasonably close to the target coverage of 95%. For this fake bank note, the shortest interval was always close to the interval that had 0.0% and 5.0%

Table 3.2: Coverages (%) of confidence intervals formed by the studentized pivotal method for the contributions of individual variables to the MI of the first fake bank note with samples of size 20. Coverages of tail areas and median widths of confidence intervals are also given.

	First three variables											
	Length				Left				Right			
Tails[a]	CCI[b]	CLT[c]	CRT[d]	Width[e]	CCI[b]	CLT[c]	CRT[d]	Width[e]	CCI[b]	CLT[c]	CRT[d]	Width[e]
(0.0,5.0)	98.3	0.0	1.7	20.8	91.1	0.0	8.9	24.9	94.7	0.0	5.3	26.0
(0.5,4.5)	98.2	0.4	1.4	22.1	91.8	0.1	8.1	28.5	94.7	0.4	4.9	28.1
(1.0,4.0)	98.1	0.6	1.3	23.7	91.8	0.4	7.8	32.2	94.6	0.7	4.7	31.5
(1.5,3.5)	97.6	1.3	1.1	25.6	91.8	1.0	7.2	37.7	94.2	1.5	4.3	36.0
(2.0,3.0)	97.5	1.5	1.0	28.0	92.2	1.1	6.7	44.4	93.6	2.5	3.9	41.9
(2.5,2.5)	97.0	2.0	1.0	31.2	92.5	1.6	5.9	56.3	93.3	3.1	3.6	52.9
(3.0,2.0)	96.2	2.8	1.0	35.6	92.8	1.8	5.4	78.8	93.5	3.4	3.1	71.2
(3.5,1.5)	96.5	3.1	0.4	43.0	92.8	2.3	4.9	120.3	93.6	3.7	2.7	97.9
(4.0,1.0)	96.0	3.9	0.1	60.3	93.3	2.7	4.0	234.6	93.8	4.1	2.1	181.9
(4.5,0.5)	95.5	4.4	0.1	117.3	94.0	3.1	2.9	720.0	94.7	4.3	1.0	563.7
(5.0,0.0)	95.0	5.0	0.0	2545.8	96.8	3.2	0.0	2776.5	95.6	4.4	0.0	2419.1
Shortest	97.0	1.3	1.7	20.7	90.8	0.3	8.9	24.9	93.4	1.3	5.3	25.9
	Last three variables											
	Bottom				Top				Diagonal			
Tails[a]	CCI[b]	CLT[c]	CRT[d]	Width[e]	CCI[b]	CLT[c]	CRT[d]	Width[e]	CCI[b]	CLT[c]	CRT[d]	Width[e]
(0.0,5.0)	96.0	0.0	4.0	32.8	96.9	0.0	3.1	31.0	99.0	0.0	1.0	26.7
(0.5,4.5)	96.6	0.1	3.3	31.7	97.1	0.1	2.8	30.2	98.8	0.3	0.9	26.4
(1.0,4.0)	96.7	0.3	3.0	31.6	97.2	0.3	2.5	30.6	98.9	0.6	0.5	26.8
(1.5,3.5)	97.0	0.5	2.5	31.7	97.0	0.8	2.2	31.0	98.5	1.1	0.4	27.3
(2.0,3.0)	96.5	1.5	2.0	32.4	96.8	1.3	1.9	31.3	97.6	2.0	0.4	28.1
(2.5,2.5)	96.1	2.1	1.8	33.5	96.7	1.7	1.6	32.5	97.4	2.2	0.4	29.6
(3.0,2.0)	96.1	2.6	1.3	35.1	97.1	1.9	1.0	34.1	97.0	2.7	0.3	31.2
(3.5,1.5)	96.7	2.6	0.7	36.6	96.6	2.6	0.8	36.0	96.4	3.4	0.2	33.3
(4.0,1.0)	96.6	3.0	0.4	39.6	96.3	3.3	0.4	39.8	96.0	4.0	0.0	37.1
(4.5,0.5)	96.0	3.9	0.1	45.9	96.0	4.0	0.0	46.4	95.8	4.2	0.0	46.0
(5.0,0.0)	95.8	4.2	0.0	6137.0	95.6	4.4	0.0	5531.1	95.4	4.6	0.0	2901.6
Shortest	94.3	2.0	3.7	30.7	96.0	1.1	2.9	29.3	97.1	1.9	1.0	25.7

[a] The nominal coverage in (left-tail, right-tail).

[b] Coverage of confidence interval

[c] Coverage of left-tail

[d] Coverage of right-tail

[e] Median width of confidence intervals

nominal coverages in the left and right tails, respectively, and far different from the equal-tailed interval. This asymmetry in the nominal tail coverages was also found with the other fake bank notes that were examined and marked asymmetry

was also found with most variables in each of the other datasets. Greater nominal coverage in the upper tail thus seems a trait of the shortest confidence interval for individual contributions to an MI. Comparison of the equal-tailed and shortest confidence intervals reveals substantial variation in their relative lengths, with the shortest confidence interval only a little shorter than the equal-tailed interval for some variables (such as 'bottom') and much larger for others — the equal-tailed interval is more than twice the width of the shortest interval for the variables 'left' and 'right'.

Condensed results for all datasets and each sample size are presented in Tables 3.3 and 3.4. They give the average coverage across variables for the equal-tailed and shortest 95% confidence intervals. For the bank note data, each average is based on 60 separate coverages, as the six variables and ten fake bank notes gave 60 Mahalanobis distances. Averages are based on 90, 50, 4 and 4 coverages for the Athletes, Skulls, Psychological tests and Flea beetles data, respectively. (With the last two datasets we examined a single difference between two means rather than the Mahalanobis distances for ten items.) The tables also give a comparison of the width of intervals relative to the width of intervals given by Method A. Specifically, for each Mahalanobis distance, the median width of the equal-tailed and shortest 95% confidence intervals were calculated for each bootstrap method and divided by the median width of the corresponding intervals given by Method A. Averages of these ratios are presented in brackets in the tables. (Hence, for example, each average that is given for the Athletes data is based on 90 ratios.) Table 3.3 gives average coverages and average width ratios for the contributions of variables and Table 3.4 gives averages for the proportion of the MI attributed

to each variable.

To meet the definition of a confidence interval, a method must be conservative rather than liberal, so the coverage of its confidence intervals should preferably be above the nominal value of 95%, rather than below it. In Tables 3.3 and 3.4, average coverages that achieve at least the nominal level are marked blue. It is readily seen that Methods A and B (the new methods) almost always achieve their nominal coverage, while the other methods do not. The obvious question is whether the new methods achieve their higher coverage at the expense of giving wider confidence intervals. Looking at Table 3.3, the bias-corrected percentile method does typically give narrower intervals than Method A; its width ratios (in brackets) are almost always less than or equal to 1.00. However, its average coverages are too far below 95% for the method to be preferred to alternatives. The other methods typically give wider intervals than method A — much wider in the cases of the pivotal methods, but also slightly wider with both the percentile method and Method B. Hence, Method A clearly has better results than other methods in Table 3.3, with Method B a close second.

While the pattern of average coverages in Table 3.4 is similar to that in Table 3.3, its pattern of width ratios is a little different. In Table 3.4 the pivotal methods still typically give much wider intervals than the new methods and the bias corrected method (which gives poor coverage) still has the narrowest intervals, but now the percentile method gives slightly narrower intervals than the new methods, and differences between the widths of the new methods favour Method A less consistently. Nevertheless, taking both coverage and interval width into account, Method A is again the best method, with Method B a very close second.

Table 3.3: Average coverage (%) of 95% confidence intervals for individual contributions given by the percentile, bias-corrected percentile, non-studentized and studentized pivotal methods, and Methods A and B. Average of the ratio of the median widths of intervals relative to the median widths of intervals given by Method A are shown in brackets. Population distributions are multivariate normal.

Sample Size	Dataset	Percentile method	Bias-corre. Percentile	Non-student. pivotal	Studentized pivotal	Method A	Method B
Equal Tailed Intervals							
20	Bank notes	80.9 (2.01)	90.7 (0.84)	87.9 (3.24)	92.7 (1.63)	97.7 (1.00)	94.4 (0.68)
20	Athletes	89.1 (1.69)	91.8 (0.76)	86.2 (5.24)	95.4 (3.28)	99.3 (1.00)	99.2 (1.97)
20	Skulls	85.4 (1.70)	91.0 (0.94)	87.3 (4.62)	91.6 (2.28)	97.7 (1.00)	94.1 (0.80)
20	Psychol. Tests	91.7 (1.24)	91.2 (0.98)	82.9 (7.49)	92.7 (4.80)	97.2 (1.00)	96.5 (1.06)
20	Flea Beetles	88.7 (1.30)	94.6 (1.08)	94.9 (1.66)	96.1 (1.05)	97.0 (1.00)	95.9 (1.05)
50	Bank notes	88.3 (1.23)	91.5 (0.98)	87.5 (5.86)	92.7 (3.89)	96.1 (1.00)	95.1 (1.02)
50	Athletes	92.1 (1.27)	94.0 (0.99)	88.8 (6.43)	95.6 (3.70)	98.6 (1.00)	97.8 (0.99)
50	Skulls	90.8 (1.18)	91.8 (1.00)	87.6 (5.55)	93.9 (3.77)	96.7 (1.00)	96.2 (1.03)
50	Psychol. Tests	94.3 (1.09)	92.1 (0.97)	85.6 (6.47)	93.5 (4.63)	96.2 (1.00)	96.7 (1.07)
50	Flea Beetles	92.6 (1.09)	94.9 (1.02)	95.2 (1.16)	93.5 (1.00)	95.0 (1.00)	95.7 (1.06)
80	Bank notes	91.0 (1.13)	92.2 (0.98)	88.8 (5.42)	94.3 (3.69)	95.9 (1.00)	96.2 (1.04)
80	Athletes	93.3 (1.15)	93.8 (1.00)	89.6 (5.51)	95.2 (3.19)	97.3 (1.00)	97.0 (1.01)
80	Skulls	92.0 (1.11)	92.1 (1.00)	88.3 (5.19)	94.3 (3.50)	95.8 (1.00)	96.4 (1.06)
80	Psychol. Tests	94.5 (1.05)	92.7 (0.98)	87.9 (6.51)	93.7 (5.14)	96.0 (1.00)	96.6 (1.05)
80	Flea Beetles	93.3 (1.06)	94.9 (1.01)	95.2 (1.09)	95.1 (1.03)	94.8 (1.00)	95.9 (1.06)
100	Bank notes	91.5 (1.10)	92.1 (0.99)	88.9 (5.22)	93.9 (3.68)	95.7 (1.00)	95.7 (1.05)
100	Athletes	93.5 (1.12)	94.1 (1.00)	89.8 (4.96)	95.1 (2.96)	96.9 (1.00)	96.6 (1.02)
100	Skulls	92.9 (1.09)	92.8 (1.00)	89.1 (4.89)	94.7 (3.31)	96.0 (1.00)	96.6 (1.05)
100	Psychol. Tests	95.2 (1.04)	92.0 (0.97)	88.7 (6.00)	94.3 (5.45)	96.0 (1.00)	96.7 (1.07)
100	Flea Beetles	93.5 (1.04)	94.2 (1.01)	94.8 (1.07)	95.3 (1.04)	94.7 (1.00)	95.9 (1.06)
200	Bank notes	93.6 (1.05)	93.5 (0.99)	90.7 (4.35)	94.7 (3.00)	95.9 (1.00)	95.8 (1.06)
200	Athletes	94.0 (1.05)	94.1 (1.00)	90.7 (3.66)	94.9 (2.19)	95.5 (1.00)	96.1 (1.04)
200	Skulls	94.0 (1.04)	93.5 (0.99)	90.8 (3.63)	94.9 (2.48)	95.5 (1.00)	96.3 (1.06)
200	Psychol. Tests	95.2 (1.02)	91.9 (0.99)	89.7 (6.19)	94.2 (4.67)	95.9 (1.00)	96.8 (1.07)
200	Flea Beetles	94.6 (1.02)	95.0 (1.00)	95.5 (1.03)	96.1 (1.05)	94.7 (1.00)	96.3 (1.06)
Shortest Intervals							
20	Bank notes	92.0 (1.91)	89.4 (0.81)	94.3 (1.80)	92.8 (1.16)	98.9 (1.00)	96.1 (0.72)
20	Athletes	96.8 (1.68)	89.8 (0.75)	92.4 (2.31)	93.4 (1.86)	99.7 (1.00)	99.9 (1.80)
20	Skulls	93.8 (1.64)	89.6 (0.91)	93.8 (2.33)	92.3 (1.49)	98.3 (1.00)	95.3 (0.83)
20	Psychol. Tests	95.3 (1.21)	90.3 (0.94)	92.3 (3.18)	92.3 (2.20)	96.7 (1.00)	96.8 (1.04)
20	Flea Beetles	91.4 (1.27)	94.0 (1.05)	97.4 (1.38)	96.2 (1.01)	94.5 (1.00)	94.0 (1.04)
50	Bank notes	93.3 (1.20)	91.0 (0.94)	93.0 (2.67)	92.5 (1.94)	96.6 (1.00)	96.2 (1.01)
50	Athletes	95.5 (1.27)	92.7 (0.97)	93.9 (2.93)	95.1 (1.90)	98.0 (1.00)	97.4 (0.99)
50	Skulls	93.9 (1.16)	91.4 (0.97)	93.3 (2.61)	93.7 (1.90)	96.1 (1.00)	96.1 (1.02)
50	Psychol. Tests	95.2 (1.08)	91.3 (0.95)	92.8 (3.07)	92.7 (2.22)	95.7 (1.00)	96.2 (1.06)
50	Flea Beetles	93.5 (1.07)	94.2 (1.01)	96.6 (1.07)	92.2 (1.00)	93.8 (1.00)	94.2 (1.05)
80	Bank notes	93.8 (1.12)	91.7 (0.96)	93.3 (2.54)	94.0 (1.89)	96.2 (1.00)	96.6 (1.02)
80	Athletes	95.1 (1.15)	92.7 (0.99)	93.9 (2.59)	94.6 (1.72)	96.5 (1.00)	96.4 (1.00)
80	Skulls	93.9 (1.09)	91.5 (0.98)	93.3 (2.46)	93.9 (1.80)	95.3 (1.00)	96.2 (1.04)
80	Psychol. Tests	95.2 (1.04)	91.9 (0.96)	93.4 (2.94)	93.6 (2.25)	95.9 (1.00)	96.2 (1.04)
80	Flea Beetles	93.9 (1.04)	94.1 (1.00)	95.3 (1.03)	94.3 (1.02)	94.1 (1.00)	95.4 (1.05)
100	Bank notes	93.5 (1.09)	91.8 (0.97)	93.2 (2.48)	93.9 (1.87)	95.7 (1.00)	96.0 (1.04)
100	Athletes	94.9 (1.11)	93.2 (0.99)	94.1 (2.40)	94.3 (1.63)	96.2 (1.00)	96.1 (1.01)
100	Skulls	94.2 (1.07)	92.3 (0.98)	93.7 (2.35)	94.1 (1.74)	95.5 (1.00)	96.3 (1.04)
100	Psychol. Tests	95.1 (1.04)	90.1 (0.97)	92.9 (2.87)	93.5 (2.41)	95.3 (1.00)	96.1 (1.07)
100	Flea Beetles	93.3 (1.03)	93.9 (1.00)	95.2 (1.03)	94.5 (1.04)	93.9 (1.00)	95.3 (1.06)
200	Bank notes	94.6 (1.04)	93.2 (0.98)	93.9 (2.14)	94.6 (1.63)	95.8 (1.00)	95.9 (1.05)
200	Athletes	94.4 (1.05)	93.4 (1.00)	94.2 (1.93)	94.5 (1.38)	94.8 (1.00)	95.3 (1.03)
200	Skulls	94.4 (1.03)	93.0 (0.98)	94.0 (1.93)	94.7 (1.51)	95.1 (1.00)	95.9 (1.05)
200	Psychol. Tests	95.5 (1.02)	90.9 (0.99)	92.8 (2.76)	92.3 (2.11)	95.5 (1.00)	96.4 (1.06)
200	Flea Beetles	94.4 (1.02)	94.1 (1.00)	94.9 (1.01)	95.6 (1.05)	94.1 (1.00)	95.8 (1.06)

Table 3.4: Average coverage (%) of 95% confidence intervals for percentage of contributions given by the percentile, bias-corrected percentile, non-studentized and studentized pivotal methods, and Methods A and B. Average of the ratio of the median widths of intervals relative to the median widths of intervals given by Method A are shown in brackets. Population distributions are multivariate normal.

Sample Size	Dataset	Percentile method	Bias-corre. Percentile	Non-student. pivotal	Studentized pivotal	Method A	Method B
Equal Tailed Intervals							
20	Bank notes	92.5 (0.54)	88.4 (0.50)	81.3 (1.45)	91.3 (1.22)	97.6 (1.00)	94.8 (0.65)
20	Athletes	94.2 (0.64)	84.7 (0.62)	80.1 (2.40)	93.9 (2.38)	97.2 (1.00)	97.4 (1.15)
20	Skulls	91.4 (0.60)	86.5 (0.57)	82.1 (1.44)	90.7 (1.26)	96.1 (1.00)	92.8 (0.71)
20	Psychol. Tests	93.7 (0.76)	88.1 (0.71)	82.2 (1.98)	91.9 (1.80)	97.5 (1.00)	97.0 (0.93)
20	Flea Beetles	93.3 (0.77)	92.3 (0.78)	94.5 (1.06)	94.9 (0.84)	98.2 (1.00)	96.5 (0.93)
50	Bank notes	93.4 (0.77)	89.8 (0.73)	85.4 (3.06)	92.2 (2.40)	97.0 (1.00)	95.7 (0.92)
50	Athletes	94.7 (0.74)	90.7 (0.74)	86.6 (3.25)	94.7 (2.39)	98.1 (1.00)	97.2 (0.93)
50	Skulls	93.2 (0.80)	89.9 (0.77)	86.0 (2.49)	93.1 (2.09)	96.6 (1.00)	95.8 (0.94)
50	Psychol. Tests	94.1 (0.91)	90.4 (0.85)	85.5 (3.05)	92.3 (2.51)	96.3 (1.00)	96.7 (1.03)
50	Flea Beetles	94.0 (0.91)	93.6 (0.92)	95.4 (1.03)	92.7 (0.94)	95.7 (1.00)	96.0 (1.02)
80	Bank notes	93.7 (0.86)	90.8 (0.84)	87.2 (3.59)	93.7 (2.69)	96.6 (1.00)	96.4 (0.99)
80	Athletes	94.5 (0.84)	91.5 (0.84)	88.2 (3.48)	94.4 (2.34)	97.0 (1.00)	96.6 (0.96)
80	Skulls	93.3 (0.88)	90.6 (0.86)	87.5 (2.95)	93.6 (2.36)	95.8 (1.00)	96.2 (1.01)
80	Psychol. Tests	95.0 (0.94)	91.6 (0.90)	88.1 (3.63)	93.4 (3.02)	96.1 (1.00)	96.5 (1.03)
80	Flea Beetles	94.5 (0.94)	94.2 (0.95)	95.3 (1.01)	94.2 (0.99)	95.8 (1.00)	95.9 (1.04)
100	Bank notes	93.6 (0.90)	90.9 (0.87)	87.6 (3.71)	93.6 (2.84)	96.0 (1.00)	96.1 (1.02)
100	Athletes	94.6 (0.87)	92.2 (0.87)	88.7 (3.41)	94.3 (2.27)	96.7 (1.00)	96.3 (0.98)
100	Skulls	93.9 (0.91)	91.6 (0.89)	88.4 (3.03)	93.9 (2.40)	95.9 (1.00)	96.3 (1.02)
100	Psychol. Tests	95.2 (0.95)	91.4 (0.91)	88.8 (3.70)	93.5 (3.49)	96.3 (1.00)	96.4 (1.04)
100	Flea Beetles	93.9 (0.96)	93.5 (0.96)	94.6 (1.01)	95.4 (1.00)	94.5 (1.00)	96.7 (1.03)
200	Bank notes	95.0 (0.95)	93.1 (0.93)	90.2 (3.60)	94.5 (2.62)	96.2 (1.00)	96.1 (1.04)
200	Athletes	94.2 (0.94)	92.8 (0.94)	90.3 (2.96)	94.5 (1.97)	95.3 (1.00)	95.9 (1.02)
200	Skulls	94.5 (0.95)	93.0 (0.93)	90.6 (2.90)	94.6 (2.13)	95.5 (1.00)	96.1 (1.04)
200	Psychol. Tests	95.6 (0.97)	91.9 (0.96)	90.0 (4.40)	94.1 (3.60)	96.0 (1.00)	96.8 (1.05)
200	Flea Beetles	94.1 (0.98)	94.0 (0.98)	94.8 (1.00)	95.9 (1.03)	94.5 (1.00)	96.2 (1.04)
Shortest Intervals							
20	Bank notes	93.5 (0.59)	87.0 (0.54)	89.6 (1.26)	91.4 (1.04)	98.4 (1.00)	95.6 (0.66)
20	Athletes	93.5 (0.67)	82.7 (0.67)	86.5 (2.11)	92.4 (2.14)	96.0 (1.00)	95.9 (1.14)
20	Skulls	91.4 (0.65)	85.0 (0.62)	88.3 (1.34)	90.5 (1.14)	97.4 (1.00)	93.4 (0.73)
20	Psychol. Tests	93.3 (0.80)	86.5 (0.74)	89.4 (1.78)	90.7 (1.49)	96.9 (1.00)	96.4 (0.94)
20	Flea Beetles	90.6 (0.80)	89.9 (0.81)	95.1 (1.03)	93.4 (0.86)	96.3 (1.00)	94.3 (0.95)
50	Bank notes	94.0 (0.80)	89.2 (0.76)	91.2 (2.06)	91.6 (1.59)	97.9 (1.00)	96.4 (0.93)
50	Athletes	93.9 (0.79)	89.4 (0.79)	92.0 (2.31)	94.3 (1.67)	97.7 (1.00)	96.7 (0.94)
50	Skulls	93.0 (0.83)	88.9 (0.80)	91.5 (1.88)	92.4 (1.52)	96.5 (1.00)	95.7 (0.95)
50	Psychol. Tests	93.8 (0.92)	89.4 (0.86)	91.6 (2.24)	91.0 (1.83)	95.7 (1.00)	95.7 (1.03)
50	Flea Beetles	92.4 (0.91)	92.5 (0.92)	95.5 (0.98)	91.2 (0.94)	94.7 (1.00)	94.6 (1.02)
80	Bank notes	93.9 (0.88)	90.2 (0.85)	91.8 (2.18)	93.4 (1.68)	97.0 (1.00)	96.7 (0.99)
80	Athletes	93.5 (0.86)	90.4 (0.86)	92.8 (2.24)	93.9 (1.56)	96.5 (1.00)	95.9 (0.97)
80	Skulls	93.1 (0.89)	89.7 (0.87)	92.0 (2.00)	92.8 (1.57)	95.5 (1.00)	95.9 (1.01)
80	Psychol. Tests	94.9 (0.95)	90.6 (0.91)	92.7 (2.43)	92.9 (1.95)	95.9 (1.00)	95.9 (1.02)
80	Flea Beetles	93.8 (0.95)	93.6 (0.95)	95.3 (0.98)	93.2 (0.99)	95.0 (1.00)	95.4 (1.04)
100	Bank notes	93.6 (0.91)	90.2 (0.88)	92.0 (2.20)	93.5 (1.72)	95.9 (1.00)	96.2 (1.01)
100	Athletes	93.9 (0.89)	91.3 (0.89)	93.2 (2.14)	93.7 (1.51)	96.1 (1.00)	95.7 (0.98)
100	Skulls	93.5 (0.92)	90.8 (0.89)	92.6 (1.98)	93.3 (1.56)	95.6 (1.00)	96.1 (1.01)
100	Psychol. Tests	94.9 (0.96)	90.0 (0.92)	92.4 (2.44)	92.6 (2.07)	96.0 (1.00)	95.9 (1.04)
100	Flea Beetles	93.2 (0.95)	92.8 (0.96)	94.5 (0.99)	94.7 (1.00)	94.0 (1.00)	96.0 (1.03)
200	Bank notes	94.8 (0.96)	92.5 (0.94)	93.5 (2.01)	94.4 (1.57)	96.0 (1.00)	96.0 (1.04)
200	Athletes	93.6 (0.94)	92.2 (0.94)	93.7 (1.83)	94.3 (1.34)	94.8 (1.00)	95.2 (1.02)
200	Skulls	94.0 (0.96)	92.2 (0.94)	93.6 (1.79)	94.1 (1.44)	95.0 (1.00)	95.8 (1.03)
200	Psychol. Tests	95.1 (0.98)	90.6 (0.96)	92.7 (2.48)	92.6 (2.00)	95.7 (1.00)	96.4 (1.05)
200	Flea Beetles	93.5 (0.98)	93.5 (0.98)	94.0 (0.99)	95.5 (1.03)	94.0 (1.00)	95.6 (1.05)

Table 3.5: Average coverage (%) of 95% equal-tailed and shortest confidence intervals for Method A and Method B, for sample sizes of 500 and 1000.

Dataset	Equal tailed intervals				Shortest intervals			
	Method A		Method B		Method A		Method B	
	500	1000	500	1000	500	1000	500	1000
Contribution of individual variables								
Bank notes	95.5	95.3	96.3	96.2	95.2	94.9	96.0	95.9
Athletes	95.6	95.2	96.5	96.1	95.1	94.7	95.8	95.6
Skulls	95.5	95.3	96.4	96.1	95.0	94.9	96.0	95.8
Psychol. Tests	95.5	95.3	96.4	96.5	95.4	94.1	96.1	95.2
Flea Beetles	95.5	94.6	96.4	95.9	94.9	94.2	95.9	95.7
Percentage contribution of variables								
Bank notes	95.7	95.0	96.0	95.9	95.4	94.6	96.3	95.6
Athletes	95.7	95.3	96.4	96.2	95.0	94.7	95.8	95.7
Skills	95.7	95.2	96.5	96.1	95.2	94.8	96.1	95.8
Psychol. Tests	95.2	95.1	96.2	96.3	94.9	93.9	95.8	95.1
Flea Beetles	95.7	95.1	96.4	95.9	95.3	94.4	96.1	95.5

Theoretical results about asymptotic properties have not been derived for the new methods (A and B) of forming confidence intervals. To explore their behaviour for larger sample sizes, further simulations were conducted with these methods that were identical to those reported above, but for sample sizes of 500 and 1000. Results are presented in Table 3.5. It can be seen that the coverage is always close to the nominal level of 95%, especially for the larger sample size of 1000. These results suggest that coverage will tend to 95% as sample size increases. The coverages for Method A are generally closer to 95% than those of Method B, albeit by marginal amounts.

As Method A seems the best method, we examined more closely how it might be used. Specifically, we compared the shortest confidence intervals that it gave with the equal-tailed intervals it gave. For each Mahalanobis distance, the width of the equal-tailed 95% interval was divided by the width of the shortest 95% con-

Table 3.6: Relative frequency distribution (%) for the width ratio of confidence intervals given by equal-tailed interval relative to shortest interval using Method A for shortest intervals that are not one-sided

Sample size	Dataset	width ratio						
		1.0-1.05	1.05-1.1	1.1-1.25	1.25-1.5	1.5-1.75	1.75-2.0	2 and over
Contribution of individual variables								
	Bank notes	97.62	2.38	0.00	0.00	0.00	0.00	0.00
	Athletes	87.80	4.88	0.00	7.32	0.00	0.00	0.00
20	Skulls	97.84	2.08	0.09	0.00	0.00	0.00	0.00
	Psychol. Tests	95.21	4.50	0.29	0.00	0.00	0.00	0.00
	Flea beetles	93.06	6.53	0.41	0.00	0.00	0.00	0.00
	Bank notes	96.61	3.22	0.17	0.00	0.00	0.00	0.00
	Athletes	88.86	10.55	0.59	0.00	0.00	0.00	0.00
100	Skulls	94.36	5.41	0.22	0.00	0.00	0.00	0.00
	Psychol. Tests	91.52	8.12	0.35	0.00	0.00	0.00	0.00
	Flea beetles	99.10	0.90	0.00	0.00	0.00	0.00	0.00
Percentage contribution of variables								
	Bank notes	86.64	11.82	1.54	0.00	0.00	0.00	0.00
	Athletes	93.32	6.11	0.57	0.00	0.00	0.00	0.00
20	Skulls	54.81	22.13	21.31	1.75	0.00	0.00	0.00
	Psychol. Tests	71.77	18.02	9.67	0.54	0.00	0.00	0.00
	Flea beetles	86.14	12.21	1.60	0.05	0.00	0.00	0.00
	Bank notes	93.66	5.80	0.54	0.00	0.00	0.00	0.00
	Athletes	81.94	16.33	1.73	0.00	0.00	0.00	0.00
100	Skulls	89.15	9.22	1.60	0.03	0.00	0.00	0.00
	Psychol. Tests	93.24	6.52	0.23	0.00	0.00	0.00	0.00
	Flea beetles	98.68	1.33	0.00	0.00	0.00	0.00	0.00

fidence intervals. The frequency distributions of these ratios differed substantially depending upon whether or not the shortest confidence interval was a one-sided interval (with 0 as its lower endpoint). Table 3.6 gives the relative frequency distributions when the shortest interval *is not* one-sided. It can be seen the equal-tailed interval is generally only slightly longer than the shortest interval, especially when the contribution of each variable (rather than percentage contribution) is the quantity of interests, when the equal-tailed interval is seldom more than 5% longer

Table 3.7: Relative frequency distribution (%) for the width ratio of confidence intervals given by equal-tailed interval relative to shortest interval using Method A for shortest intervals that are one-sided

Sample	Dataset	width ratio						
size		1.0-1.05	1.05-1.1	1.1-1.25	1.25-1.5	1.5-1.75	1.75-2.0	2 and over
Contribution of individual variables								
	Bank notes	2.69	12.94	30.48	31.53	16.88	4.36	1.13
	Athletes	0.41	3.54	22.76	51.00	18.66	2.92	0.70
20	Skulls	4.42	11.47	36.39	35.55	10.33	1.53	0.31
	Psychol. Tests	4.53	11.62	46.94	35.93	0.97	0.00	0.00
	Flea beetles	13.39	32.70	50.87	3.04	0.00	0.00	0.00
	Bank notes	1.86	9.38	58.25	30.33	0.17	0.00	0.00
	Athletes	3.95	15.33	58.67	21.89	0.16	0.00	0.00
100	Skulls	2.66	10.85	57.27	29.01	0.21	0.00	0.00
	Psychol. Tests	1.86	9.71	54.75	33.61	0.07	0.00	0.00
	Flea beetles*	-	-	-	-	-	-	-
Percentage contribution of variables								
	Bank notes	7.59	13.96	35.59	35.72	6.64	0.48	0.03
	Athletes	4.49	7.34	44.83	40.48	2.74	0.10	0.01
20	Skulls	6.40	13.35	40.93	33.07	5.63	0.55	0.07
	Psychol. Tests	3.66	8.87	41.82	41.89	3.73	0.04	0.00
	Flea beetles	9.97	21.95	56.15	11.67	0.26	0.00	0.00
	Bank notes	1.51	7.62	53.30	36.84	0.74	0.00	0.00
	Athletes	2.87	12.24	57.31	27.06	0.51	0.01	0.00
100	Skulls	2.02	8.81	53.55	34.95	0.68	0.00	0.00
	Psychol. Tests	2.67	8.70	51.58	36.42	0.63	0.00	0.00
	Flea beetles*	-	-	-	-	-	-	-

* Shortest confidence intervals for the flea beetles data were never one-sided for this sample size

than the shortest interval. As people are unfamiliar with interpreting confidence intervals that are neither equal-tailed nor one-sided, there will seldom be much justification for presenting a shortest confidence interval if it is not one-sided.

Table 3.7 gives the relative frequency distributions when the shortest interval *is* one-sided. Now the equal-tailed interval is 10%-50% wider than the one-sided interval in most cases. We earlier noted that a one-sided confidence interval for a squared quantity is attractive when 0 is contained in an equal-tailed interval for the un-squared quantity, which typically happens when the shortest interval is one-sided. Hence, when the shortest confidence interval is one-sided, there seems

good reason to report it in preference to an equal-tailed interval.

Tables 3.6 and 3.7 only present results for sample sizes of 20 and 100. Results for sample sizes of 50, 80 and 200 were also produced but, for brevity, are not presented here because results did not vary appreciably with sample size.

3.6 Simulation Study: Skew Distribution

In the last section we compared transformation and examined their performance when the underlying population distributions were multivariate normal. Here we extend that work and examine the sensitivity of its results to departures from normal distributions. Specifically, we take the same population distributions as before and use the sinh-arcsinh transformation (Jones and Pewsey, 2009) to construct skew population distributions that retain features of the original distributions — each variable keeps its mean and variance and the correlation structure is broadly similar. Details are given in the next subsections.

3.6.1 Simulation procedure for skew distributions: bank note data

We will refer to the bank note data constructed here as the skew-genuine and skew-fake bank notes, to distinguish them from the genuine and fake bank notes from which we start. As before, we calculated the mean and covariance of the 100 genuine bank note measurements and took these as the mean ($\boldsymbol{\mu}$) and variance ($\boldsymbol{\Sigma}$) of an MVN distribution. We then generated 100,000 observations from an MVN($\mathbf{0}, \boldsymbol{\Sigma}$) distribution. Denote these observations as $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{100\,000}$ and put

$\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})^\top$. The sinh-arcsinh was then applied separately to each component of each $\boldsymbol{y}$, putting

$$y_{ij}^{\#} = \sinh\left[\delta^{-1}\left\{\sinh^{-1}(y_{ij}) + \epsilon\right\}\right] \tag{3.24}$$

for $i = 1, \ldots, 100\,000$; $j = 1, \ldots, p$.

Let $Y_j^{\#}$ denote a scalar variable whose sample values are $y_{1j}^{\#}, \ldots, y_{100\,000j}^{\#}$. The parameters ϵ and δ in equation (3.24) respectively affect the skewness and tailweight of the distribution of $Y_j^{\#}$. The parameter δ is always positive and ϵ has a range of $(-\infty, +\infty)$. The positive value of ϵ yields positive skewness and negative value yields negative skewness and skewness increases with increasing ϵ. Whereas tailweight increases with decreasing δ, $\delta < 1$ yields tailweight that is heavier than the Normal distribution. For the bank note data we set $\epsilon = 1.0$ and $\delta = 0.8$. Let $\bar{y}_j^{\#}$ and $s(y_j^{\#})$ denote the sample mean and sample standard deviation of $Y_j^{\#}$. Also let μ_j denote the jth component of $\boldsymbol{\mu}$ and σ_j^2 denote the jth diagonal element of $\boldsymbol{\Sigma}$. For $i = 1, \ldots, 100\,000$; $j = 1, \ldots, p$ put

$$x_{ij}^{\#} = \mu_j + \sigma_j \left\{y_{ij}^{\#} - \bar{y}_j^{\#}\right\} / s\left(y_j^{\#}\right) \tag{3.25}$$

and $\boldsymbol{x}_i^{\#} = \left(x_{i1}^{\#}, \ldots, x_{ip}^{\#}\right)^\top$.

We suppose that the complete population of skew-genuine bank notes consists of 100 000 notes and that $\boldsymbol{x}_i^{\#}$ is the vector of measurements on the ith note ($i = 1, \ldots, 100\,000$). There are only two differences between the simulation method used now and the simulation method used in Section 3.5.

1. In Section 3.5, sample datasets were generated from MVN$(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Here, $\boldsymbol{x}_1^{\#}, \ldots, \boldsymbol{x}_{100\,000}^{\#}$ are treated as the population and a sample data set is generated by sampling without replacement from $\boldsymbol{x}_1^{\#}, \ldots, \boldsymbol{x}_{100\,000}^{\#}$.

2. The distribution of the skew-fake bank notes should be similar in shape to that of the skew-genuine bank notes, but with a different mean. Let $\boldsymbol{\mu}^{\diamond} = \left(\mu_1^{\diamond}, \ldots, \mu_p^{\diamond}\right)^{\top}$ denote the sample mean of the 100 fake bank notes and let $\boldsymbol{t} = (t_1, \ldots, t_p)^{\top}$ denote the deviations from this mean for one fake bank note. Analogous to equations (3.24) and (3.25), for $j = 1, \ldots, p$ put

$$t_j^{\#} = \sinh\left[\delta^{-1}\left\{\sinh^{-1}(t_j) + \epsilon\right\}\right] \tag{3.26}$$

and

$$v_j^{\#} = \mu_j^{\diamond} + \sigma_j \left\{t_j^{\#} - \bar{y}_j^{\#}\right\} / s\left(y_j^{\#}\right) \tag{3.27}$$

where σ_j, $\bar{y}_j^{\#}$ and $s(y_j^{\#})$ take the same values as in equation (3.25). We take $(v_1^{\#}, \ldots, v_p^{\#})^{\top}$ as the vector of values of the skew-fake bank note. In the simulations, a skew-fake bank note is constructed from each of the first ten fake bank notes.

For the population of skew-genuine bank notes, the Pearson's moment coefficients of skewness for the six measurements were 1.016, 1.020, 0.981, 1.384, 1.410 and 1.150. Figures 3.2(a) and 3.2(b) show the marginal probability density functions (*p.d.f.*s) of the third and the fifth measurements. It can be seen that the *p.d.f.*s have clear skewness.

3.6.2 Simulation procedure for skew distributions: other datasets

The simulation procedure used to construct skew distributions for the bank note dataset was also used for the Athletes data and the Tibetan skull data. As in the earlier study, the male athletes or the skulls from the Sikkim area took the role of

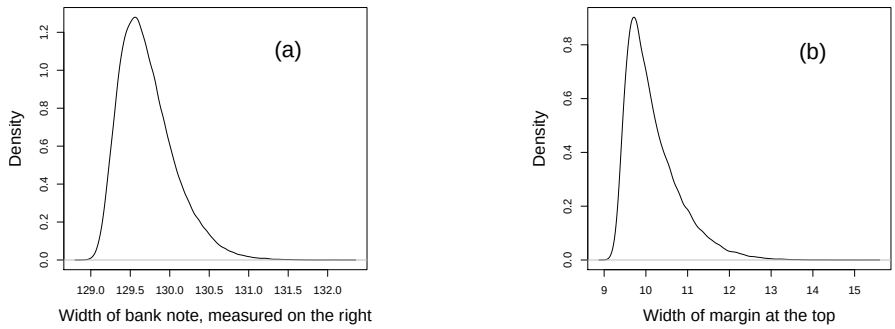

Figure 3.2: Probability density functions with coefficients of skewness of (a) 0.981 and (b) 1.410.

the genuine bank notes, while the first ten female athletes or the first ten skulls from the Lhasa district took the role of the fake bank notes.

For the Psychological test and Flea beetle datasets, in which two means are compared, $\mathbf{\Sigma}$ was set equal to the pooled sample covariance matrix and two populations of 100 000 skew-data were constructed. For the Psychological test, for one population μ_j in equation (3.25) was set equal to the jth component of the sample mean for men and for the other population it was obtained from the sample mean for women. For the Tibetan skulls, μ_j was taken as the jth component of either the sample mean for skulls from the Sikkim area (for one population) or the sample mean for skulls from the Lhasa district (for the other population).

In applying the sinh-arcsinh transformation, the parameters ϵ and δ were varied across our five datasets so as to vary the degree of skewness and thickness of tails. Table 3.8 shows the values chosen for ϵ and δ and lists the Pearson's moment coefficients of skewness and kurtosis for each variable in the datasets. (The Pearson's moment coefficient of kurtosis is 3 for a normal distribution and larger for heavier tails).

Table 3.8: Parameters ϵ and δ and the Pearson's moment coefficient of skewness for each variable in each dataset. The Pearson's moment coefficient of kurtosis for each variable is given in parentheses

Dataset	ϵ	δ	Skewness (kurtosis)
Bank notes	1.0	0.8	1.02(4.44), 1.02(4.56), 0.98(4.37), 1.38(5.41), 1.41(5.62), 1.15(4.84)
Athletes	0.8	0.9	1.57(5.50), 1.57(5.55), 0.73(3.76), 1.24(4.82), 1.48(5.30), 1.44(5.19), 1.56(5.36), 1.52(5.43), 1.58(5.53)
Skulls	0.5	0.9	1.24(4.67), 1.24(4.71), 1.22(4.65), 1.22(4.69), 1.24(4.73)
Psychol. Tests	0.4	0.8	1.29(5.53), 1.33(5.46), 1.35(5.61), 1.33(5.54)
Flea beetles	0.6	0.9	1.40(5.11), 1.39(5.03), 1.38(5.04), 1.39(5.01)

3.6.3 Results: Skew distributions

Tables 3.9 and 3.10 give results for the skew population distributions that are equivalent to results presented in Tables 3.3 and 3.4 for the MVN population distributions. Table 3.9 gives average coverages of intervals for the contributions of individual variables and hence corresponds to Table 3.3; Table 3.10 gives similar results for the proportion of the MI attributable to each variable and so corresponds to Table 3.4. As before, the median widths of intervals given by each method were compared with the median widths of intervals given by Method A. The average of these ratios are given in brackets in the tables. Average coverages above the nominal level of 95% are again marked blue.

The motivation for these simulations was to examine the robustness of results to departures from normality in the population distributions. Hence we focus on comparing Tables 3.9 and 3.10 with Tables 3.3 and 3.4. Regarding average coverage, the same main features in Tables 3.3 and 3.4 were also found in Tables 3.9 and 3.10. Specifically:

- Methods A and B almost always achieve the nominal coverage.

Table 3.9: Average coverage (%) of 95% confidence intervals for individual contributions given by the percentile, bias-corrected percentile, non-studentized and studentized pivotal methods, and Methods A and B. Average of the ratio of the median widths of intervals relative to the median widths of intervals given by Method A are shown in brackets. Population distributions are skew.

Sample Size	Dataset	Percentile method	Bias-corre. Percentile	Non-student. pivotal	Studentized pivotal	Method A	Method B
Equal Tailed Intervals							
20	Bank notes	78.0 (2.11)	89.4 (0.80)	85.6 (2.87)	90.2 (1.38)	96.8 (1.00)	92.8 (0.63)
20	Athletes	85.9 (1.76)	90.5 (0.73)	84.6 (4.38)	95.3 (2.76)	97.7 (1.00)	97.7 (2.16)
20	Skulls	84.1 (1.90)	89.1 (0.90)	84.1 (3.92)	90.7 (1.93)	98.6 (1.00)	94.9 (0.80)
20	Psychol. Tests	88.8 (1.34)	89.7 (0.99)	78.5 (7.24)	91.6 (4.66)	96.3 (1.00)	95.6 (1.10)
20	Flea Beetles	84.8 (1.47)	92.9 (1.15)	92.7 (1.81)	95.3 (1.14)	95.9 (1.00)	95.6 (1.09)
50	Bank notes	85.6 (1.30)	89.6 (0.99)	84.4 (5.56)	91.7 (3.55)	95.9 (1.00)	94.5 (1.04)
50	Athletes	89.0 (1.37)	93.0 (1.03)	88.0 (5.43)	94.7 (2.85)	97.2 (1.00)	96.1 (1.03)
50	Skulls	89.2 (1.30)	89.9 (1.04)	85.7 (5.81)	93.2 (3.66)	96.5 (1.00)	95.9 (1.11)
50	Psychol. Tests	92.6 (1.13)	91.2 (0.96)	82.6 (7.00)	93.4 (4.86)	95.6 (1.00)	96.6 (1.10)
50	Flea Beetles	89.7 (1.14)	93.4 (1.04)	93.0 (1.20)	93.9 (1.01)	94.6 (1.00)	95.4 (1.08)
80	Bank notes	89.2 (1.18)	90.9 (0.99)	86.0 (5.51)	93.0 (3.68)	95.4 (1.00)	95.4 (1.08)
80	Athletes	90.1 (1.21)	93.2 (1.02)	88.8 (4.17)	94.5 (2.49)	95.9 (1.00)	95.3 (1.04)
80	Skulls	91.0 (1.18)	90.9 (1.03)	87.1 (5.42)	93.7 (3.54)	95.9 (1.00)	96.1 (1.11)
80	Psychol. Tests	93.9 (1.07)	92.0 (0.96)	85.8 (6.31)	93.1 (4.82)	95.6 (1.00)	96.4 (1.10)
80	Flea Beetles	91.9 (1.08)	94.4 (1.02)	93.8 (1.11)	94.6 (1.04)	94.8 (1.00)	95.9 (1.07)
100	Bank notes	90.3 (1.14)	91.4 (0.99)	86.9 (5.17)	93.6 (3.59)	95.5 (1.00)	95.8 (1.09)
100	Athletes	90.6 (1.16)	93.3 (1.01)	89.3 (3.78)	94.6 (2.36)	95.4 (1.00)	95.2 (1.05)
100	Skulls	92.3 (1.14)	91.9 (1.02)	88.7 (5.15)	94.5 (3.59)	95.9 (1.00)	96.5 (1.11)
100	Psychol. Tests	94.3 (1.05)	92.6 (0.95)	87.5 (5.63)	94.1 (4.84)	95.3 (1.00)	96.7 (1.09)
100	Flea Beetles	93.1 (1.06)	94.5 (1.02)	94.4 (1.09)	95.2 (1.05)	94.8 (1.00)	96.0 (1.08)
200	Bank notes	92.2 (1.07)	92.1 (0.99)	88.7 (4.37)	94.2 (3.07)	94.9 (1.00)	95.9 (1.09)
200	Athletes	92.1 (1.07)	93.5 (1.00)	90.8 (2.93)	94.3 (2.17)	94.3 (1.00)	95.0 (1.06)
200	Skulls	93.9 (1.07)	92.7 (1.01)	90.5 (4.42)	94.8 (3.05)	95.6 (1.00)	96.5 (1.09)
200	Psychol. Tests	95.3 (1.03)	91.6 (0.96)	88.0 (5.71)	93.7 (4.58)	95.3 (1.00)	96.6 (1.08)
200	Flea Beetles	94.2 (1.03)	95.3 (1.01)	95.0 (1.04)	96.1 (1.05)	95.7 (1.00)	96.5 (1.07)
Shortest Intervals							
20	Bank notes	90.4 (2.04)	88.3 (0.79)	93.7 (1.63)	91.3 (1.04)	99.2 (1.00)	95.5 (0.68)
20	Athletes	96.4 (1.75)	89.5 (0.73)	92.3 (2.02)	93.6 (1.66)	99.8 (1.00)	99.9 (1.97)
20	Skulls	93.2 (1.82)	87.8 (0.87)	91.6 (2.04)	91.0 (1.33)	98.6 (1.00)	96.2 (0.83)
20	Psychol. Tests	93.3 (1.31)	89.5 (0.95)	91.8 (3.16)	92.5 (2.27)	96.1 (1.00)	96.3 (1.07)
20	Flea Beetles	89.5 (1.43)	92.2 (1.10)	95.6 (1.49)	95.4 (1.10)	94.1 (1.00)	94.2 (1.06)
50	Bank notes	90.9 (1.27)	89.6 (0.94)	91.6 (2.56)	92.0 (1.87)	95.8 (1.00)	95.3 (1.02)
50	Athletes	94.3 (1.36)	92.7 (1.00)	93.9 (2.62)	95.3 (1.71)	97.6 (1.00)	97.0 (1.02)
50	Skulls	92.9 (1.26)	89.8 (1.00)	91.9 (2.75)	92.9 (1.96)	95.5 (1.00)	95.7 (1.08)
50	Psychol. Tests	93.8 (1.10)	90.5 (0.93)	91.3 (2.96)	92.1 (2.29)	94.9 (1.00)	96.2 (1.08)
50	Flea Beetles	91.6 (1.11)	93.0 (1.02)	95.1 (1.09)	92.8 (1.00)	93.1 (1.00)	94.2 (1.06)
80	Bank notes	92.6 (1.16)	90.9 (0.96)	92.0 (2.55)	93.0 (1.91)	95.3 (1.00)	95.7 (1.06)
80	Athletes	93.8 (1.19)	93.1 (0.99)	94.2 (2.21)	94.9 (1.52)	95.8 (1.00)	95.7 (1.02)
80	Skulls	93.3 (1.15)	90.8 (1.00)	92.5 (2.61)	93.4 (1.88)	95.0 (1.00)	95.8 (1.08)
80	Psychol. Tests	94.6 (1.07)	91.3 (0.95)	92.9 (2.92)	92.9 (2.25)	95.1 (1.00)	96.0 (1.09)
80	Flea Beetles	92.7 (1.06)	93.7 (1.01)	95.0 (1.04)	93.6 (1.03)	93.8 (1.00)	94.7 (1.07)
100	Bank notes	93.1 (1.12)	91.4 (0.97)	92.4 (2.46)	93.6 (1.88)	95.3 (1.00)	95.9 (1.07)
100	Athletes	93.6 (1.14)	93.1 (0.99)	94.5 (2.05)	94.8 (1.46)	95.4 (1.00)	95.5 (1.03)
100	Skulls	94.1 (1.12)	91.4 (0.99)	93.1 (2.49)	93.9 (1.87)	95.1 (1.00)	96.2 (1.09)
100	Psychol. Tests	94.9 (1.05)	91.4 (0.93)	93.4 (2.66)	93.5 (2.13)	94.7 (1.00)	95.9 (1.08)
100	Flea Beetles	93.5 (1.05)	94.1 (1.01)	95.0 (1.03)	94.7 (1.04)	93.7 (1.00)	95.3 (1.07)
200	Bank notes	93.3 (1.06)	91.8 (0.98)	93.0 (2.20)	94.0 (1.70)	94.7 (1.00)	95.7 (1.07)
200	Athletes	93.8 (1.06)	93.3 (0.99)	94.7 (1.70)	94.0 (1.36)	94.6 (1.00)	95.3 (1.05)
200	Skulls	94.4 (1.06)	92.1 (0.99)	93.5 (2.22)	94.2 (1.68)	94.8 (1.00)	96.1 (1.08)
200	Psychol. Tests	95.2 (1.02)	90.2 (0.95)	92.3 (2.69)	92.1 (2.09)	95.1 (1.00)	96.1 (1.07)
200	Flea Beetles	94.5 (1.02)	95.0 (1.00)	95.5 (1.01)	95.9 (1.05)	95.1 (1.00)	96.2 (1.06)

Table 3.10: Average coverage (%) of 95% confidence intervals for percentage of contributions given by the percentile, bias-corrected percentile, non-studentized and studentized pivotal methods, and Methods A and B. Average of the ratio of the median widths of intervals relative to the median widths of intervals given by Method A are shown in brackets. Population distributions are skew.

Sample Size	Dataset	Percentile method	Bias-corre. Percentile	Non-student. pivotal	Studentized pivotal	Method A	Method B
Equal Tailed Intervals							
20	Bank notes	92.6 (0.58)	86.5 (0.53)	79.0 (1.48)	88.8 (1.26)	96.6 (1.00)	93.2 (0.68)
20	Athletes	93.2 (0.68)	83.2 (0.68)	79.8 (2.45)	93.0 (2.48)	96.5 (1.00)	96.9 (1.14)
20	Skulls	90.8 (0.63)	83.2 (0.58)	79.4 (1.42)	89.2 (1.30)	96.0 (1.00)	92.9 (0.81)
20	Psychol. Tests	93.3 (0.70)	86.6 (0.64)	79.2 (1.79)	90.2 (1.67)	96.5 (1.00)	95.8 (0.95)
20	Flea Beetles	92.9 (0.74)	91.4 (0.75)	92.7 (1.03)	93.9 (0.87)	97.3 (1.00)	96.4 (0.94)
50	Bank notes	92.9 (0.72)	88.2 (0.69)	83.0 (2.72)	90.8 (2.14)	96.2 (1.00)	95.1 (0.93)
50	Athletes	93.5 (0.70)	89.7 (0.72)	86.1 (2.73)	93.7 (1.93)	96.8 (1.00)	95.9 (0.93)
50	Skulls	92.2 (0.74)	88.0 (0.71)	84.9 (2.17)	92.3 (1.80)	96.2 (1.00)	95.3 (0.95)
50	Psychol. Tests	94.4 (0.86)	89.6 (0.80)	83.0 (2.98)	91.7 (2.47)	95.4 (1.00)	96.6 (1.06)
50	Flea Beetles	92.4 (0.88)	91.6 (0.89)	93.6 (1.01)	92.3 (0.92)	94.8 (1.00)	95.5 (1.04)
80	Bank notes	93.5 (0.82)	89.6 (0.79)	85.2 (3.28)	92.3 (2.53)	95.8 (1.00)	95.6 (1.02)
80	Athletes	92.9 (0.78)	90.6 (0.79)	87.6 (2.63)	93.5 (1.80)	95.6 (1.00)	95.1 (0.96)
80	Skulls	93.2 (0.83)	89.8 (0.80)	87.0 (2.62)	92.8 (2.10)	95.9 (1.00)	95.8 (1.01)
80	Psychol. Tests	94.7 (0.91)	91.0 (0.86)	86.3 (3.42)	92.6 (2.88)	95.6 (1.00)	96.7 (1.07)
80	Flea Beetles	94.2 (0.92)	93.9 (0.92)	94.9 (1.00)	94.1 (0.97)	95.3 (1.00)	96.5 (1.05)
100	Bank notes	94.0 (0.85)	90.6 (0.83)	86.2 (3.41)	93.0 (2.62)	95.7 (1.00)	96.0 (1.04)
100	Athletes	93.2 (0.82)	91.3 (0.82)	88.3 (2.60)	93.8 (1.79)	95.3 (1.00)	95.1 (0.99)
100	Skulls	93.8 (0.86)	90.6 (0.84)	88.2 (2.76)	93.5 (2.23)	95.9 (1.00)	96.1 (1.03)
100	Psychol. Tests	95.2 (0.93)	91.6 (0.86)	87.4 (3.47)	93.8 (3.06)	95.4 (1.00)	96.6 (1.07)
100	Flea Beetles	94.2 (0.94)	93.8 (0.94)	94.8 (1.00)	94.6 (0.99)	95.1 (1.00)	96.2 (1.06)
200	Bank notes	94.0 (0.93)	91.4 (0.91)	88.3 (3.49)	93.9 (2.60)	95.0 (1.00)	96.0 (1.07)
200	Athletes	93.1 (0.90)	92.3 (0.90)	90.3 (2.37)	93.7 (1.82)	94.4 (1.00)	95.0 (1.03)
200	Skulls	94.6 (0.93)	92.1 (0.91)	90.3 (3.08)	94.4 (2.37)	95.7 (1.00)	96.5 (1.05)
200	Psychol. Tests	95.9 (0.96)	91.3 (0.92)	88.2 (4.13)	93.7 (3.55)	96.1 (1.00)	96.9 (1.07)
200	Flea Beetles	95.0 (0.97)	94.8 (0.97)	94.9 (1.00)	95.6 (1.03)	95.4 (1.00)	96.2 (1.06)
Shortest Intervals							
20	Bank notes	93.2 (0.62)	84.9 (0.57)	87.2 (1.29)	89.1 (1.08)	97.6 (1.00)	93.5 (0.69)
20	Athletes	92.9 (0.72)	81.2 (0.73)	85.0 (2.18)	91.6 (2.29)	96.5 (1.00)	96.9 (1.12)
20	Skulls	90.2 (0.66)	80.9 (0.61)	84.2 (1.33)	88.1 (1.19)	95.0 (1.00)	92.1 (0.80)
20	Psychol. Tests	93.0 (0.75)	85.2 (0.68)	88.0 (1.62)	89.7 (1.43)	95.6 (1.00)	95.1 (0.96)
20	Flea Beetles	89.9 (0.78)	88.7 (0.79)	93.5 (1.01)	92.5 (0.89)	94.8 (1.00)	94.1 (0.95)
50	Bank notes	92.9 (0.76)	87.3 (0.73)	89.3 (1.91)	90.4 (1.51)	96.2 (1.00)	95.3 (0.94)
50	Athletes	92.8 (0.75)	88.8 (0.77)	92.0 (2.06)	94.2 (1.51)	97.0 (1.00)	96.0 (0.94)
50	Skulls	91.5 (0.78)	86.3 (0.75)	89.8 (1.76)	91.2 (1.45)	95.6 (1.00)	94.9 (0.97)
50	Psychol. Tests	93.7 (0.88)	87.9 (0.81)	90.8 (2.13)	90.4 (1.80)	95.0 (1.00)	95.9 (1.05)
50	Flea Beetles	91.2 (0.88)	90.3 (0.89)	93.7 (0.97)	90.7 (0.92)	93.5 (1.00)	94.2 (1.03)
80	Bank notes	93.4 (0.84)	89.0 (0.81)	90.4 (2.09)	92.1 (1.64)	95.8 (1.00)	95.8 (1.02)
80	Athletes	92.3 (0.82)	89.8 (0.83)	92.8 (1.88)	93.8 (1.37)	95.7 (1.00)	95.1 (0.97)
80	Skulls	92.3 (0.85)	88.4 (0.83)	91.2 (1.92)	92.1 (1.54)	95.2 (1.00)	95.3 (1.02)
80	Psychol. Tests	94.2 (0.93)	89.5 (0.87)	92.3 (2.36)	91.8 (1.90)	95.3 (1.00)	96.1 (1.07)
80	Flea Beetles	93.4 (0.92)	92.9 (0.93)	95.0 (0.97)	93.3 (0.97)	94.3 (1.00)	95.5 (1.05)
100	Bank notes	93.8 (0.87)	89.7 (0.84)	91.3 (2.10)	92.8 (1.67)	95.7 (1.00)	96.1 (1.04)
100	Athletes	92.7 (0.84)	90.6 (0.85)	93.4 (1.80)	94.1 (1.33)	95.6 (1.00)	95.3 (0.99)
100	Skulls	93.1 (0.88)	89.2 (0.85)	91.8 (1.94)	92.6 (1.58)	95.3 (1.00)	95.6 (1.03)
100	Psychol. Tests	94.8 (0.94)	90.2 (0.87)	92.8 (2.20)	92.7 (1.85)	95.1 (1.00)	96.1 (1.06)
100	Flea Beetles	93.3 (0.94)	92.9 (0.94)	94.9 (0.98)	94.0 (0.99)	94.5 (1.00)	95.6 (1.06)
200	Bank notes	93.6 (0.94)	90.8 (0.92)	92.2 (2.05)	93.6 (1.61)	94.7 (1.00)	95.7 (1.06)
200	Athletes	93.3 (0.91)	91.8 (0.91)	94.1 (1.61)	93.6 (1.30)	94.8 (1.00)	95.3 (1.03)
200	Skulls	94.0 (0.94)	91.1 (0.92)	93.0 (1.96)	93.5 (1.56)	95.1 (1.00)	96.0 (1.05)
200	Psychol. Tests	95.2 (0.97)	90.2 (0.92)	92.1 (2.40)	92.0 (1.95)	95.6 (1.00)	96.7 (1.07)
200	Flea Beetles	94.6 (0.97)	94.3 (0.97)	94.7 (0.99)	95.2 (1.03)	94.9 (1.00)	95.9 (1.06)

- With other methods, the average coverage is typically below the nominal level.

- In particular, the average coverage of the bias-corrected percentile method is often well below the nominal level.

Similarly, with respect to the width of intervals, the results in Tables 3.9 and 3.10 show the same basic patterns as in Tables 3.3 and 3.4:

- In Table 3.9 (as in Table 3.3), the bias-corrected percentile method typically gives narrower confidence intervals than Method A (at the expense of having poor coverage). The pivotal methods generally give much wider interval than Method A and the percentile method and Method B give slightly wider intervals than Method A.

- In Table 3.10 (as in Table 3.4), the pivotal methods still typically give much wider intervals than the new methods, the percentile method gives slightly narrower intervals than the new methods, and differences in widths of the new methods favour Method A. Again, the bias corrected method has the narrowest intervals, but with poor coverage.

The only noteworthy difference between results with the skew distributions and those with the MVN distributions is that average coverages were slightly smaller with the skew distributions. With the new methods, averages coverages were still almost always above the nominal level, so average coverages for these methods were better with the skew distributions than with the MVN distributions; the other methods gave average coverages that were usually below the nominal level for the MVN distributions, so these were further below the nominal level for the

skew distributions.

With the MVN distributions, the overall conclusion was that Method A had better results than other methods, with Method B a close second. This conclusion also holds for the skew population distributions. In general, the results found with the MVN population distributions were robust to the introduction of skewness and higher kurtosis.

3.7 Concluding comments

Motivation for this chapter is the potential importance of the Garthwaite-Koch partition (as illustrated, for example, in the work of Rogers (2015)) and the need of a method for forming confidence intervals for the quantities it yields. In the simulations the two new methods almost always gave confidence intervals whose coverage was conservative, while the coverages of the four standard methods that were examined were generally liberal. Nevertheless, the widths of the intervals given by the new methods tended to be much smaller than those given by the non-studentized and studentized pivotal methods, and similar in size to those of the percentile method. The only method that gave appreciably narrower intervals than the new methods was the bias-corrected percentile, but the coverage of its intervals was typically well below the nominal 95% level. These results held both for MVN population distributions and for skew population distributions with heavy tails, suggesting some robustness of these results to departures from normality. Consequently, in this study the new methods clearly outperformed the standard methods.

In the study, Method A performed marginally better than Method B — there

was little to choose between them in terms of their coverages but Method A tended to give slightly narrower intervals. Method A is also computationally a little simpler and a little faster than Method B (unlike Method B, it does not require second-level bootstrap sampling), so it is the method we recommend for constructing bootstrap confidence intervals for both the contributions and the percentage contributions determined by the Garthwaite-Koch partition.

The widths of equal-tailed and shortest confidence intervals given by Method A were compared. It was found the shortest interval was generally not markedly narrower than the equal-tailed interval when the shortest interval was a two-sided confidence interval, but differences tended to be much greater when the shortest interval was a one-sided confidence interval. In the present context with squared quantities, the shortest interval has further merit when it is one-sided, as it gives coherence with the interval for the un-squared quantity. Hence, reporting the shortest interval in preference to the equal-tailed interval should be strongly considered when the shortest interval is one-sided.

The good performance of the new methods begs two obvious questions: "Why did the new methods perform well for this application?" and "For what types of application are they likely to be useful?" Regarding the first question, none of the pivotal quantities used in this chapter are exactly pivotal (under the strict definition of a pivotal quantity), but the simulations indicate that the W_j are closer to giving pivotal quantities than $\log W_j^2$ or $\text{logit}(W_j^2/\sum_{i=1}^{p} W_i^2)$. Hence the new methods performed better than the non-studentized and studentized pivotal methods. Also, pivotal methods will outperform percentile methods when conditions do not hold for the latter to work well *and* the pivotal methods use good

pivotal quantities. So it seems that the W_j are reasonably good pivotal quantities for the application of interest here.

Regarding the second question, the benefit of the new methods is that they enable the use of a broader range of pivotal quantities than can be used with standard bootstrap methods. Thus, the question may be re-phrased as "When will broadening the range of pivotal quantities prove advantageous?" Davison and Hinkley (1997) note the importance of variance-stabilization in choosing the quantity (θ) to bootstrap. They write (p. 111), "Experience suggests that bootstrap methods for confidence limits and significance tests ... are most effective when θ is essentially a location parameter, which is approximately induced by a variance-stabilizing transformation." This suggests that the new methods should be considered when the quantity of interest (such as W_j^2 or $W_j^2 / \sum_{i=1}^{p} W_i^2$) can be constructed from simpler quantities that are essentially location parameters. For example, the methods should be considered if the quantity of interest can be expressed as a function of the means of several inter-related variables that have been scaled to have unit sample variances.

We believe that Method A and Method B should prove useful in practice. While Method A performed marginally better than method B in this application, this will not always be the case. The difference between the two methods is similar to the difference between the non-studentized and studentized pivotal methods: Method A uses pivotal quantities whose variances may fluctuate across samples, while Method B uses pivotal quantities that have been standardised to have consistent sample variances. This has little benefit in the present application because the variances of the pivotal quantities used by Method A do not vary appreciably

across bootstrap samples (they derive from the Mahalanobis distance — a scale invariant quantity). Further research is needed to evaluate the methods in other applications.

Chapter 4

Relative importance of variables in regression sum of squares

4.1 Introduction

An important question that statistical consultants and researchers commonly face after conducting a multiple regression analysis is which variable contributes most to predict or explain the criterion variable. For example, a chemist may raise the question of the relative importance of temperature and concentration in determining the rate of reaction. The term *importance* is recognized in the literature as having various possible meanings. A predictor may be considered important if the corresponding regression parameter is statistically significant. A second definition judges a predictor as more or less important on the basis of its practical impact on the response. Soofi et al. (2000) claimed that relative importance is related to statistical estimation. However, Kruskal and Majors (1989) took a sample from a population of papers that had relative importance (or the equivalent) in their

titles. The papers come from many different fields and Kruskal and Majors were unhappy to find that a substantial fraction (one-fifth) of them used statistical significance to measure relative importance. It has been argued that the question of relative importance is even more common than the question of statistical significance (e.g., Healy, 1990).

Numerous methods have been proposed for evaluating the relative importance of regressors. Darlington (1968) gave an overview of the methods for evaluating practical importance that were available at that time. These included using squared zero-order correlations, squared standardized regression coefficients, product measures, usefulness, and a measure proposed by Engelhart (1936). We describe these five measures in the next section. When predictors are uncorrelated, they lead to the same result and the sum over all the predictor variables for any of the five measures is equal to R^2. However, the measures can lead to different results for correlated regressors and among them only the product measure and Engelhart's measure satisfy the condition that sum of the individual contributions over all regressors is equal to model R^2. Among the five measures, only Engelhart's measure considers the contribution of individual predictors and the joint effect of each pair of predictors, but all the measures ignore the interaction effect of all possible combinations of predictor variables. This is regretable, as people are interested in the total contribution of a predictor and this is the sum of the direct contribution of that particular predictor and all possible joint contributions (interactions) with other predictors.

The order in which the regressors enter the model is important for correlated regressors. For example, the sum of the squares due to X_3 from the order $X_1X_2X_3$

and the order $X_1X_3X_2$ will be different, and hence the contribution of X_3 to R^2 will also be different. So the relative importance based on only one fixed order is not appropriate. It is better to take the average of the sequential sums of squares over all $p!$ possible ordering of the p regressors. Lindeman et al. (1980) proposed an unweighted average of sequential sums of squares over $p!$ possible orderings and hence the name LMG (Lindeman, Merenda and Gold). According to Johnson and LeBreton (2004) LMG was the first measure that was theoretically meaningful and consistently provided sensible results.

Budescu (1993) and Azen and Budescu (2003) proposed a number of measures in work referred to as "dominance analysis". The most widely used of these measure is a general dominance measure that is identical to the LMG measure. The LMG/general dominance measure is computationally intensive and Lindeman et al. (1980) claimed during the introduction of LMG that the method may not be feasible for more than 5 or 6 predictors. The general dominance measure formulates the task in a different way to the LMG measure, and is computationally much faster than the LMG method. However, neither measure can be calculated exactly when there are more than about 25 variables and then we have found that random sampling can be used with the LMG method and gives good accuracy.

Gibson (1962) and Johnson (1966) suggested using transformed orthogonal variables for measuring the relative importance of a set of predictors that are highly correlated to the original set of predictors. Green et al. (1978) felt the method had limitations and suggested estimating the relative importance of the original predictors by regressing the orthogonal predictors on the original predictors. Instead of regressing the orthogonal predictors on the original predictors, J.

W. Johnson (2000) proposed regressing the original predictors on the orthogonal predictors. However, the methods ignore the correlation between the response variable and the predictor variables in obtaining the orthogonal predictors.

Reviews of work on relative importance are given in Johnson and LeBreton (2004), Nathans et al. (2012), Grömping (2007), Grömping (2015), Bi (2012), and Kraha et al. (2012), and a good older review is given by Darlington (1968). As noted by Johnson and LeBreton (2004), there is no unique solution to the problem of evaluating relative importance, so identifying good measures must be based on the logic behind their development, their properties and shortcomings, and the apparent sensibility of the results they yield.

In this chapter we develop new measures of relative importance and compare them with well-regarded alternatives. The new measures are based on transformations that yield orthogonal variables that are closely related to the original regressors. In consequence they have much in common with the orthogonal counterparts measure proposed by Gibson (1962) and the relative weights measure of Johnson (2000). The main difference is that the new measures use the values of both the regressors *and* the response in determining the transformation, while the measures of Gibson (1962) and Johnson (2000) ignore the response when determining the transformation and use only the values of the regressors. Intuitively, there should be benefits in letting the response influence the transformation, as the purpose of the transformation is to help evaluate the relationship between regressor and the response.

The new measures proposed here are compared with the orthogonal counterparts measure (Gibson, 1962), and the relative weights measure (Johnson, 2000)

and also with the general dominance measure proposed by Budescu (1993). Comparison is made through examples and by examining theoretical properties.

In Section 4.2 we briefly discuss simple methods for measuring importance. We briefly discuss the more complicated methods for evaluating importance in Section 4.3. The variable transformation methods are described in Section 4.4. In Section 4.5 we describe the new methods and they are compared with other three measures in Section 4.7. Some of the measures have the *rotation invariance property*, whereby an orthogonal rotation can be applied to some variables without affecting the relative weights assigned to un-rotated variables. The rotation invariance property for the first new measure is proved in Section 4.6. Concluding comments are given in Section 4.8.

4.2 Simple methods of relative importance

In this section we describe the methods reviewed by Darlington (1968). These measures are the simple measures and should be used only when the regressors are uncorrelated. For correlated regressors, these measures have serious drawbacks and so the methods are not considered further after this section, but reviewed here for completeness. The methods are illustrated using a real dataset.

We assume that the response, Y, and regressors $X_1, \ldots, X_p$ are related through the regression equation

$$Y | \boldsymbol{X} = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon, \tag{4.1}$$

where $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ and ϵ is random error and has variance σ^2. We suppose

Table 4.1: Correlation matrix of the body fat data

Variable	BF	TST	TC	MC
BF	1.000	0.843	0.878	0.142
TST	0.843	1.000	0.924	0.458
TC	0.878	0.924	1.000	0.085
MC	0.142	0.458	0.085	1.000

there are n data, so that the model can be written in matrix form as:

$$\boldsymbol{y}|\,\mathbf{X} = \beta_0\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.2}$$

where $\mathbf{1}$ is an $n\times 1$ vector of 1's, $\boldsymbol{y}$ is an $n\times 1$ vector of responses, $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ is an $n \times p$ matrix of known values of $X_1, \ldots, X_p$, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients (whose values are unknown) and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of independent random errors. The coefficient β_0 is irrelevant for the regressors' relative importance so, to simplify notation, throughout this chapter we assume that Y and $X_1, \ldots, X_p$ have been centered to have sample means of 0. Then the least squares estimate of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{y}$ and $\text{var}(\widehat{\boldsymbol{\beta}}) = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$.

4.2.1 Example data

Data for illustrating the methods was obtained from Neter et al. (1983). There are four measurements collected from 20 healthy women, aged between 25 and 34 years. The measurements were: BF (Body Fat), TST (Triceps Skinfold Thickness), TC (Thigh Circumference), MC (Midarm Circumference). We take BF as the response variable. The correlation matrix of the body fat data is shown in Table 4.1. There is a high correlation between TST and TC and the correlation structure creates problems in allocating relative importance.

The fitted standardized multiple regression model is:

$$\widehat{BF} = 4.264TST - 2.929TC - 1.561MC \tag{4.3}$$

having $R^2 = 0.801$.

4.2.2 Zero-order correlation (validities)

Zero-order correlation is the simplest measure of importance. It measures the degree and direction of the linear relationship between a regressor variable and the response variable when all other regressors are ignored in the regression model, so that it is unaffected by the other regressors of the model. In the case of uncorrelated predictors, the sum of the squared zero-order correlations is equal to the model R^2, and thus can be used for initial rank ordering of the individual contributions of predictor variables to the model. However, for correlated regressors, shared variance in the response variable is added up several times and thus the sum of the squared zero-order correlations is often greater than R^2 for the model with all regressors together (Bi, 2012). This is also clear from the example data:

Variable	TST	TC	MC
$r^2_{yx_j}$	0.711	0.771	0.020

$Note$: $r^2_{yx_j}$ is the squared zero-order correlation between the response and the jth regressor.

These $r^2_{yx_j}$'s are the R^2 values of each regressor. The sum of the squared zero-order correlations is 1.502, which is about twice the size of the model R^2 of 0.801. In contrast, the reverse relation can happen if some of the regressors are suppressors (Hamilton, 1987). If a regressor has zero or near zero correlation with the response but is correlated with one or more of the regressors then that variable is a suppressor.

4.2.3 Standardized regression coefficients (beta weights)

Standardized regression coefficients are commonly used for evaluating the contribution of each predictor variable. Beta weights are easily computed and when the predictor variables are uncorrelated they are simply equal to the zero-order correlations. In such a case, squared beta weights can be used to determine the relative importance of each predictor — the squared beta weights sum to the full model's R^2 so there is no need to calculate a complicated measures in order to rank predictor variables. However, predictor variables are usually correlated and beta weight for a particular predictor will depend on which other predictor variables are in the model. When a predictor shares the explained variance with one or more predictors in the model (Pedhazur, 1997), then a predictor variable that has a high positive correlation with the response variable may have a near-zero beta weight. Alternatively, a predictor variable with a low (positive) zero-order correlation may have a large positive beta weight. As Darlington (1968) notes, it is possible to have a negative beta weight for a predictor that has a positive zero-order correlation. The following are the squared beta weights determined from the example dataset:

Variable	TST	TC	MC
$\widehat{\beta}_j^2$	18.179	8.577	2.438

Note : $\widehat{\beta}_j^2$ is the squared beta weight for the jth predictor.

TST and TC are approximately equally correlated with BF. However, TST contributes twice as much as TC in predicting BF and the beta weight ($\widehat{\beta}_j$) for TC is negative, even though TC has a positive correlation with the criterion. This happens due to the high collinearity between TST and TC. Thus interpretation

of beta weight is sometimes an invalid measure of the importance of collinear regressors. Moreover, if we add or remove variables from the model, then the sign of the beta weights can change.

4.2.4 Product measures

Hoffman (1960) proposed the product measure (named by Bring, 1996), which is the product of the zero-order correlations with corresponding beta weights. Pratt (1987) justified this measure as a relative importance measure and showed that the sum of the product measures over all predictor variables is equal to the model R^2, irrespective of correlated or uncorrelated predictors. A major disadvantage of this measure is that it may produce a negative importance value for a predictor variable even though that predictor contributes substantially to the criterion (Darlington, 1968). Thomas et al. (1998) claimed that the negative value of the product measure can only happen for high multicollinearity. Basically, this measure shares the limitations of both the zero-order correlations and the beta weights (Bring, 1996; Darlington, 1968). The example data illustrates this.

Variable	TST	TC	MC
$\widehat{\beta}_j r_{yx_j}$	3.595	-2.572	-0.222

Note : $\widehat{\beta}_j r_{yx_j}$ is the product of the beta weight for the jth predictor and the corresponding zero-order correlation.

Since the beta weights for the variables TC and MC are negative while the zero-order correlations are positive, the product measures for these variables are negative. So calculation of percentages of importance is not possible. If one or more of the predictors has a negative product measure value, then the product measures of all variables have no meaningful interpretation. Pratt (1987) mentioned that this

measure would be valid only if both the zero-order correlation and beta weight for a predictor variable have same sign.

4.2.5 Usefulness

The increase (decrease) in R^2 from adding (removing) a predictor to (from) a model that already contains all other predictors is referred to as the usefulness of that particular predictor (Darlington, 1968). If the regressors are highly correlated, the usefulness of a predictor can exceed the squared zero-order correlation and a predictor with the lowest zero-order correlation can have a higher usefulness than some of the other predictor variables. For correlated predictors, the sum of the usefulness over all predictors is typically far less than the model R^2 (Grömping, 2006). The table below gives the increase in R^2 from adding each variable for example dataset:

Variable	TST	TC	MC
Usefulness	0.026	0.015	0.023

Though MC has a lower zero-order correlation than TC, the percentage of importance assigned from the usefulness measure to the variable TC is even smaller than that of MC. MC has a usefulness value of 0.023, which is greater than the squared zero-order correlation of 0.020 (see, Subsection 4.2.2). Also, sum of the usefulness of the three predictors is 0.064, which is far less than the model R^2 of 0.801.

4.2.6 Engelhart's measure

Engelhart (1936) assigns a contribution (squared beta weight) to each predictor

variable and also a joint effect to each pair of predictor variables. The sum of the contributions (individual and joint) is equal to the model R^2 irrespective of correlated or uncorrelated predictors. Engelhart expressed model R^2 by

$$R^2 = \widehat{\beta}_1^2 + \ldots + \widehat{\beta}_p^2 + 2\widehat{\beta}_1\widehat{\beta}_2 r_{12} + \ldots + 2\widehat{\beta}_{p-1}\widehat{\beta}_p r_{(p-1)p}. \tag{4.4}$$

If a regression model has p predictors then it will produce p individual contributions and $[p(p-1)]/2$ joint effect terms. So if the number of predictors increases then the total number of joint contributions increases rapidly. For high multicollinearity, the joint effect can be negative (as with the product measure) and hence have no meaningful interpretation (Darlington, 1968). This is also clear from the example dataset:

Variable	TST	TC	MC	$TST * TC$	$TST * MC$	$TC * MC$
Contribution	18.179	8.577	2.438	-23.072	-6.095	0.774

Because of high multicollinearity between TST and TC, the joint effect of them is negative, even though both of them are highly correlated with the response variable, BF. The sum of the contributions is equal to the model R^2 of 0.801. Since some of the joint contributions are negative it is not possible to calculate the percentage contributions.

4.3 Relative importance based on sequential sums of squares

In the previous section, we discussed simple methods for evaluating relative importance of predictor variables in multiple regression. In this section, we discuss

the LMG and general dominance measures, which are based on averaging over orderings.

For correlated regressors, the change in the regression sum of squares for adding any predictor depends on the order in which the predictor enter the model. So, to determine the contribution of a predictor to the total sum of squares, it is better to take the average of the sums of squares over all possible $p!$ orderings of p predictors.

In the remainder of this chapter we will assume that Y and each X variable have been standardised to have unit length. That is, $\boldsymbol{y}^\top \boldsymbol{y} = \boldsymbol{x}_j^\top \boldsymbol{x}_j = 1$ for $j = 1, \ldots, p$. So the total sum of squares is 1. Thus regression sum of squares is equal to the model R^2.

4.3.1 LMG measure

The method was proposed by Lindeman et al. (1980, p.120), and hence has the name LMG. They proposed taking a simple average of the sequential sums of squares over all orderings. Kruskal (1987) independently decomposed R^2 into non-negative partitions based on averaging the squared semipartial correlations over orderings and popularized the LMG method. This method is computationally intensive — as the number of predictors increases the number of possible ordering also increases rapidly.

To specify a formula for the LMG measure, suppose the regression sum of squares for a model with the set S of regressors is denoted by $SSR(S)$. Also let $SSR(M|S)$ denote the additional sum of squares from adding the set M of

regressors to the set S, defined as

$$SSR(M|S) = SSR(M \cup S) - SSR(S). \tag{4.5}$$

Let $\mathcal{F}$ be the set of all $p!$ orderings of regressors and let $r = (r_1, \ldots, r_p)$ denote an ordering ($r \in \mathcal{F}$). Let (r_j) denote the position of X_j in that ordering. Suppose $S^r_{(r_j)-1}$ denotes the set of the first $(r_j) - 1$ variables that entered into the model in the order r. Then the sum of squares of X_j in the order r is given by

$$\begin{aligned} SSR(\{X_j\}|S^r_{(r_j)-1}) &= SSR(\{X_j\} \cup S^r_{(r_j)-1}) - SSR(S^r_{(r_j)-1}) \\ &= SSR(S^r_{(r_j)}) - SSR(S^r_{(r_j)-1}). \end{aligned} \tag{4.6}$$

So for p regressors, the LMG for regressor X_j is given as

$$\text{LMG}(X_j) = \frac{1}{p!} \sum_{r \in \mathcal{F}} SSR(\{X_j\}|S^r_{(r_j)-1}). \tag{4.7}$$

When Lindeman et al. proposed this relative importance measure in 1980, fitting models with all combinations of variables was only practical when the number of variables was fewer than 5 or 6. Since then, advances in computer power has substantially increased that number, so currently it takes only 0.28 seconds to fit all submodels of 12 regressors using software developed by Grömping (2006). However, it is not possible ot use Grömping's software with models containing 25 regressors.

4.3.2 Dominance Analysis (DA) measure

Dominance analysis measures relative importance and was originally proposed by Budescu (1993) and refined and extended by Azen and Budescu (2003). DA is based on the comparison of R^2 for all possible subset models. Budescu (1993) defined the contribution of variables as the squared semipartial correlation, i.e., the

increase in R^2 from adding a new variable to a model that contains other variables. For example, the contribution of X_1 to the subset model $\{X_3\}$ is $R^2_{Y.X_1X_3} - R^2_{Y.X_3}$, where $R^2_{Y.X_1X_3}$ is the value of R^2 when Y is regressed on X_1 and X_3, while $R^2_{Y.X_3}$ is the R^2- value when Y is regressed on X_3.

When performing a regression analysis, the standard summary statistics include an estimate of the regression coefficients, the standard error of the coefficients and the t- statistics, as well as the p- values corresponding to the t- statistics. The t- statistics are used to check whether individual regression coefficients significantly differ from 0 and the square of the t- statistics are F- statistics. The increase in R^2 from adding a variable to the model that has other variables in the model is used to calculate the dominance analysis measure and is also the numerator of the F-statistic for testing whether a single parameter should be added to the model.

There are three types of dominance, namely complete dominance, conditional dominance and general dominance (Azen and Budescu, 2003). The strongest type of dominance, termed complete dominance, rarely occurs. If the additional contribution of X_j is always higher than X_k for all possible subset models, then X_j completely dominates X_k. According to Budescu (1993), if X_j completely dominates X_k and X_k completely dominates X_r then X_j completely dominates X_r, i.e., dominance is transitive. Complete dominance does not exist in most cases, because typically the additional contribution of X_j is greater than that of X_k for some of the subset models, but the reverse happens for the remaining subset models. A weaker type of dominance, called conditional dominance, compares the average additional contributions to all subset models of sizes 0 to $p-1$. If the

average additional contribution of X_j is greater than X_k within each model size, then X_j conditionally dominates X_k. Like complete dominance, conditional dominance often does not exist for all pairs of predictors. Lastly, the weakest type of dominance, called general dominance, is computed by averaging all the conditional statistics. The sum of the general dominance over all predictors is equal to the model R^2. If X_j has the greater overall average additional contribution than X_k then X_j generally dominates X_k.

Table 4.2: Dominance Analysis for Three Variables

Subset model ($\boldsymbol{X}$)	$R^2_{Y.\boldsymbol{X}}$	Submodel size (k)	Additional contribution of:		
			X_1	X_2	X_3
Null	–	0	$R^2_{Y.X_1}$	$R^2_{Y.X_2}$	$R^2_{Y.X_3}$
Average (CD_0)			$R^2_{Y.X_1}$	$R^2_{Y.X_2}$	$R^2_{Y.X_3}$
X_1	$R^2_{Y.X_1}$	1	–	$R^2_{Y.X_1X_2} - R^2_{Y.X_1}$	$R^2_{Y.X_1X_3} - R^2_{Y.X_1}$
X_2	$R^2_{Y.X_2}$	1	$R^2_{Y.X_1X_2} - R^2_{Y.X_2}$	–	$R^2_{Y.X_2X_3} - R^2_{Y.X_2}$
X_3	$R^2_{Y.X_3}$	1	$R^2_{Y.X_1X_3} - R^2_{Y.X_3}$	$R^2_{Y.X_2X_3} - R^2_{Y.X_3}$	–
Average (CD_1)			$\frac{-R^2_{Y.X_2}-R^2_{Y.X_3}+R^2_{Y.X_1X_2}+R^2_{Y.X_1X_3}}{2}$	$\frac{-R^2_{Y.X_1}-R^2_{Y.X_3}+R^2_{Y.X_1X_2}+R^2_{Y.X_2X_3}}{2}$	$\frac{-R^2_{Y.X_1}-R^2_{Y.X_2}+R^2_{Y.X_1X_3}+R^2_{Y.X_2X_3}}{2}$
X_1X_2	$R^2_{Y.X_1X_2}$	2	–	–	$R^2_{Y.X_1X_2X_3} - R^2_{Y.X_1X_2}$
X_1X_3	$R^2_{Y.X_1X_3}$	2	–	$R^2_{Y.X_1X_2X_3} - R^2_{Y.X_1X_3}$	–
X_2X_3	$R^2_{Y.X_2X_3}$	2	$R^2_{Y.X_1X_2X_3} - R^2_{Y.X_2X_3}$	–	–
Average (CD_2)			$-R^2_{Y.X_2X_3} + R^2_{Y.X_1X_2X_3}$	$-R^2_{Y.X_1X_3} + R^2_{Y.X_1X_2X_3}$	$-R^2_{Y.X_1X_2} + R^2_{Y.X_1X_2X_3}$
Overall average (GD)			$GD(X_1) = \frac{\sum_{k=0}^{2} CD_k(X_1)}{3}$	$GD(X_2) = \frac{\sum_{k=0}^{2} CD_k(X_2)}{3}$	$GD(X_3) = \frac{\sum_{k=0}^{2} CD_k(X_3)}{3}$

where $GD(X_1) = \frac{2R^2_{Y.X_1}-R^2_{Y.X_2}-R^2_{Y.X_3}+R^2_{Y.X_1X_2}+R^2_{Y.X_1X_3}-2R^2_{Y.X_2X_3}+2R^2_{Y.X_1X_2X_3}}{6}$,

$GD(X_2) = \frac{-R^2_{Y.X_1}+2R^2_{Y.X_2}-R^2_{Y.X_3}+R^2_{Y.X_1X_2}-2R^2_{Y.X_1X_3}+R^2_{Y.X_2X_3}+2R^2_{Y.X_1X_2X_3}}{6}$,

$GD(X_3) = \frac{-R^2_{Y.X_1}-R^2_{Y.X_2}+2R^2_{Y.X_3}-2R^2_{Y.X_1X_2}+R^2_{Y.X_1X_3}+R^2_{Y.X_2X_3}+2R^2_{Y.X_1X_2X_3}}{6}$.

Table 4.2 gives a general picture of all three types of dominance for a three regressors multiple regression model. This table makes the dominance analysis procedure simpler to understand. The first column of Table 4.2 defines the subset model, giving the variable(s) already in the model; the second column is the R^2-value of the subset model; the third column indicates the size of the subset model, which varies between 0 and 2. The last three columns corresponding to each subset model represents the additional contributions of X_1, X_2 and X_3 respectively as a result of adding that particular predictor to a subset model. For example, $R^2_{Y.X_1X_2} - R^2_{Y.X_2}$ is the increase in R^2 when X_1 is added to the model that already contains X_2; it is called the contribution of X_1 to the subset model X_2. Similarly, $R^2_{Y.X_1X_2X_3} - R^2_{Y.X_2X_3}$ is the increase in R^2 when X_1 is added to the model that already contains X_2 and X_3. The first three coloured rows give averages and represent the average increase in R^2 from adding a specified variable to all the subset models of a given size that do not contain the specified variable. These averages are the conditional dominance values. For example, the conditional dominance value of variable X_1 to subset models of size 1 is $(-R^2_{Y.X_2} - R^2_{Y.X_3} + R^2_{Y.X_1X_2} + R^2_{Y.X_1X_3})/2$. The last coloured row is the average of the conditional dominance values for each variable and these are called the general dominance values. The expression of the general dominance values are given below the table.

General dominance coincides with the LMG measure proposed by Lindeman et al. (1980). The most commonly stated criticism of the DA measure is that it is computationally demanding. This is because there are $(2^p - 1)$ regression models that should be fitted in order to evaluate the relative importance of all variables. Azen (2003) provided a **SAS** macro for the computation of dominance

analysis that reduces the number of regression that must be calculated. The **R** package "**yhat**" of Nimon and Oswald (2013) can be used to calculate dominance analysis along with other statistics such as commonality analysis. However, with 20 regressors the general dominance measure could not be calculated using the '**yhat**' package.

The number of variables that are included in regression models has increased substantially, especially with interest in 'big-data'. One possibility is to examine a sample of submodels rather than examining all possible submodels. Simulations we have conducted suggest that the LMG measure can be well-approximated by examining 500 random sequences for entering variables into the regression model. However, it is very difficult to take a random sample from the dominance analysis procedure in order to get approximate results and reduce the number of model fitting. A procedure for taking random sequences of samples from all possible $p!$ orderings for LMG is discussed in appendix A, where we also explain why random samples cannot be obtained with the general dominance formulation.

4.4 Variables transformation methods

In variable transformation methods the correlated regressors are transformed to closely related orthogonal variables (surrogate of original regressors). Then the orthogonal variables are used as regressors instead of using the original regressors. The transformed orthogonal variables are related with the original regressors on a one-to-one basis. They are computationally far less demanding than the methods based on sequential sums of squares.

4.4.1 Orthogonal Counterparts (OC) measure

Gibson (1962) and R. M. Johnson (1966) suggested a method for obtaining a set of orthonormal predictors that are closely related on a one-to-one basis with the original set of predictors. The new predictors can be considered as 'orthogonal counterparts' to the original regressors. To approximate the relative importance of the original predictors, the response variable is regressed on the new orthonormal variables. The proportion of the predictable variance in the response that is accounted for by each orthogonal counterpart can be taken as the importance measure of the original regressors. We have discussed Johnson's (1966) transformation in Chapter 2.

The best-fitting (in the sense of least squares) orthogonal approximation of $\mathbf{X}$ (Johnson, 1966) can be obtained by

$$\mathbf{Z} = \mathbf{U}\mathbf{V}^\top, \tag{4.8}$$

where $\mathbf{U}$ and $\mathbf{V}$ are the same as in equation (2.2) and $\mathbf{Z}$ can be found in equation (2.9). The orthogonal vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ are called the 'orthogonal counterparts' of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$.

Let $\widehat{\boldsymbol{\beta}}_Z = (\widehat{\beta}_{Z_1}, \ldots, \widehat{\beta}_{Z_p})^\top$ denote the vector of regression coefficients from regressing Y on $\boldsymbol{Z}$, so

$$\widehat{\boldsymbol{\beta}}_Z = \left(\mathbf{Z}^\top\mathbf{Z}\right)^{-1}\mathbf{Z}^\top\boldsymbol{y} = \mathbf{Z}^\top\boldsymbol{y}. \tag{4.9}$$

Then $\widehat{\beta}_{Z_j}$ is called the beta weight of Z_j $(j = 1, \ldots, p)$ and the squared beta weight, $\widehat{\beta}^2_{Z_j}$, is the variation in Y that is explained by Z_j. Hence the squared beta weights are a natural measure of the relative importance of the Z variables (cf. Subsection 4.2.3). Each Z variable is paired with an X variable, and the

Orthogonal Counterparts (OC) measure takes these squared beta weights as a measure of the importance of the X variables, defining the relative importance of X_j as $\widehat{\beta}^2_{Z_j}$. The sum of these importance weights equals the variation in Y that is explained by a multiple linear regression with $X_1, \ldots, X_p$ as the independent variables (or, equivalently, with $Z_1, \ldots, Z_p$ as the independent variables).

J. W. Johnson (2000) argues that the OC measure can assign relative weights that are inappropriate when the original X variables are highly correlated, and gives examples where some variables are assigned weights that seem too low. However, the OC measure appears to give sensible weight to the X variables when the correlations between variables are not high. Also, recent work by Garthwaite and Koch (2016) implies that the OC measure has an attractive 'rotation invariance' property. (See Subsection 2.4.2 for details of rotation invariance property.) When some variables have strong collinearities, they can be transformed into non-collinear variables via orthogonal rotation of coordinate axes. Only axes corresponding to variables involved in the collinearities need to be rotated, and Garthwaite and Koch (2016) show that the rotation has no effect on the Z variables that correspond to un-rotated axes. The predictable variation in Y is also unaffected by the rotation, so the OC measure has the property that the relative importance is unchanged for those X variables associated with un-rotated axes (Further detail is given in Subsection 2.4.2). This has the following implications for the OC measure.

- Sometimes collinear variables can be transformed into meaningful variables that are not collinear through a rotation of the axes associated with them. This can lead to relative weights that are a transparently reasonable repre-

sentation of the importance of the different variates. Moreover, the relative weights are unchanged for those variables that are not involved in the rotation.

- Since axes *could* be rotated to remove collinearities without affecting the relative weights of the other variables, *multicollinearities do not affect the relative weights that the OC measure gives to variables not involved in the collinearities.*

The OC measure is compared with the new measures, the relative weight measure and the general dominance measure using numerical example in Section 4.7.

4.4.2 Green et al.'s δ^2

The OC measure of Johnson (1966) and Gibson (1962) regress Y on the orthonormal predictors $\boldsymbol{Z}$ and use squared betas as a measure of variable importance. However, they did not relate the orthonormal predictors, $\boldsymbol{Z}$, with the original predictors, $\boldsymbol{X}$. Realizing the limitations of OC measure, Green et al. (1978) proposed a measure of variable importance where they regress the columns of $\mathbf{Z}$ on $\mathbf{X}$ in addition to regressing Y on $\boldsymbol{Z}$. Suppose $\boldsymbol{\Gamma}$ is the matrix of regression coefficients when $\boldsymbol{Z}$ is regressed on $\boldsymbol{X}$. That is,

$$\boldsymbol{\Gamma} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Z}. \tag{4.10}$$

Where γ_{jk} denote the jth regression coefficient when Z_k is regressed on $\boldsymbol{X}$. Then the relative contribution of X_j to predict Z_k is obtained by

$$\gamma_{jk}^{*\,2} = \frac{\gamma_{jk}^2}{\sum_{j=1}^{p} \gamma_{jk}^2}. \tag{4.11}$$

If $\widehat{\beta}_{Z_k}$ is the beta weight (obtained from OC measure) of Z_k $(k = 1, \ldots, p)$ when regressing Y on $\boldsymbol{Z}$, then the relative importance of X_j as suggested by Green et al. (1978) is given by

$$\delta_j^2 = \sum_{k=1}^{p} {\gamma_{jk}^*}^2 \widehat{\beta}_{Z_k}^2. \tag{4.12}$$

Each δ_j^2 is always non-negative and sum of the δ_j^2's is equal to R^2.

Jackson (1980) criticized the method, because "the γ_{jk}^2's are coefficients from regressions on correlated variables; ... cannot meaningfully and unambiguously assign importance to the X_j's any more than could the $\widehat{\beta}_j$'s from a regression of Y on the X_j's". Green et al. (1980) replied that δ_j^2's are at least better than previous measures. However, this method is not considered further in this thesis because subsequent work by Johnson (2000) is clearly better, as Johnson (2000) demonstrates.

4.4.3 Relative Weights (RW) measure

The Relative Weights (RW) measure of J. W. Johnson (2000) is based on the same Z variables that are calculated for the OC measure. That is $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ are the orthogonal variables that minimize $\sum_{j=1}^{p} (\boldsymbol{x}_j - \boldsymbol{z}_j)^\top (\boldsymbol{x}_j - \boldsymbol{z}_j)$ and, as they are orthogonal, the relative importance of Z_j in predicting Y is clearly $\widehat{\beta}_{Z_j}^2$. However, while the OC measure simply takes $\widehat{\beta}_{Z_j}^2$ as the relative importance of X_j, the measure of Johnson (2000) takes into account all the correlations between the X and Z variables. From the criterion that determines the Z variables, the correlation between X_j and Z_j should be high, but this correlation could still be well below 1 if the X variables display collinearities or high intercorrelations. Also, X_j might not be the only X variable that has a marked correlation with Z_j.

Let λ_{jk} denote the correlation between X_j and Z_k. The transformation from X to Z has the unexpected symmetry property that $\lambda_{jk} = \lambda_{kj}$ for all $j,\ k$ (see, for example, Johnson (1966)). As $\mathbf{Z}$ is a linear transformation of $\mathbf{X}$ and vice-versa, $\sum_{j=1}^{p} \lambda_{jk}^2 = 1$ and the symmetry of λ_{jk} leads to the useful consequence that $\sum_{j=1}^{p} \lambda_{jk}^2 = \sum_{k=1}^{p} \lambda_{jk}^2 = 1$ (see Subsection 2.1). The RW measure divides the relative importance of Z_k amongst the X variables (as $\boldsymbol{z}_k$ is a linear combination of $\boldsymbol{x}_j$'s) according to the square of their correlations with Z_k, so the relative importance weight that X_j derives from Z_k is $\lambda_{jk}^2 \widehat{\beta}_{Z_k}^2$. (This indeed partitions the relative importance of Z_k, as $\sum_{j=1}^{p} \lambda_{jk}^2 \widehat{\beta}_{Z_k}^2 = \widehat{\beta}_{Z_j}^2 \sum_{j=1}^{p} \lambda_{jk}^2 = \widehat{\beta}_{Z_k}^2$.) The full relative importance weight of X_j is obtained by summing the relative importance weights that it derives from all the Z variables. Thus, under the RW measure, the relative importance of X_j is given by

$$\text{RW of } X_j = \sum_{k=1}^{p} \lambda_{jk}^2 \widehat{\beta}_{Z_k}^2. \tag{4.13}$$

Johnson's (2000) relative weights of regressors obtained from equation (4.13) coincides with the proposals of Fabbris (1980) and Genizi (1993) (see Nimon and Oswald (2013)). The latest version of Grömping's **R** package "**relaimpo**" contains the metric **genizi**. Also the **R** package "**yhat**" developed by Nimon and Oswald (2013) can be used to calculate RW of regressors.

The λ_{jk}^2 in equation (4.13) may be regarded as the squares of regression coefficients rather than the squares of correlation coefficients, as

$$E\left(X_j | \boldsymbol{Z}\right) = \lambda_{j1} Z_1 + \ldots + \lambda_{jp} Z_p \tag{4.14}$$

when X_j is regressed on $Z_1, \ldots, Z_p$. When Johnson (2000) proposed the RW measure he used the regression model in equation (4.14) to motivate its construction.

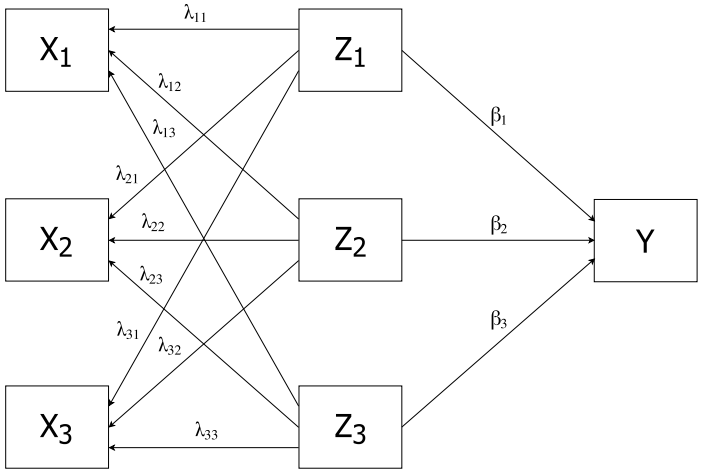

Figure 4.1: Relationships between the X, Z and Y variables for three regressors when Y is regressed on the Z variables and each X variable is regressed on the Z variables

However, we prefer to view the λ_{jk}^2 as squared correlations because correlation is a symmetric relationship while the regression equation (4.14) is a one-directional relationship and shows how the Z variables determines X_j. When viewed as a regression, the relationships between the X, Z and Y variables is illustrated in Figure 4.1 and shows no direct link between the X and Y variables. When the λ_{jk}^2 are viewed as squared correlations, the links between the X and Z variables are two-directional association, thus giving links from the X variables to Y.

Applications in which the RW measure has been used are reported in Johnson and LeBreton (2004) and Krasikova et al. (2011). Part of the attraction of the RW measure is that it typically gives similar results to the general dominance measure of Budescu (1993), even though Budescu's measure and the RW measure are calculated in very different ways. As Johnson (2000, p.15) suggests, "it is encouraging that two measures that have very different definitions and calculations produce very similar solutions", and Johnson and LeBreton, 2004 (2004, p.251)

argue that the closeness of results indicates that the two measures are measuring the same construct. Thomas et al. (2014) show that for two variables, RW of Johnson (2000) and the relative importance obtained by LMG (and hence general dominance of Budescu (1993)) are algebraically identical.

4.5 New measures of Relative Importance

Three new measures are proposed here. All are based on transformations that yield orthogonal variables — the first and third are similar to the OC measure of Gibson (1962) and R. M. Johnson (1966); the second is very similar to the RW measure of J. W. Johnson (2000). The main difference is that the new measures use transformations that are determined by cross-products of the X and Y variables, rather than ignoring Y in choosing the transformation. The third new measure uses weights to alter the balance of the different cross-products when forming orthogonal variables.

The estimated regression coefficient $\widehat{\boldsymbol{\beta}}$ for regressing Y on $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ is

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \boldsymbol{y} \tag{4.15}$$

and the regression sum of squares (RegSS) is

$$\boldsymbol{y}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}. \tag{4.16}$$

Let $(y_1, \ldots, y_n)^\top = \boldsymbol{y}$ and let $\mathbf{Y}$ be an $n \times n$ diagonal matrix with diagonal elements $y_1, \ldots, y_n$. The RegSS can also be rewritten as:

$$\mathbf{1}^\top \mathbf{Y} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \mathbf{1} \tag{4.17}$$

where $\mathbf{1}$ is a vector of ones.

Both the OC and RW measures construct orthogonal vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ that corresponds closely to the original predictors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ on a one-to-one basis. The way the $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ are chosen ignores the values of Y, even though the reason for constructing $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ is to partition the RegSS. With our new measures, a set of orthogonal vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p$ is chosen so that $\mathbf{Y}\boldsymbol{w}_j$ is closely related to $\mathbf{Y}\boldsymbol{x}_j$. Suppose Y is regressed on $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p$ and that w_{ij} is the ith component of $\boldsymbol{w}_j$. Then $\boldsymbol{w}_j$'s contribution to the RegSS from the ith sample is $(y_i w_{ij})^2$. Our new measures take $(y_i w_{ij})^2$ as a first estimate of the contribution of X_j to the RegSS from the ith sample. (The OC and RW measures equivalently take $(y_i z_{ij})^2$ as a first estimate of X_j's contribution to the RegSS from the ith sample, where z_{ij} is the ith component of $\boldsymbol{z}_j$.) Hence, as $y_i w_{ij}$ is the ith component of $\mathbf{Y}\boldsymbol{w}_j$, it is appropriate to focus on $\mathbf{Y}\boldsymbol{w}_1, \ldots, \mathbf{Y}\boldsymbol{w}_p$ in the criterion for choosing $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p$. It is for this reason that we want to focus on the distance between $\mathbf{Y}\boldsymbol{w}_j$ and $\mathbf{Y}\boldsymbol{x}_j$ (we want them to be closely related in some sense) rather than focusing on the distance between $\boldsymbol{w}_j$ and $\boldsymbol{x}_j$. We also want $\mathbf{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p)$ to be a linear transformation of $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, so that regression models with $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p$ as explanatory variables and with $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ as explanatory variables give identical predictions, residuals and regression sums of squares. Hence, analogous to equation (2.18), we choose $\mathbf{W}$ so that $\sum_{j=1}^{p} (\mathbf{Y}\boldsymbol{w}_j)^\top (\mathbf{Y}\boldsymbol{x}_j)$ is maximized subject to the constraints that $\mathbf{W}^\top\mathbf{W} = \mathbf{I}_p$ and $\mathbf{W} = \mathbf{X}\mathbf{A}$ for some $p \times p$ non-singular matrix $\mathbf{A}$.

The following lemma and theorem give the transformation for obtaining $\mathbf{W}$ from $\mathbf{X}$ and $\mathbf{Y}$.

Lemma 1. *If* $\mathbf{W} = \mathbf{X}\mathbf{A}$ *and* $\mathbf{W}^\top\mathbf{W} = \mathbf{I}_p$, *then* $\mathbf{W} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\mathbf{G}$ *where* $\mathbf{G}$ *is a* $p \times p$ *orthogonal matrix. The converse also holds, i.e., if* $\mathbf{W} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\mathbf{G}$

and $\mathbf{G}$ *is an orthogonal matrix, then* $\mathbf{W}^\top\mathbf{W} = \mathbf{I}_p$.

Proof. For the first part of the lemma, let $\mathbf{G} = (\mathbf{X}^\top\mathbf{X})^{1/2}\mathbf{A}$. Then $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1/2}\mathbf{G}$ and $\mathbf{W} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1/2}\mathbf{G}$. Also $\mathbf{I}_p = \mathbf{W}^\top\mathbf{W} = \mathbf{G}^\top(\mathbf{X}^\top\mathbf{X})^{-1/2}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1/2}\mathbf{G} = \mathbf{G}^\top\mathbf{G}$. This implies that $\mathbf{G}$ is orthogonal, as required. The converse is immediate: if $\mathbf{W} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1/2}\mathbf{G}$ and $\mathbf{G}$ is an orthogonal matrix, then $\mathbf{W}^\top\mathbf{W} = \mathbf{G}^\top(\mathbf{X}^\top\mathbf{X})^{-1/2}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1/2}\mathbf{G} = \mathbf{G}^\top\mathbf{G} = \mathbf{I}_p$. □

Theorem 4. *Under the constraints that* $\mathbf{W} = \mathbf{X}\mathbf{A}$ *and* $\mathbf{W}^\top\mathbf{W} = \mathbf{I}_p$*, the value of* $\mathbf{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p)$ *that maximizes* $\sum_{j=1}^p (\mathbf{Y}\boldsymbol{w}_j)^\top (\mathbf{Y}\boldsymbol{x}_j)$ *is*

$$\mathbf{W} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\mathbf{G}, \tag{4.18}$$

where

$$\mathbf{G} = \boldsymbol{\Psi}\left(\boldsymbol{\Psi}^\top\boldsymbol{\Psi}\right)^{-1/2} \tag{4.19}$$

and

$$\boldsymbol{\Psi} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\mathbf{X}^\top\mathbf{Y}\mathbf{Y}\mathbf{X}. \tag{4.20}$$

Proof. From Lemma 1, $\mathbf{W} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\mathbf{G}$ where $\mathbf{G}$ is an orthogonal matrix. Put $\mathbf{G} = \left(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_p\right)$, so $\boldsymbol{w}_j = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\boldsymbol{g}_j$. Also, define $\boldsymbol{\psi}_j = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}\mathbf{X}^\top\mathbf{Y}\mathbf{Y}\boldsymbol{x}_j$ for $j = 1, \ldots, p$. Then $\sum_{j=1}^p (\mathbf{Y}\boldsymbol{w}_j)^\top (\mathbf{Y}\boldsymbol{x}_j) = \sum_{j=1}^p (\boldsymbol{g}_j^\top(\mathbf{X}^\top\mathbf{X})^{-1/2}\mathbf{X}^\top\mathbf{Y}(\mathbf{Y}\boldsymbol{x}_j) = \sum_{j=1}^p \boldsymbol{g}_j^\top\boldsymbol{\psi}_j$. As $\mathbf{G}$ is an orthogonal matrix, it is immediate from Theorem 1 in Garthwaite et al. (2012) that $\sum_{j=1}^p \boldsymbol{g}_j^\top\boldsymbol{\psi}_j$ is maximized when $\mathbf{G} = \boldsymbol{\Psi}\left(\boldsymbol{\Psi}^\top\boldsymbol{\Psi}\right)^{-1/2}$, where $\boldsymbol{\Psi} = \left(\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p\right)$. Thus equation (4.20) defines $\boldsymbol{\Psi}$. □

We should note that the X variables are standardized but $\|\mathbf{Y}\boldsymbol{x}_j\|$ typically varies with j. Hence the X variables are given equal importance in maximising $\sum_{j=1}^p \boldsymbol{x}_j^\top\boldsymbol{z}_j$ or minimizing $\sum_{j=1}^p (\boldsymbol{x}_j - \boldsymbol{z}_j)^\top(\boldsymbol{x}_j - \boldsymbol{z}_j)$ (as with the OC and RW

measures) but here, in maximising $\sum_{j=1}^{p} (\mathbf{Y}\boldsymbol{w}_j)^\top (\mathbf{Y}\boldsymbol{x}_j)$, X_j is given greater importance when $\|\mathbf{Y}\boldsymbol{x}_j\|$ is larger than when it is small. This has the benefit that those X variables that are most highly correlated with Y are given greater weight when choosing the $\boldsymbol{w}_j$. (We could scale the X variables so that $\|\mathbf{Y}\boldsymbol{x}_j\|$ is the same for each X_j, but that would lose this benefit.)

4.5.1 First new measure (NM1)

In the same way that the OC measure views $\boldsymbol{z}_j$ as the counterpart of $\boldsymbol{x}_j$ ($j = 1, \ldots, p$), our first New Measure (NM1) views $\mathbf{Y}\boldsymbol{w}_j$ as the counterpart of $\mathbf{Y}\boldsymbol{x}_j$ ($j = 1, \ldots, p$). The RegSS when Y is regressed on $\boldsymbol{w}_j$ is $\{\mathbf{1}^\top \mathbf{Y}\boldsymbol{w}_j\}^2 = \left(\boldsymbol{y}^\top \boldsymbol{w}_j\right)^2$. As $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p\}$ are a set of orthonormal vectors, $\sum_{j=1}^{p} (\boldsymbol{y}^\top \boldsymbol{w}_j)^2$ is the RegSS both when Y is regressed on $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p$ and when Y is regressed on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$. NM1 defines the relative importance of X_j as

$$\text{NM1: Relative importance of } X_j = \left(\boldsymbol{y}^\top \boldsymbol{w}_j\right)^2. \tag{4.21}$$

Like the OC measure, NM1 has a rotation invariance property. Specifically, if an orthogonal rotation is applied to some of the X variables, the relative importance of the other X variables is unchanged if relative importance is measured using NM1. Further detail of rotation invariance is given in Chapter 2 and the proof of rotation invariance for NM1 is given in Section 4.6. As with the OC measure, it means that collinearities do not affect the relative importances that NM1 gives to variables not involved in the collinearities.

4.5.2 Second new measure (NM2)

While NM1 allocates all the RegSS of W_k to X_k, our second new method, NM2, divides the RegSS of W_k between the X variables according to their association with W_k. As $\mathbf{Z} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1/2}$, from equation (4.18) we have that $\mathbf{W} = \mathbf{ZG}$. (This is an attractive representation of $\mathbf{W}$ because $\mathbf{Z}$ is a set of orthonormal vectors and $\mathbf{G} = \left(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_p\right)$ is an orthogonal matrix.) Thus,

$$\boldsymbol{w}_k = \mathbf{Z}\boldsymbol{g}_k. \tag{4.22}$$

As noted in Section 4.5, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ correspond closely to $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ on a one-to-one basis, so $\boldsymbol{w}_k$ should generally be highly correlated with $\mathbf{X}\boldsymbol{g}_k$. Also $\boldsymbol{g}_j$ and $\boldsymbol{g}_k$ are orthogonal for $j \neq k$, typically $\boldsymbol{w}_k$ will not be closely associated with $\mathbf{X}\boldsymbol{g}_j$ for $j \neq k$.

NM2 divides the RegSS of W_k between $X_1, \ldots, X_p$ to reflect the squares of the sample correlations between $\mathbf{X}\boldsymbol{g}_j$ and $\boldsymbol{w}_k$ $(j = 1, \ldots, p)$. Let r_{jk} denote the sample correlation between $\mathbf{X}\boldsymbol{g}_j$ and $\boldsymbol{w}_k$. It is readily shown that

$$r_{jk} = \frac{\boldsymbol{g}_j^\top \left(\mathbf{X}^\top\mathbf{X}\right)^{1/2} \boldsymbol{g}_k}{\left[\boldsymbol{g}_j^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{g}_j\right]^{1/2}}. \tag{4.23}$$

The proportion of W_k's RegSS that NM2 attributes to X_j is $r_{jk}^2 / \sum_{l=1}^p r_{lk}^2$, so NM2 defines the relative importance of X_j as:

$$\text{NM2: Relative importance of } X_j = \sum_{k=1}^{p} \frac{r_{jk}^2 \left(\boldsymbol{y}^\top\boldsymbol{w}_k\right)^2}{\sum_{l=1}^p r_{lk}^2}. \tag{4.24}$$

If X_j has low correlations with other X variables, the NM1 and NM2 will give similar importance to X_j. However the relative importance that they assign to X_j can differ markedly if X_j is highly correlated with some of the X variables. This can be seen in the Section 4.7.

4.5.3 Third new measure (NM3)

If an X variable has a small regression coefficient in the multiple regression of y on all the X variables, then dropping that variable from the regression model can be attractive. With the NM1 and NM2 measures (and also the OC and RW measures), the orthogonal counterparts of all variables can change markedly if any X variables are discarded, which might be undesirable in some situations. Our third new measure, NM3, takes account of the size of regression coefficients when forming orthogonal counterparts, so that the inclusion or exclusion of variables with small regression coefficient has little effect on the orthogonal counterparts of other variables.

As in equation (4.15), let $\widehat{\boldsymbol{\beta}}$ denote the estimated regression coefficient for regressing Y on $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ and put $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)^\top$. While NM1 and NM2 choose $\mathbf{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p)$ to maximize $\sum_{j=1}^p (\mathbf{Y}\boldsymbol{w}_j)^\top (\mathbf{Y}\boldsymbol{x}_j)$, with NM3 we choose $\mathbf{W}^\# = (\boldsymbol{w}_1^\#, \ldots, \boldsymbol{w}_p^\#)$ to maximize $\sum_{j=1}^p |\widehat{\beta}_j| (\mathbf{Y}\boldsymbol{w}_j^\#)^\top (\mathbf{Y}\boldsymbol{x}_j)$. Thus, with NM3, the importance of the correlation between $(\mathbf{Y}\boldsymbol{w}_j^\#)$ and $(\mathbf{Y}\boldsymbol{x}_j)$ depends upon the size of $\widehat{\beta}_j$.

Now $\sum_{j=1}^p |\widehat{\beta}_j| (\mathbf{Y}\boldsymbol{w}_j^\#)^\top (\mathbf{Y}\boldsymbol{x}_j) = \sum_{j=1}^p (\mathbf{Y}\boldsymbol{w}_j^\#)^\top (\mathbf{Y}\boldsymbol{x}_j^\#)$ where $\boldsymbol{x}_j^\# = |\widehat{\beta}_j| \boldsymbol{x}_j$. If we let $\mathbf{X}^\# = (\boldsymbol{x}_1^\#, \ldots, \boldsymbol{x}_p^\#)$, then analogy to Theorem (4) yields the following result.

Corollary 1. *Under the constraints that* $\mathbf{W}^\# = \mathbf{X}^\# \mathbf{A}$ *and* $\left(\mathbf{W}^\#\right)^\top \mathbf{W}^\# = \mathbf{I}_p$, *the value of* $\mathbf{W}^\# = (\boldsymbol{w}_1^\#, \ldots, \boldsymbol{w}_p^\#)$ *that maximizes* $\sum_{j=1}^p |\widehat{\beta}_j| (\mathbf{Y}\boldsymbol{w}_j^\#)^\top (\mathbf{Y}\boldsymbol{x}_j)$ *is obtained by replacing* $\mathbf{W}$ *with* $\mathbf{W}^\#$ *and* $\mathbf{X}$ *with* $\mathbf{X}^\#$ *in equations (4.18) - (4.20).*

NM3 views $(\mathbf{Y}\boldsymbol{w}_j^\#)$ as the counterpart of $(\mathbf{Y}\boldsymbol{x}_j^\#)$ $(j = 1, \ldots, p)$ and evaluates

the relative importance of X_j as the value of R^2 when Y is regressed on $\boldsymbol{w}_j^{\#}$. Thus,

$$\text{NM3: Relative importance of } X_j = \left(\boldsymbol{y}^\top \boldsymbol{w}_j^{\#}\right)^2. \tag{4.25}$$

NM3 and the DA measure are the only measures we examine that explicitly use the multiple regression of Y on $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$. Using this regression model seems sensible, since the purpose of the measures is to evaluate the contribution of each variable to this regression.

4.6 Rotation invariance property

With the majority of measures of importance, rotating some explanatory variables will change the relative importance of *every* variable. However, results in Garthwaite and Koch (2016) show that with the OC measure only the relative importances of variables involved in the rotation are changed — the relative importances are unchanged for those variables that are not involved in the rotation (see also Subsection 2.4.2). Theorem 5 (below) shows that NM1 also has this rotation invariance property.

Lemma 2. *If* $\mathbf{H}$ *is a positive-definite matrix and* $\boldsymbol{\Gamma}$ *is an orthogonal matrix of the same dimension as* $\mathbf{H}$*, then* $\left(\boldsymbol{\Gamma}^\top \mathbf{H} \boldsymbol{\Gamma}\right)^{-1/2} = \boldsymbol{\Gamma}^\top \mathbf{H}^{-1/2} \boldsymbol{\Gamma}$.

Proof. $(\boldsymbol{\Gamma}^\top \mathbf{H}^{1/2} \boldsymbol{\Gamma}).(\boldsymbol{\Gamma}^\top \mathbf{H}^{1/2} \boldsymbol{\Gamma}) = \boldsymbol{\Gamma}^\top \mathbf{H}^{1/2} (\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top) \mathbf{H}^{1/2} \boldsymbol{\Gamma} = \boldsymbol{\Gamma}^\top \mathbf{H} \boldsymbol{\Gamma}$, so $(\boldsymbol{\Gamma}^\top \mathbf{H} \boldsymbol{\Gamma})^{1/2} = \boldsymbol{\Gamma}^\top \mathbf{H}^{1/2} \boldsymbol{\Gamma}$. Hence, $(\boldsymbol{\Gamma}^\top \mathbf{H} \boldsymbol{\Gamma})^{-1/2} = (\boldsymbol{\Gamma}^\top \mathbf{H}^{1/2} \boldsymbol{\Gamma})^{-1} = \boldsymbol{\Gamma}^{-1} \mathbf{H}^{-1/2} (\boldsymbol{\Gamma}^\top)^{-1} = \boldsymbol{\Gamma}^\top \mathbf{H}^{-1/2} \boldsymbol{\Gamma}$.

□

Lemma 3. *Suppose* $\mathbf{X}^* = \mathbf{X}\boldsymbol{\Gamma}$*. Under the constraints that* $\mathbf{W}^*$ *is a linear transformation of* $\mathbf{X}^*$ *and that* $(\mathbf{W}^*)^\top \mathbf{W}^* = \mathbf{I}_p$*, the value of* $\mathbf{W}^* = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p)^\top$ *that*

maximises $\sum_{j=1}^{p}(\mathbf{Y}\boldsymbol{w}_j^*)^\top(\mathbf{Y}\boldsymbol{x}_j^*)$ *is*

$$\mathbf{W}^* = \mathbf{W}\boldsymbol{\Gamma}, \tag{4.26}$$

where $\mathbf{W}$ *is defined by equations (4.18), (4.19) and (4.20).*

Proof. Let $\boldsymbol{\Psi}^* = [(\mathbf{X}^*)^\top\mathbf{X}^*]^{-1/2}(\mathbf{X}^*)^\top\mathbf{Y}\mathbf{Y}\mathbf{X}^*$ and put $\mathbf{G}^* = \boldsymbol{\Psi}^*[(\boldsymbol{\Psi}^*)^\top\boldsymbol{\Psi}^*]^{-1/2}$. Now, $[(\mathbf{X}^*)^\top\mathbf{X}^*]^{-1/2}(\mathbf{X}^*)^\top = [\boldsymbol{\Gamma}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\Gamma}]^{-1/2}\boldsymbol{\Gamma}^\top\mathbf{X}^\top = \boldsymbol{\Gamma}^\top[\mathbf{X}^\top\mathbf{X}]^{-1/2}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top\mathbf{X}^\top$ (from Lemma 2), so

$$\left[(\mathbf{X}^*)^\top\mathbf{X}^*\right]^{-1/2}(\mathbf{X}^*)^\top = \boldsymbol{\Gamma}^\top\left[\mathbf{X}^\top\mathbf{X}\right]^{-1/2}\mathbf{X}^\top. \tag{4.27}$$

Hence, $\boldsymbol{\Psi}^* = \boldsymbol{\Gamma}^\top[\mathbf{X}^\top\mathbf{X}]^{-1/2}\mathbf{X}^\top\mathbf{Y}\mathbf{Y}\mathbf{X}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^\top\boldsymbol{\Psi}\boldsymbol{\Gamma}$, where $\boldsymbol{\Psi}$ is defined in equation (4.20). Thus $\mathbf{G}^* = \boldsymbol{\Gamma}^\top\boldsymbol{\Psi}\boldsymbol{\Gamma}[\left(\boldsymbol{\Gamma}^\top\boldsymbol{\Psi}\boldsymbol{\Gamma}\right)^\top\left(\boldsymbol{\Gamma}^\top\boldsymbol{\Psi}\boldsymbol{\Gamma}\right)]^{-1/2} = \boldsymbol{\Gamma}^\top\boldsymbol{\Psi}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top[\boldsymbol{\Psi}^\top\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top\boldsymbol{\Psi}]^{-1/2}\boldsymbol{\Gamma}$ (from Lemma 2), so $\mathbf{G}^* = \boldsymbol{\Gamma}^\top\boldsymbol{\Psi}[\boldsymbol{\Psi}^\top\boldsymbol{\Psi}]^{-1/2}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^\top\mathbf{G}\boldsymbol{\Gamma}$, where $\mathbf{G}$ is defined by equation (4.19).

The proof of Theorem 4 does not require the fact that the X variables have been standardized to have unit variance. Hence the result of the theorem also applies to $\mathbf{W}^*$ and $\mathbf{X}^*$. It follows that $\mathbf{W}^* = (\mathbf{X}^*)^\top[(\mathbf{X}^*)^\top\mathbf{X}^*]^{-1/2}\mathbf{G}^*$ so, from equation (4.27), $\mathbf{W}^* = \mathbf{X}[\mathbf{X}^\top\mathbf{X}]^{-1/2}\boldsymbol{\Gamma}\mathbf{G}^*$. As $\mathbf{G}^* = \boldsymbol{\Gamma}^\top\mathbf{G}\boldsymbol{\Gamma}$, this gives $\mathbf{W}^* = \mathbf{X}\left[\mathbf{X}^\top\mathbf{X}\right]^{-1/2}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top\mathbf{G}\boldsymbol{\Gamma}$, so $\mathbf{W}^* = \mathbf{W}\boldsymbol{\Gamma}$, where $\mathbf{W}$ is defined in equation (4.18). □

If say, just the first d of p explanatory variables are rotated, then the rotation matrix $\boldsymbol{\Gamma}$ has the block-diagonal structure

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Gamma}_d & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-d} \end{pmatrix}, \tag{4.28}$$

where Γ_d is an orthogonal matrix of order d and $\mathbf{I}_{p-d}$ is a $(p-d)$ order identity matrix. Then, from Lemma 3, $\left(\boldsymbol{w}_{d+1}^*, \ldots, \boldsymbol{w}_p^*\right) = (\boldsymbol{w}_{d+1}, \ldots, \boldsymbol{w}_p)$. When Y is regressed on $\left(\boldsymbol{w}_1^*, \ldots, \boldsymbol{w}_p^*\right)$, the contribution of $\boldsymbol{w}_j^*$ to the RegSS is the same as the

RegSS from a simple regression of Y on $\boldsymbol{w}_j^*$, because $\left(\boldsymbol{w}_1^*, \ldots, \boldsymbol{w}_p^*\right)$ are an orthogonal set of vectors. Under NM1, this RegSS is taken as the relative importance of X_j^* in a regression of Y on $\left(X_1^*, \ldots, X_p^*\right)$. Similarly, when Y is regressed on $(X_1, \ldots, X_p)$, NM1 evaluates the relative importance of X_j as the RegSS from a simple regression of Y on $\boldsymbol{w}_j$. As $\boldsymbol{w}_j^* = \boldsymbol{w}_j$ for $j = d+1, \ldots, p$, NM1 has the rotation invariance property given in the following theorem.

Theorem 5. *If an orthogonal rotation is applied to some of the X variables, the relative importance of the other X variables is unchanged if relative importance is measured using NM1.*

4.7 Examples

In this section we apply the measures of relative importance to several datasets. In Subsection 4.7.1 we examine straightforward application of the measures, using five datasets that have clear structures. In Subsection 4.7.2 we examine how relative importance changes under orthogonal rotation of some variables and under variable selection.

4.7.1 Fixed models

Each dataset consists of 1000 data drawn from a multivariate normal distribution with a mean vector of zeros and variance-covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ varies with the dataset. The first component of a datum is the response, Y, and the other components are the explanatory variables, $X_1, \ldots, X_p$. We first describe each dataset by giving the sample correlation matrix $\widehat{\mathbf{R}}$, the multiple regression model that relates Y to the explanatory variables, the value of R^2 for that regression, and

the regression coefficients for simple regressions when Y is regressed separately on one X variable at a time. We also note salient features of the dataset. After this brief description of the datasets, we tabulate the relative importance that the different measure allocate to each variable. The results are then discussed.

Example 4.1

In this first dataset, X_2 and X_3 are useful predictors of Y and X_1 correlates with both X_2 and X_3. But the regression coefficient for X_1 is small in both a simple regression of Y on X_1 and a multiple regression of Y on X_1, X_2 and X_3.

The sample correlation matrix is

$$\widehat{\mathbf{R}} = \begin{array}{c} \begin{array}{cccc} Y & X_1 & X_2 & X_3 \end{array} \\ \left(\begin{array}{cccc} 1.000 & -0.007 & 0.501 & -0.489 \\ -0.007 & 1.000 & 0.676 & 0.717 \\ 0.501 & 0.676 & 1.000 & 0.304 \\ -0.489 & 0.717 & 0.304 & 1.000 \end{array} \right) \end{array} \begin{array}{c} Y \\ X_1 \\ X_2 \\ X_3 \end{array}$$

The fitted standardized multiple regression model is:

$$\hat{Y} = 0.063X_1 + 0.684X_2 - 0.743X_3 \qquad (R^2 = 0.706)$$

and the simple regression models are

$$\hat{Y} = -0.007X_1, \quad \hat{Y} = 0.501X_2, \quad \text{and } \hat{Y} = -0.489X_3.$$

Example 4.2

In a sense, this example is the opposite of Example 4.1. Now Y correlates highly with X_1 and its correlations with X_2 and X_3 are much lower. Also, X_1 has much

the biggest regression coefficient in a multiple regression of Y on X_1, X_2 and X_3. Again, there is marked correlation between the X variables.

The sample correlation matrix is

$$\widehat{\mathbf{R}} = \begin{array}{c} \begin{array}{cccc} Y & X_1 & X_2 & X_3 \end{array} \\ \left(\begin{array}{cccc} 1.000 & 0.847 & 0.419 & 0.382 \\ 0.847 & 1.000 & 0.701 & 0.697 \\ 0.419 & 0.701 & 1.000 & 0.483 \\ 0.382 & 0.697 & 0.483 & 1.000 \end{array} \right) \end{array} \begin{array}{c} Y \\ X_1 \\ X_2 \\ X_3 \end{array}$$

The fitted standardized multiple regression model is:

$$\hat{Y} = 1.380X_1 - 0.351X_2 - 0.411X_3 \qquad (R^2 = 0.865)$$

and the simple regression models are

$$\hat{Y} = 0.847X_1, \quad \hat{Y} = 0.419X_2, \quad \text{and } \hat{Y} = 0.382X_3.$$

Example 4.3

There are just two explanatory variables in this dataset. The Y variable is highly correlated with X_1 but uncorrelated with Z_1. Together, X_1 and X_2 give a multiple regression equation that predicts Y perfectly.

The sample correlation matrix is

$$\widehat{\mathbf{R}} = \begin{array}{c} \begin{array}{ccc} Y & X_1 & X_2 \end{array} \\ \left(\begin{array}{ccc} 1.000 & 0.893 & 0.450 \\ 0.893 & 1.000 & 0.803 \\ 0.450 & 0.803 & 1.000 \end{array} \right) \end{array} \begin{array}{c} Y \\ X_1 \\ X_2 \end{array}$$

The fitted standardized multiple regression model is:

$$\hat{Y} = 1.499X_1 - 0.755X_2 \qquad (R^2 = 1.00)$$

and the simple regression models are

$$\hat{Y} = 0.893X_1, \quad \text{and } \hat{Y} = 0.450X_2.$$

Example 4.4

The correlation between the X variables in this example is the same as in Example 4.2, apart from sampling variation. Now, however, the three X variables make similar contributions in a multiple regression, and their simple regressions (with Y as the dependent variable) are also similar.

The sample correlation matrix is

$$\widehat{\mathbf{R}} = \begin{array}{c} \begin{array}{cccc} Y & X_1 & X_2 & X_3 \end{array} \\ \left(\begin{array}{cccc} 1.000 & 0.854 & 0.801 & 0.792 \\ 0.854 & 1.000 & 0.704 & 0.679 \\ 0.801 & 0.704 & 1.000 & 0.505 \\ 0.792 & 0.679 & 0.505 & 1.000 \end{array} \right) \end{array} \begin{array}{c} Y \\ X_1 \\ X_2 \\ X_3 \end{array}$$

The fitted standardized multiple regression model is:

$$\hat{Y} = 0.335X_1 + 0.376X_2 + 0.375X_3 \qquad (R^2 = 0.884)$$

and the simple regression models are

$$\hat{Y} = 0.854X_1, \quad \hat{Y} = 0.801X_2, \quad \text{and } \hat{Y} = 0.792X_3.$$

Example 4.5

In this example, Y is almost a perfect linear function of the last five X variables $(X_2, \ldots, X_6)$ while Y is more highly correlated with X_1 than the other X variables. Also, the largest correlations between the X variables are the correlations involving X_1.

The sample correlation matrix is

$$\widehat{\mathbf{R}} = \begin{array}{c} \begin{array}{ccccccc} Y & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 \end{array} \\ \left(\begin{array}{ccccccc} 1.000 & 0.805 & 0.681 & 0.669 & 0.702 & 0.702 & 0.698 \\ 0.805 & 1.000 & 0.581 & 0.572 & 0.601 & 0.605 & 0.598 \\ 0.681 & 0.581 & 1.000 & 0.352 & 0.361 & 0.392 & 0.386 \\ 0.669 & 0.572 & 0.352 & 1.000 & 0.372 & 0.384 & 0.365 \\ 0.702 & 0.601 & 0.361 & 0.372 & 1.000 & 0.428 & 0.422 \\ 0.702 & 0.605 & 0.392 & 0.384 & 0.428 & 1.000 & 0.400 \\ 0.698 & 0.598 & 0.386 & 0.365 & 0.422 & 0.400 & 1.000 \end{array} \right) \begin{array}{c} Y \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{array} \end{array}$$

The fitted standardized multiple regression model is:

$$\hat{Y} = 0.010X_1 + 0.277X_2 + 0.266X_3 + 0.272X_4 + 0.261X_5 + 0.269X_6 \qquad (R^2 = 0.936)$$

and the simple regression models are

$$\hat{Y} = 0.805X_1, \quad \hat{Y} = 0.681X_2, \quad \hat{Y} = 0.669X_3$$

$$\hat{Y} = 0.702X_4, \quad \hat{Y} = 0.702X_5, \quad \text{and } \hat{Y} = 0.698X_6.$$

Results from the five examples.

The relative importance given to each variable by the six different measures are given for each example in Table 4.3. Advocates of the RW measure argue that one

Table 4.3: Relative importances given by the orthogonal counterparts (OC), relative weights (RW) and general dominance (GD) measures and by three new measures (NM1, NM2 and NM3) in Examples 4.1–4.5

	OC	RW	GD	NM1	NM2	NM3
Example 4.1						
X_1	0.000	0.099	0.078	0.000	0.094	0.000
X_2	0.350	0.304	0.314	0.366	0.322	0.358
X_3	0.355	0.302	0.314	0.340	0.289	0.347
Example 4.2						
X_1	0.856	0.642	0.665	0.764	0.669	0.864
X_2	0.008	0.115	0.101	0.061	0.105	0.000
X_3	0.002	0.108	0.099	0.040	0.091	0.001
Example 4.3						
X_1	1.000	0.798	0.798	0.930	0.830	0.989
X_2	0.000	0.202	0.202	0.070	0.170	0.011
Example 4.4						
X_1	0.305	0.301	0.316	0.316	0.307	0.281
X_2	0.292	0.293	0.287	0.289	0.292	0.307
X_3	0.288	0.290	0.282	0.280	0.285	0.296
Example 4.5						
X_1	0.124	0.137	0.170	0.172	0.159	0.001
X_2	0.163	0.160	0.152	0.143	0.147	0.184
X_3	0.155	0.153	0.145	0.146	0.149	0.177
X_4	0.167	0.164	0.158	0.160	0.162	0.199
X_5	0.163	0.160	0.155	0.160	0.161	0.187
X_6	0.165	0.162	0.156	0.155	0.157	0.188

of its strengths is that it generally gives similar results to the general dominance (GD) measure. Table 4.3 shows that this was also the case for our examples, but the table shows that the NM2 measure also gives similar results to GD. Indeed, for Examples 4.2, 4.4 and 4.5 the relative importances assigned by GD are a little closer to those of NM2 than to those of RW. The results of the other measures (OC, NM1 and NM3) are often fairly similar to each other, especially those of OC and NM3, as in Examples 4.2 and 4.3. At the same time, NM1 is notably similar to GD in examples 4.4 and 4.5, and NM3 gives radically different results to all

other measures in Example 4.5.

In some of the examples, a variable's contribution to predicting Y was small but it was correlated with variables that were better predictors. Then the variable's relative importance was higher when measured by RW, GD or NM2 than when measured by OC, NM1 or NM3. This can be seen in Example 4.1, where X_1 is a poor predictor, and in Example 4.2, where X_2 and X_3 are poor predictors. In Example 4.1, the multiple regression suggests that X_1 makes some contribution to the prediction (albeit small) so the relative importance value it receives from OC, NM1 and NM3 seem too close to 0. Similarly, the relative importance values they give to X_2 and X_3 in Example 4.2 also seem too low.

The NM1 and NM3 measures, though conceptually quite similar to the OC measure, can give evaluations that are clearly more sensible than those of the OC measure. This is illustrated in Example 4.3, where the OC measure evaluates the relative importance of X_1 as 100% and the relative importance of X_2 as 0%. This is inappropriate, since X_1 on its own cannot explain all the variation in Y, while the combination of X_1 and X_2 *can* explain all the variation in Y, clearly showing that X_2 contributes usefully to the multiple regression model. The NM1 and NM3 measures evaluate the contribution of X_2 as small, but non-zero. The larger values given to X_2 by the RW, GD and NM2 measures are perhaps a better reflection of X_2's contribution, since on its own X_2 explains 20.3% of the variation in Y.

In Example 4.4 the three X variables all make similar contributions to the prediction of Y and the six measures all seem to reflect this appropriately. In Example 4.5, it is arguable whether X_1 is useful for predicting Y. On the one hand, X_1 makes little contribution to the multiple regression model while, on the

other hand, it is the best predictor of Y from simple regression. NM3 gives X_1 a relative importance that is close to 0, which might be considered appropriate in view of the multiple regression model. Other measures give it a much higher relative importance; indeed, GD and NM1 evaluate it as the most important predictor which, to us, seems inappropriate. Example 4.5 also shows that the RW and GD measures are not always in close agreement: while GD evaluates X_1 as the most important variable in the regression model, RW evaluates it as the least important.

4.7.2 Orthogonal rotation and variable selection

Two examples are examined in this subsection. In the first, two of the explanatory variables are highly correlated and we consider both the model with the original variables and the model that results from rotating the correlated variables. Measures of relative importance are applied to both models and their differences are examined. In the second example, one variable has a regression coefficient that does not differ significantly from 0 (at the 5% level of significance). We examine how dropping this variable from the model effects the relative importances of the other variables.

Example 4.6 Orthogonal rotation

The Longley dataset (Longley, 1967) is well-used as an example of highly collinear regression. The dataset contains annual values of various US macroeconomic variables for the years 1947-1962. Here we use five of its variables: *npe* (number of thousands of people employed), GNP_1 (GNP implicit price deflator), GNP_2 (GNP

in millions of dollars), *npue* (number of thousands of unemployed people) and *npa* (number of people in the armed forces). We take *npe* as the response variable and initially take the other four variables as the explanatory variables.

The following is the sample correlation matrix for these variables:

$$\widehat{\mathbf{R}} = \begin{array}{c} \begin{array}{ccccc} npe & GNP_1 & GNP_2 & npue & npa \end{array} \\ \begin{pmatrix} 1.000 & 0.971 & 0.984 & 0.502 & 0.457 \\ 0.971 & 1.000 & 0.992 & 0.621 & 0.465 \\ 0.984 & 0.992 & 1.000 & 0.604 & 0.446 \\ 0.502 & 0.621 & 0.604 & 1.000 & -0.177 \\ 0.457 & 0.465 & 0.446 & -0.177 & 1.000 \end{pmatrix} \end{array} \begin{array}{c} npe \\ GNP_1 \\ GNP_2 \\ npue \\ npa \end{array}$$

The fitted standardized multiple regression model is:

$$\widehat{npe} = 0.173GNP_1 + 0.998GNP_2 - 0.227npue - 0.109npa, \qquad (4.29)$$

for which $R^2 = 0.986$.

The correlation matrix shows that there is a strong collinearity between two of the explanatory variables, GNP_1 and GNP_2. Collinearity can radically affect the values of parameter estimates and will inflate their variances. Transforming variables to remove collinearity is consequently attractive and here we replace GNP_1 and GNP_2 by the variables

$$X_1 = (GNP_1 + GNP_2)/\sqrt{2} \quad \text{and} \quad X_1 = (GNP_1 - GNP_2)/\sqrt{2}.$$

This is equivalent to multiplying the original variables by the orthogonal rotation

Table 4.4: Relative importances of variables before and after rotation

	OC	RW	GD	NM1	NM2	NM3
Relative importances before rotation						
GNP_1	0.400	0.390	0.390	0.527	0.361	0.088
GNP_2	0.526	0.417	0.411	0.371	0.393	0.888
npue	0.023	0.099	0.104	0.046	0.139	0.005
npa	0.036	0.079	0.081	0.042	0.092	0.004
Relative importances after rotation						
X_1	0.922	0.682	0.687	0.891	0.550	0.967
X_2	0.004	0.014	0.015	0.007	0.183	0.004
npue	0.023	0.161	0.156	0.046	0.146	0.004
npa	0.036	0.128	0.128	0.042	0.107	0.011

matrix,

$$\mathbf{\Gamma} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The new variables X_1 and X_2 are uncorrelated.

Regressing *npe* on the transformed set of variables gives the equation

$$\widehat{npe} = 1.681X_1 - 0.054X_2 - 0.227npue - 0.109npa. \tag{4.30}$$

Theory implies that the regression coefficients of the unrotated components (*npue* and *npa*) should be unchanged — comparison of equations (4.29) and (4.30) shows that this is indeed the case. Also, the R^2 value is again 0.986. However, with some measures of relative importance, the importances of *npue* and *npa* in the pre-rotation model (equation (4.29)) will differ from their importances in the post-rotation model (equation (4.30)). This can be seen in Table 4.4, where the relative importances given by our six measures of importance are presented.

In line with theory, the table shows that the relative importances given by the

OC and NM1 measures to *npue* and *npa* are unchanged by the rotation of GNP_1 and GNP_2. With the other measures, the relative importances given to *npue* and *npa* do change, though the degree of change varies with the measure. With NM3 the importance values change by a large proportion (e.g., from 0.004 to 0.011), though the changes are small in absolute terms. With the RW and GD measures the changes are quite large — noticeably larger (at least three times larger) than with the NM2 measure. Interestingly, values given by the NM2 measure are straddled by the before/after values given by the RW and GD measures, and are quite close to the averages of the before/after values given by both the RW measure and the GD measure. For example, the RW measure gives before/after values of 0.079 and 0.128 to *npa*, and their average is relatively close to the values 0.092 and 0.107 that NM2 gives to *npa*.

Example 4.7 Variable selection

Wood (1973) presents data from a process variable study of a petroleum refinery unit. The dependent variable (Y) is the octane value of the petroleum produced and there are four independent variables: three relate to feed composition (X_1, X_2, X_3) and the fourth relates to process conditions (X_4). Eighty-two observations were taken, giving the following sample correlation matrix:

$$\widehat{\mathbf{R}} = \begin{array}{c} \begin{array}{ccccc} Y & X_1 & X_2 & X_3 & X_4 \end{array} \\ \left(\begin{array}{ccccc} 1.000 & -0.870 & 0.392 & -0.638 & 0.629 \\ -0.870 & 1.000 & -0.589 & 0.449 & -0.337 \\ 0.392 & -0.589 & 1.000 & -0.298 & 0.161 \\ -0.638 & 0.449 & -0.298 & 1.000 & -0.722 \\ 0.629 & -0.337 & 0.161 & -0.722 & 1.000 \end{array} \right) \end{array} \begin{array}{c} Y \\ X_1 \\ X_2 \\ X_3 \\ X_4 \end{array}$$

After standardizing the variables, regression of Y on the four independent variables gave

$$\widehat{Y} = -0.824X_1 - 0.172X_2 - 0.097X_3 + 0.309X_4, \quad (R^2 = 0.905) \tag{4.31}$$

as the regression model. There is clear evidence that X_1, X_2 and X_4 should be included in the regression model ($p < 0.0002$ for each of these three variables) but whether X_3 should be included is debatable. The null hypothesis that the regression coefficient for X_3 is zero is rejected only at significance level 0.07. Omitting X_3 from the model gives the regression equation

$$\widehat{Y} = -0.841X_1 - 0.163X_2 + 0.372X_4, \quad (R^2 = 0.902). \tag{4.32}$$

The top half of Table 4.5 displays the real importance assigned to the different X variables by the different measures when all four X variables are included in the regression model. Surprisingly, all but one of the measures gives X_3 a higher relative importance than X_2, even though X_3 is the variable whose inclusion in the model is tenuous. The NM3 measure is the exception. It gives X_3 a relative importance of 0.0, which concords fully with the inference that X_3 can reasonably be omitted from the regression model.

Table 4.5: Relative importances of variables before and after omitting X_3

	OC	RW	GD	NM1	NM2	NM3
Relative importances before omitting X_3						
X_1	0.593	0.515	0.518	0.488	0.439	0.685
X_2	0.012	0.066	0.064	0.009	0.051	0.060
X_3	0.121	0.146	0.150	0.160	0.178	0.000
X_4	0.180	0.179	0.173	0.250	0.238	0.161
Relative importances after omitting X_3						
X_1	0.640	0.570	0.578	0.604	0.542	0.665
X_2	0.016	0.075	0.075	0.017	0.072	0.074
X_4	0.246	0.257	0.248	0.281	0.287	0.162

The lower half of Table 4.5 shows the relative importances assigned to X_1, X_2 and X_4 after X_3 has been omitted from the model. In the whole of the table, the RW and GD measures are strikingly similar in all their evaluations. It is also the case that all measures evaluate X_1 as the most important variable and X_4 as the second most important (both before and after omitting X_3). In other respects though, there is limited agreement across measures. For example, NM1 and NM2 agree quite closely in their evaluations of X_1 and X_4, but NM1 is similar to OC in its evaluation of X_2, while NM2's evaluations of X_2 are similar to those of RW, GD and NM3.

With most measures, the relative importance of X_3 is far greater than the difference between the R^2 values of the models in equations (4.31) and (4.32). Hence, with those measures the omission of X_3 must substantially increase the relative importance of at least one X variable. As X_3 has higher absolute correlation with X_4 than with X_1 or X_2, it might be anticipated that omitting X_3 would increase the relative importance of X_4 more than that of X_1 or X_2. This is indeed the case for the OC, RW and GD measures, but not for NM1, NM2 or NM3. It seems

then, that the effects on relative importance of omitting a variable are somewhat unpredictable and can vary markedly with the choice of measure.

4.8 Concluding comments

Six measures for evaluating the relative importance of predictor variables in a regression have been examined. From the examples presented in Section 4.7, it is clear that usually there is some consensus between them — variables given a high relative importance by one measure are usually given a high relative importance by other measures, and similarly for low relative importance. At the same time, in each example there were differences between the measures in their evaluations, and some differences were substantial.

Occasionally, common sense shows that an evaluation is unreasonable. For instance, in Example 4.3 the OC measure evaluated the relative importance of X_1 as 100% and that of X_2 as 0%. This is clearly inappropriate, as all the variation in Y could not be explained by X_1 on its own, but could be explained by the combination of X_1 and X_2. Often though, the evaluations of the different measures all seem reasonable and how to choose between them is not clear-cut, because there are no known 'correct' evaluations with which to make comparison. As noted by Johnson and LeBreton (2004, p.240), "Because there is no unique mathematical solution to the problem [of evaluating relative importances], these indices [measures] must be evaluated on the basis of the logic behind their development, the apparent sensibility of the results they provide, and whatever shortcomings can be identified."

The following arguments favour different measures.

1. The GD and RW measures have been the most widely recommended measures in recent years, partly because they typically give similar evaluations, suggesting that there is an underlying construct that they both appraise. The examples presented here support that rationale, as they give further evidence that the two measures generally give similar results — there is only one case (variable X_1 in Example 4.5) where the GD and RW evaluations differ appreciably. Most often, the GD measure is closer to RW than to any other measure, though there were examples where GD was closer to NM2.

2. The OC measure has the benefit of simplicity, so that the relationship between a variable and the evaluation of its relative importance is fairly direct. A further merit of the OC measure is that it has the rotation invariance property.

3. In constructing the new measures (NM1–3), the aim was to improve upon the OC and RW measures by letting Y influence the transformation to orthogonality, rather than determining the transformation from just the values of the regressors. This was motivated by the observation that the transformation's purpose is to help evaluate the relationship between Y and the regressors, so both should be taken into account in forming the transformation. On that basis, NM1 is to preferred over OC, since in other respects the construction of the two measures are very similar. Similarly for NM2 and RW. Regarding NM1 and OC, the two measures tend to give similar results but, when there are larger differences, the NM1 evaluations tend to be closer to the consensus of all six measures, perhaps giving further reason to favour NM1 over OC.

4. In the examples, a feature of NM3 is that it gave low relative importance to variables that might reasonably be omitted from the regression, which could be considered an attractive characteristic. In Example 4.7, for instance, it gave X_3 a relative importance of 0.000 while other measures gave it a relative importance of 0.121 or more. Similarly, in both Examples 4.1 and 4.5, predictions of Y are not improved by including X_1 in the regression model but, in Example 4.1, OC, NM1 and NM3 gave X_1 a low evaluation and, in Example 4.5, only NM3 gave it a low evaluation.

The new measures presented here and ideas behind them could be adapted to give other measures of potential value. In particular, any of the OC, RW and NM2 measures could be modified to use regression coefficients as weights when forming orthogonal counterparts, in the same way that NM3 is derived from NM1. The weighting scheme could also be generalised to use the $(|\widehat{\beta}_j|)^\alpha$ as weights (where the $\widehat{\beta}_j$ are the multiple regression coefficients). Setting α equal to 0 would correspond to 'no weighting', and increasing α would increase the importance of the weighting.

Chapter 5

Identifying variables underlying multicollinearity

5.1 Introduction

When the regressors are orthogonal, the use and interpretation of a multiple regression model is straightforward. But in real life problems, many of the regressors are correlated. There is no precise definition of collinearity in the literature. Originally it meant that there is a perfect linear relationship between two or more regressors. In practice, such exact collinearity rarely happens. A broader sense of collinearity refers to a near linear relationship between two or more regressors. The term near linear relationship implies that one of the regressor can be approximately expressed as a linear combination of one or more of the remaining regressors. For this chapter, we use the term multicollinearity or collinearity to indicate near linear relationship among some or all regressors of the model.

Collinearity has a number of practical consequences. When there is an exact

collinearity, the ordinary least squares (OLS) estimate of $\boldsymbol{\beta}$ is not unique and variances of the individual estimates are infinite (Gujarati, 2003, p.345–347). When there is approximate collinearity, OLS estimates have large variances and covariances, and one or more important regressors may have low t-ratios and much wider confidence intervals (statistically insignificant), even if the overall model parameters are significant and the model has a high R^2 value. Also, OLS estimates and their variances will be sensitive to minor changes in the data (Gujarati, 2003, p350), so the addition of new observations to the data or deletion of observations from the data can change the regressors that variable selection chooses to include in the model. It also inflates some of the OLS estimates in absolute value, i.e., the estimated vector $\widehat{\boldsymbol{\beta}}$ is generally longer than the true parameter vector $\boldsymbol{\beta}$. As a result, the average of the squared distance between the parameters and the estimates can become large (Ofir and Khuri, 1986). In some cases one or more of the estimates may even produce incorrect signs of the OLS estimates, contradicting the relationship of these regressors and the response variable (Hocking, 2003, p.153).

In summary, inferences and prediction can be misleading in the presence of multicollinearity. So identification of and remedy of collinearity deserves more attention. This chapter focuses on the diagnostics and identification of collinear sets.

Several methods for detecting multicollinearity have been proposed in the literature. None of them can uniquely identify collinearity in the data. Detailed description of the methods will provide clearer guidelines for the researchers. The simplest of them examines the pairwise correlations between regressors. This

method is applicable for identifying pairwise collinearity but is not suitable for the detection of more complex collinearity. Farrar and Glauber (1967) proposed a three-stage procedure to detect and locate interdependence and identify its pattern. However, their tests are seriously criticized by Haitovsky (1969), Wichers (1975), Kumar (1975), and O'Hagan and McCabe (1975). Most researchers prefer to use condition indices and variance inflation factors, which are related to the correlation matrix of the regressors.

The above methods are used either for the diagnosis of collinearity or detecting the number of collinear sets. However, they cannot identify the collinear sets, i.e., the sets of regressors that are collinear. Gunst and Mason (1977) and Belsley et al. (1980) suggested two procedures for identifying collinear sets. Both of them are based on the eigenvalues and eigenvectors of the correlation matrix of the regressors. However, Belsley et al.'s (1980) procedure provides more detailed information about collinearity sets than Gunst and Mason's (1977) procedure. More recently, Garthwaite et al. (2012) proposed a procedure for identifying the collinear sets that is based on the inverse of the square-root matrix of the correlation matrix of the regressors. Basically, this procedure partition the variance inflation factors into individual contribution of regressors. Hence, also, this procedure provides very detailed information about collinearity sets. For example, if there is a collinearity among three variables, then three rows of the transformation matrix will provide this information.

Good reviews of works are given in Belsley et al. (1980), Ofir and Khuri (1986), Gujarati (2003) and Montgomery et al. (2015). The above mentioned three procedures for identifying collinear sets are compared through examples from three

published studies that identify collinear sets using either procedure.

In designed experiments, the values of the explanatory variables are usually fixed at values that avoid multicollinearities. This is the case, for example, in standard designs involving factors, such as Latin squares or randomized block designs. With such designs, analysis of variance is commonly used to test for the importance of variables and $\mathbf{X}^\top\mathbf{X}$ is a non-singular matrix. Our methods can be applied to the data from such experiments. However, in some designed experiments (for example, when there is confounding) $\mathbf{X}^\top\mathbf{X}$ is singular and then our methods cannot be applied.

Diagnosis and identification of the number of collinearities in the dataset is discussed in Section 5.2. In Section 5.3 we discuss three procedures for identifying collinearity sets. Comparison of the procedures by using published examples that address multicollinearity using either procedure are given in Section 5.4. Concluding comments are given in Section 5.6.

5.2 Detection of Multicollinearity

Suppose we have a multiple regression model of Y on $X_1, \ldots, X_p$

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon. \tag{5.1}$$

For n observations from a sample, the above model can be written in matrix form as

$$\boldsymbol{y} = \beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{5.2}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of response variable, $\mathbf{1}$ is an $n \times 1$ vector of ones, $\mathbf{X}$ is an $n \times p$ matrix of known values of the regressors $X_1, \ldots, X_p$, $\boldsymbol{\beta}$ is a $p \times 1$

vector of unknown parameters and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of random error terms, with $\text{var}(\epsilon_i) = \sigma^2$ and the ϵ_i's independently distributed. For simplicity, all $\boldsymbol{X}$ variables and Y are standardized to have sample means of 0 and unit lengths. Consequently, $\mathbf{X}^\top\mathbf{X}$ is the sample correlation matrix of the regressors and $\mathbf{X}^\top\boldsymbol{y}$ is the vector of sample correlations between the response and the regressors. We denote the correlation matrix of the regressors by $\mathbf{R}_{xx}$, i.e., $\mathbf{R}_{xx} = \mathbf{X}^\top\mathbf{X}$.

Essentially, multicollinearity depends on the particular sample and is not a statistical problem. The regressors from a particular set of sample data may be multicollinear, when the regressors are not correlated in the population from which the sample was taken. There are several methods of detecting multicollinearity, but there is no unique simple method for its detection. Instead, there are some rules of thumb for detecting the number of collinearities and identifying the collinearities sets. In the next subsections, some of these rules will briefly be discussed.

The dataset used in Chapter 4 from Neter et al. (1983) is used to explain the common methods of identifying the multicollinearity. The dataset has four variables BF, TST, TC and MC. The variable BF is considered as the response variable. The estimated standardized regression model is

$$\widehat{BF} = 4.264TST - 2.929TC - 1.561MC \tag{5.3}$$

5.2.1 Inspection of R^2, F and t-statistics

If X_j and Y are highly correlated, but $H_0 : \beta_j = 0$ is not rejected, it suggests a collinearity that involves X_j. If the overall F-test for the hypothesis $H_0 : \beta_1 = \ldots = \beta_p = 0$ is significant but none or very few of the individual model parameters are significantly different from zero, it suggests multicollinear-

ity. That is, if there is an overall significant F-test but insignificant t-values for all the individual model parameters or most of the individual parameters, we may suspect that multicollinearity is present among the regressors. At the same time, a few insignificant t-values and a high R^2 does not mean multicollinearity.

Sample size has an effect on the F test. If the sample size become larger and R^2 stays the same, the significance of the F- statistic increases. Hence, for a given R^2, with a large sample size there is less uncertainty when rejecting the null hypothesis that the population regression coefficients are all zero. With a small sample size there is greater uncertainty.

With the dataset from Neter et al. (1983), the model has $R^2 = 0.801$ and $F = 22.86$ with 3 and 17 df. This is significant at 1% level of significance ($p = 0.000$) but none of the individual regression parameters are significant (p =0.157, 0.270, and 0.176 respectively), suggesting the likelihood of multicollinearity.

5.2.2 Examination of the correlation matrix of the regressors

If a model has only two regressors, the correlation coefficient determines the degree of collinearity. For a model with more than two regressors, any pairwise correlations between the regressors greater than 0.8 or 0.9 indicates multicollinearity (Farrar and Glauber, 1967, p.98). According to Gujarati (2003, p.359), high pairwise correlations are not a necessary condition for the existence of multicollinearity but they are a sufficient condition. If more than two regressors form a multicollinearity set, it is not necessary for any of the pairwise correlations to be large. Pairwise correlation can be used only to detect pairwise multicollinearity and inspection of

Table 5.1: Correlation matrix of the body fat data

Variable	BF	TST	TC	MC
BF	1.000	0.843	0.878	0.142
TST	0.843	1.000	0.924	0.458
TC	0.878	0.924	1.000	0.085
MC	0.142	0.458	0.085	1.000

pairwise correlation is not sufficient for identifying the presence of multicollinearity among more than two regressors (Montgomery et al., 2015, p.296). According to Klein (1962), multicollinearity is harmful if the pairwise correlation between two regressors is greater than or equal to the overall multiple correlation, i.e., $r_{jk} \geq R_y$, where r_{jk} is the correlation between X_j and X_k, and R_y is the multiple correlation between the response and the regressors.

The correlation table of Chapter 4 in page 92 is reproduced below. It shows that the correlation between TST and TC is high (0.924), so a collinearity is indicated between TST and TC.

5.2.3 Examination of the determinant of the correlation matrix of the regressors

The determinant of the correlation matrix of the regressors can be used as a measure of multicollinearity (Ofir and Khuri, 1986). The range of the determinant of $\mathbf{R}_{xx}$ is $0 \leq |\mathbf{R}_{xx}| \leq 1$. When the regressors are orthogonal, then $|\mathbf{R}_{xx}| = 1$. If $|\mathbf{R}_{xx}| = 0$, there is a perfect linear dependency between some of the regressors. $|\mathbf{R}_{xx}|$ becomes closer to zero if multicollinearity becomes more severe. On the basis of the distribution of $|\mathbf{R}_{xx}|$, Farrar and Glauber (1967) proposed a statistical test to check the existence and severity of multicollinearity. They

viewed multicollinearity as any departure from orthogonality. The hypothesis is $H_0 : The\ X's\ are\ orthogonal,$ versus $H_1 : The\ X's\ are\ not\ orthogonal.$ The test statistic has the form

$$\chi^2 = -\left[n - 1 - \frac{1}{6}(2p + 5)\right] log_e|\mathbf{R}_{xx}|. \tag{5.4}$$

The test statistic approximately follows a chi-square distribution with $v = \frac{1}{2}p(p-1)$ degrees of freedom. If the calculated value of chi-square, χ^2_{cal}, from a sample is greater than the tabulated value of $\chi^2_v(\alpha)$ for a significance level α, it is concluded that collinearity exists between some of the regressors. This chi-square statistic is basically Bartlett's (1954) sphericity test, which is based on the Wishart distribution. Under the assumption that $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ follows a multivariate normal distribution, the sample correlation matrix, $\mathbf{R}_{xx}$, is distributed as a Wishart distribution. The chi-square test of Farrar and Glauber (1967) has been criticized by Haitovsky (1969). Haitovsky point out that the existence of this test requires $\mathbf{X}$ to be stochastic, while one of the assumption of the linear regression model is that $\mathbf{X}$ is fixed. Also the test statistic is sensitive to sample size.

The calculated value of the chi-square test statistic from the body fat data is $\chi^2_{cal} = 100.75$, which is significant at the 1% level of significance ($\chi^2_3(0.01) = 11.34$). This again indicates the presence of multicollinearity among the regressors. The **R** package "**mctest**" can be used to calculate the chi-square test statistic for testing the overall collinearity.

5.2.4 Examination of R^2 from auxiliary regressions

The regressions of X_j $(j = 1, \ldots, p)$ on $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$ are called the auxiliary regressions of the main regression of Y on $X_1, \ldots, X_p$. One way of

identifying which particular regressor is collinear with the remaining regressors is by performing overall F-tests for the auxiliary regressions (Farrar and Glauber, 1967). For X_j, the hypothesis to be tested is $H_0 : R^2_{j.1,\ldots,p} = 0$ versus $H_1 : R^2_{j.1,\ldots,p} \neq 0$. The test statistic is

$$F = \frac{R^2_{j.1,\ldots,p}/(p-1)}{(1 - R^2_{j.1,\ldots,p})/(n-p)}. \tag{5.5}$$

Which is distributed as F with $p-1$ and $n-p$ degrees of freedom. The null hypothesis is rejected at significance level α, if the observed value of F from the sample is greater than the tabulated value $F_{(p-1),(n-p)}(\alpha)$. If the overall F-test for a particular auxiliary regression is significant it means that the particular regressor X_j is collinear with $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$.

The F-statistics corresponding to TST, TC and MC are 62.47, 45.68 and 10.81 respectively. The degrees of freedom of the F distribution are (2, 17) and the corresponding tabulated value at the 1% level of significance is 6.11. We therefore conclude that TST can be predicted from TC and MC, TC can be predicted from TST and MC and, in addition, MC can be predicted from TST and TC. The **R** package "**mctest**" can also be used to calculate F- statistics to test the significance of auxiliary regressions.

Instead of performing the overall F-test for all auxiliary regressions, one can compare the R^2 value of an auxiliary regression with the R^2 value of the main regression model. If we let R^2_y denote R^2 for the full model and R^2_j denote R^2 for the auxiliary regression ($j = 1, \ldots, p$), then Klein (1962) suggested that a particular regressor X_j ($j = 1, \ldots, p$) is collinear with $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$ if R^2_j is greater than R^2_y.

5.2.5 Examination of partial correlations

Partial correlation is the simple correlation between two variables after controlling for the effect of other regressors. If we have three variables, X_1, X_2 and X_3, the partial correlation between X_1 and X_2 is denoted by $r_{12.3}$ and is defined by

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \tag{5.6}$$

where r_{jk}'s are the simple correlations.

For more than three variables, the partial correlation can be obtained from the inverse of the correlation matrix. The partial correlation coefficient between X_j and X_k after adjusting for the other variables is defined by

$$r_{jk.} = \frac{-r^{jk}}{\sqrt{r^{jj}r^{kk}}}, \tag{5.7}$$

where r^{jk} is the (j,k)th off-diagonal element of $\mathbf{R}_{xx}^{-1}$, and r^{jj} and r^{kk} are the jth and kth diagonal elements of $\mathbf{R}_{xx}^{-1}$ respectively.

The hypothesis to be tested is $H_0 : r_{x_jx_k.x_1,\ldots,x_p} = 0$, versus $H_1 : r_{x_jx_k.x_1,\ldots,x_p} \neq 0$. As given by Farrar and Glauber (1967), the statistic

$$t_{jk.} = \frac{r_{jk.}\sqrt{n-p}}{\sqrt{1 - r_{jk.}^2}} \tag{5.8}$$

has Student's t-distribution with $v = (n-p)$ degrees of freedom and can be used to assess the interdependence pattern among the regressors. If the calculated value of the t-statistic from the sample is greater than the theoretical value $t_v(\alpha)$ for a significance level α, then the regressors X_j and X_k are considered to be collinear.

Wichers (1975, p.367) argues that the partial correlation test is not appropriate for detecting multicollinearity because entirely different multicollinearity patterns may produce the same partial correlation.

Table 5.2: t-statistics and its associated p-values (p-values are given inside the brackets)

Variable	TST	TC	MC
TST	–	5.979 (0.000)	1.975 (0.074)
TC	–	–	-0.469 (0.648)

Table 5.2 gives the t-statistics values for all pairs of partial correlations obtained from the body fat data. The test statistic is significant only for the variable pair TST and TC. On that basis the variables TST and TC are collinear. Partial correlations can be tested using the "**ppcor**" package in **R** language.

5.2.6 Variance inflation factors

Variance inflation factors (VIFs) are the most commonly used method of identifying the existence of collinearities and which variables are involved in collinearities. The VIF measures the increase of variance of the regression estimator with the increase of multicollinearity. The term VIF was first used by Marquaridt (1970).

The variance-covariance matrix of the regression estimator $\widehat{\boldsymbol{\beta}}$ of a multiple regression model is

$$\text{var}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \tag{5.9}$$

where σ^2 is the common variance of the random error term. The covariance of the estimators are related to the off-diagonal elements of $(\mathbf{X}^\top \mathbf{X})^{-1}$, while the variance of the estimators are related to the diagonal elements of $(\mathbf{X}^\top \mathbf{X})^{-1}$. That is

$$\text{var}(\widehat{\boldsymbol{\beta}}_j) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}_{jj}. \tag{5.10}$$

Suppose the regressor set, $\boldsymbol{X}$, is partitioned as

$$\boldsymbol{X} = \left(\boldsymbol{X}_{(1)} \vdots \boldsymbol{X}_{(2)}\right) \tag{5.11}$$

where $\boldsymbol{X}_{(1)} = X_1$ and $\boldsymbol{X}_{(2)} = (X_2, \ldots, X_p)^\top$. The corresponding partitioning of the correlation matrix $\mathbf{R}_{xx} = \mathbf{X}^\top\mathbf{X}$ is

$$\mathbf{X}^\top\mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1^\top\boldsymbol{x}_1 & \boldsymbol{x}_1^\top\mathbf{X}_{(2)} \\ \mathbf{X}_{(2)}^\top\boldsymbol{x}_1 & \mathbf{X}_{(2)}^\top\mathbf{X}_{(2)} \end{pmatrix}. \tag{5.12}$$

From the standard result for the inverse of a portioned matrix, the upper-left corner, r^{11} of the inverse of $\mathbf{X}^\top\mathbf{X}$ is

$$r^{11} = \left[\boldsymbol{x}_1^\top\boldsymbol{x}_1 - \boldsymbol{x}_1^\top\mathbf{X}_{(2)}\left(\mathbf{X}_{(2)}^\top\mathbf{X}_{(2)}\right)^{-1}\mathbf{X}_{(2)}^\top\boldsymbol{x}_1\right]^{-1}. \tag{5.13}$$

Now R^2 value from regressing X_1 on $\boldsymbol{X}_{(2)} = (X_2, \ldots, X_p)^\top$ is

$$R_1^2 = \frac{\boldsymbol{x}_1^\top\mathbf{X}_{(2)}\left(\mathbf{X}_{(2)}^\top\mathbf{X}_{(2)}\right)^{-1}\mathbf{X}_{(2)}^\top\boldsymbol{x}_1}{\boldsymbol{x}_1^\top\boldsymbol{x}_1} = \boldsymbol{x}_1^\top\mathbf{X}_{(2)}\left(\mathbf{X}_{(2)}^\top\mathbf{X}_{(2)}\right)^{-1}\mathbf{X}_{(2)}^\top\boldsymbol{x}_1. \tag{5.14}$$

Since $\boldsymbol{x}_1^\top\boldsymbol{x}_1 = 1$, r^{11} can be expressed as

$$r^{11} = \left(1 - R_1^2\right)^{-1} \tag{5.15}$$

and $\text{var}(\widehat{\beta}_1)$ from equation (5.10) can be rewritten as

$$\text{var}(\widehat{\beta}_1) = \sigma^2\frac{1}{(1 - R_1^2)}. \tag{5.16}$$

The factor $(1 - R_1^2)^{-1}$ is called the variance inflation factor (VIF_1) for the variance of $\widehat{\beta}_1$. In general,

$$\text{var}(\widehat{\beta}_j) = \sigma^2\frac{1}{\left(1 - R_j^2\right)}, \tag{5.17}$$

where R_j^2 is the R^2 value from regressing X_j on $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$ and σ^2 is the variance of the random error term ϵ. The factor $\left(1 - R_j^2\right)^{-1}$ is called the VIF of X_j (Farrar and Glauber, 1967). The VIF_j is the jth diagonal element of the inverse of the correlation matrix $\mathbf{R}_{xx}$ (Hocking, 2003, p.166).

It can also be shown that the VIF associated with a predictor X_j is the ratio of the variance of the estimated coefficient, var($\widehat{\beta}_j$), form a model with correlated regressors and the variance of the estimated coefficient, var($\widehat{\beta}_{j0}$), of the model when the regressors are orthogonal (Hocking, 2003, p.166). So VIF$_j$ measures the amount of inflation of the variance of $\widehat{\beta}_j$ due to the inclusion of correlated regressors in the model. For example, a VIF$_j$ of 5 tells us that the variance of $\widehat{\beta}_j$ is 5 times inflated due to its collinearity with other regressors. Also the width ratio of the confidence interval of β_j from the observed data relative to the confidence interval obtained from the model with orthogonal regressors is equal to the square root of VIF_j (Ofir and Khuri, 1986).

There is no theory-based rule for what values of VIF indicate collinearity. There is a rule of thumb that a VIF greater than 10 indicates collinearity and this rule is often adopted (Neter et al., 1983, p.392). However, Montgomery et al. (2015, p.296) claims that the VIF for a regressor greater than 5 or 10 indicates that the associated regression coefficient is poorly estimated due to collinearity. Also, according to O'Brien (2007), a VIF of even 4 indicates serious multicollinearity.

For the correlation matrix in page 141, the VIF's are 708.84 (TST), 564.34 (TC) and 104.61 (MC). Hence all three regressors are involved in a collinearity, or collinearities. General statistical packages for identifying collinearity automatically give VIFs, making them convenient to use.

5.2.7 Eigensystem analysis of the correlation matrix of the regressors

An eigensystem analysis of $\mathbf{X}^\top\mathbf{X}$ is widely used to detect multicollinearity. Here, $\mathbf{X}^\top\mathbf{X}$ is both the cross-product matrix of the standardized data matrix $\mathbf{X}$ and the correlation matrix $\mathbf{R}_{xx}$ of the regressors. Discussion of the eigensystem analysis is usually based on the correlation matrix because multicollinearity entirely depends on the correlation structure of the regressors. Suppose $\lambda_1 \geq \ldots, \geq \lambda_p$ are the ordered eigenvalues of the correlation matrix $\mathbf{R}_{xx}$, where $\sum_{j=1}^{p} \lambda_j = p$, the number of columns of $\mathbf{X}$. If the regressors are uncorrelated then $\lambda_j = 1$, for $j = 1, \ldots, p$. For correlated regressors, some of the eigenvalues are greater than 1 and some are less than 1. Since the product of the eigenvalues of a square matrix is equal to the determinant of that matrix, one or more small eigenvalues implies that the correlation matrix is near-singular. Near-singularity in the correlation matrix implies that one or more near-linear dependencies exist among the columns of the data matrix, $\mathbf{X}$.

If there is a perfect multicollinearity between some variables, one of the eigenvalues of $\mathbf{R}_{xx}$ will be exactly equal to zero. The number of collinear sets in the data is equal to the number of near zero eigenvalues of $\mathbf{R}_{xx}$. For example, if $\lambda_1 = 2.532$, $\lambda_2 = 0.460$ and $\lambda_3 = 0.008$, then the dataset has only one collinear set. On the other hand, if $\lambda_1 = 2.932$, $\lambda_2 = 0.040$ and $\lambda_3 = 0.028$, then there are two collinear sets in the data.

Some analyst prefer to use the condition number of the correlation matrix $\mathbf{R}_{xx}$ as a measure of overall multicollinearity. The condition number is generally used as a measure of numerical instability in a correlation matrix $\mathbf{R}_{xx}$ (a positive

definite matrix). The condition number of $\mathbf{R}_{xx}$ is defined as

$$\text{CN} = \frac{\text{Maximum eigenvalue}}{\text{Minimum eigenvalue}} = \frac{\lambda_1}{\lambda_p}. \tag{5.18}$$

For orthogonal regressors, all the eigenvalues of $\mathbf{R}_{xx}$ are one and hence the condition number is 1. Since the sum of the diagonal elements of the correlation matrix is equal to the number of regressors, the sum of the eigenvalues of that correlation matrix is equal to p, the number of regressors. So if a correlation matrix has some small eigenvalues, it also has some large eigenvalues. Consequently, the condition number is always greater than 1 for non-orthogonal regressors. A condition number between 100 and 1000 indicates moderate to strong multicollinearity and a condition number greater than 1000 is often used to indicate severe multicollinearity (Gujarati, 2003, p.362; Montgomery et al., 2015, p.298).

A related measure used to identify the number of collinear sets in the data is the condition index (CI). The condition indices κ_j of $\mathbf{R}_{xx}$ are

$$\kappa_j = \sqrt{\frac{\lambda_{max}}{\lambda_j}}, \text{ for } j = 1, \ldots, p. \tag{5.19}$$

The largest condition index equals the square root of the condition number. In the literature, there is no agreement on specific thresholds for condition index that strictly determine the strength (degree) of multicollinearity. However, empirical understanding of Belsley et al. (1980, p.105) suggest that κ_j's around 5 or 10 indicate weak dependencies, whereas values of κ_j's between 30 to 100 indicate moderate to strong collinearity. In addition, Rawlings (1988, p.371) suggests that κ_j's around 10 indicate weak dependencies, 30 to 100 indicate moderate to strong collinearity and serious collinearity is indicated by κ_j's greater than 100. To the best of our knowledge, there is no suggestion for condition indices between 10 and 30.

Condition indices have limited value in the diagnosis of collinearity. They can only detect the number of collinearities but cannot identify the variables which are involved in the multicollinearity. Standard statistical packages for collinearity diagnostics automatically give condition number and condition indices. The condition indices for the body fat data are 1.000, 1.488 and 53.329, which indicate that there is only one collinear set.

Thisted (1980) suggested two multicollinearity indexes. These are

$$\text{mci} = \sum_{j=1}^{p} \left(\frac{\lambda_p}{\lambda_j} \right)^2 \tag{5.20}$$

and

$$\text{pmci} = \sum_{j=1}^{p} \frac{\lambda_p}{\lambda_j} \tag{5.21}$$

where λ_p is the smallest eigenvalue of $\mathbf{R}_{xx}$. The mci is for estimation and pmci is for prediction. They satisfy the inequality $1 < \text{mci} \leq \text{pmci} \leq p$, with equality holding for orthogonal regressors. For orthogonal regressors, both measures have the value p and with increasing collinearity they go towards one. That is, high multicollinearity is indicated by values close to one; if their values are greater than two it indicate relatively little or no multicollinearity. According to Fellman et al. (2009), these measures can be used only if there is one small eigenvalue but are unreliable for several small eigenvalues. To the best of our knowledge there is no **R** package for calculating Thisted's collinearity measures.

5.2.8 Expected squared distance between $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$

For collinear data, the OLS estimates are usually large in absolute value. So the average squared distance between the true parameter values and the OLS estimates

are sometimes used as a measure of multicollinearity. Denote the squared distance by D^2, then

$$\begin{aligned} E(D^2) &= E\left[\sum_{j=1}^{p}(\widehat{\beta}_j - \beta_j)^2\right] = E\left[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right] \\ &= tr\left\{E\left[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\right]\right\} \\ &= tr\left\{\text{var}(\widehat{\boldsymbol{\beta}})\right\} \\ &= tr\left\{\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right\} \\ &= tr\left\{\sigma^2\mathbf{R}_{xx}^{-1}\right\} \\ &= \sigma^2 tr\left(\mathbf{R}_{xx}^{-1}\right). \end{aligned} \tag{5.22}$$

Since the sum of the eigenvalues of a matrix is equal to its trace, the average squared distance can be expressed in terms of the eigenvalues. Also the eigenvalues of $\mathbf{R}_{xx}^{-1}$ are the reciprocals of the eigenvalues of $\mathbf{R}_{xx}$. So equation (5.22) becomes

$$D^2 = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}. \tag{5.23}$$

One or more small eigenvalues can make the average squared distance very large, so small eigenvalues indicate collinearity. For orthogonal regressors, the λ_j's are all equal to 1 and hence $D^2 = p\sigma^2$. Hoerl and Kennard (1970) used this squared distance as a measure of multicollinearity.

Equation (5.22) can also be written as

$$D^2 = \sigma^2 \sum_{j=1}^{p} \text{VIF}_j \tag{5.24}$$

as the diagonal elements of the inverse of the correlation matrix of the regressors are the VIFs of the regressors. If the regressors are orthogonal then VIFs would be all 1 and consequently D^2 would be $p\sigma^2$. If D^2 is five times (say) greater than $p\sigma^2$, multicollinearity is present (Chatterjee and Hadi, 2006, p.245). The

average squared distance between $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$ for the body fat data is $1377.79\sigma^2$ and $p\sigma^2 = 3\sigma^2$, again suggesting a multicollinearity in the data.

5.3 Identifying collinear sets

The methods discussed above can provide only the presence and the number of collinearities among the columns of $\mathbf{X}$. However, they cannot identify the cause of each collinear sets. The methods discussed in this section aim to identify the variables involved in each collinearity.

There are three procedures for identifying the collinear sets. The first is based on the eigenvalues and the eigenvectors of the correlation matrix $\mathbf{R}_{xx}$. The second is also based on the eigenvalues and the eigenvectors of the correlation matrix, but is more sophisticated. It decomposes the variance of the regression coefficients and finally express them in proportions. The last and most recent procedure was suggested by Garthwaite et al. (2012) and is based on a transformation that partitions the VIFs into contribution of individual variables. The procedure can be used for identifying the collinear variables as well as the collinearity sets.

5.3.1 Eigenvalues and eigenvectors of $\mathbf{R}_{xx}$

It has been suggested that the eigenvectors corresponding to small eigenvalues can be used to identify the multicollinearity sets (e,g., Gunst and Mason (1977)). If $\boldsymbol{v}_j$ is the normalized eigenvector corresponding to the eigenvalue λ_j for the matrix $\mathbf{R}_{xx} = \mathbf{X}^\top\mathbf{X}$, then

$$\mathbf{X}^\top\mathbf{X}\boldsymbol{v}_j = \lambda_j\boldsymbol{v}_j. \tag{5.25}$$

Since $\boldsymbol{v}_j$ is normalized, this implies

$$(\mathbf{X}\boldsymbol{v}_j)^\top(\mathbf{X}\boldsymbol{v}_j) = \lambda_j \approx 0, \tag{5.26}$$

if λ_j is near zero. Now $(\mathbf{X}\boldsymbol{v}_j)^\top(\mathbf{X}\boldsymbol{v}_j) \approx 0$, if and only if $\mathbf{X}\boldsymbol{v}_j \approx \mathbf{0}$, i.e., $\sum_{i=1}^{p} v_{ij}\boldsymbol{x}_i \approx \mathbf{0}$, since the sum of the squares of the elements of near-zero vector is close to zero. This implies that the columns of $\mathbf{X}$ are linearly dependent. Since $\sum_{i=1}^{p} v_{ij}\boldsymbol{x}_i \approx \mathbf{0}$ is the definition of multicollinearity, small eigenvalues can identify collinearity among the regressors and collinearity sets can be described by the eigenvectors corresponding to small eigenvalues. If λ_j is near-zero, then the elements of $\boldsymbol{v}_j$ that are large identify the regressors that are involved in a collinearity.

This method is very simple and hence widely used in practice. However, combining VIFs with eigenvalues and eigenvectors is not a well-integrated approach, as a VIF is not linked to a particular eigenvector (VIFs are commonly used to determine the variables which are involved in collinearity). Garthwaite et al. (2012) illustrated that this method provides limited information about the pattern of collinearities in the data as a collinear set can be identified from only one eigenvector. Also, if two small eigenvectors are approximately equal then eigenvectors corresponding to those small eigenvalues cannot identify the form of the collinearities — it can only identify the variables that belong to at least one of the collinearities. That is, the method fails to separate specific variables involved in specific collinearities. For example, two eigenvectors may identify that variables X_1, X_2, X_3 and X_4 are involved in at least one collinearity. However, in reality, there can be a collinearity between X_1 and X_2 and another between X_3 and X_4. In addition, Belsley (1991, p.36) gave an example in which the small eigenvector elements that corresponded to a small eigenvalue did not correspond to the

absence of those variables in a collinearity.

For simplicity, we will use the term eigenvectors analysis to indicate the eigenvalues and eigenvectors analysis.

5.3.2 Regression coefficient variance-decomposition

The procedure of variance-decomposition was proposed by Silvey (1969) and was reinterpreted and extended by Belsley et al. (1980). It is closely related to the concept of eigenvectors analysis, as the method uses eigenvalues and eigenvectors to form a set of variance-decomposition proportions. The basis for the analysis is to decompose the variances of the regression estimators into a sum of terms that are associated with singular values of $\mathbf{X}$. Like eigenvectors analysis, this method is a by-product of principal component analysis.

The variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$, the OLS estimators, is

$$\text{var}\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \tag{5.27}$$

where σ^2 is the common variance of the random error term ϵ of the regression model $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

Using the singular-value decomposition of $\mathbf{X}$ (discussed in Chapter 2), equation (5.27) can be written as

$$\text{var}\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} = \sigma^2 \left(\mathbf{V}\boldsymbol{\Delta}\mathbf{U}^\top \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^\top\right)^{-1} = \sigma^2 \left(\mathbf{V}\boldsymbol{\Delta}^{-2}\mathbf{V}^\top\right). \tag{5.28}$$

The jth diagonal element of this matrix is the variance of the regression coefficient $\widehat{\beta}_j$. That is,

$$\text{var}\left(\widehat{\beta}_j\right) = \sigma^2 \sum_{k=1}^{p} \frac{v_{jk}^2}{\lambda_k} \tag{5.29}$$

Table 5.3: Matrix of variance decomposition proportions

		Variance-decomposition Proportions			
		X_1	X_2	...	X_p
Associated with	$\sqrt{\lambda_1}$	π_{11}	π_{12}	...	π_{1p}
	$\sqrt{\lambda_2}$	π_{21}	π_{22}	...	π_{2p}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$\sqrt{\lambda_p}$	π_{p1}	π_{p2}	...	π_{pp}

where v_{jk} is the jth element of the kth eigenvector. Equation (5.29) decomposes $\text{var}(\widehat{\beta}_j)$ into p components, each of them related to a single eigenvalue λ_k. Other things being equal, a small λ_k (which indicates near linear dependencies in the data) can dramatically inflate $\text{var}(\widehat{\beta}_j)$ if the corresponding entry v_{jk} is not close to zero. Here, $\sum_{k=1}^{p} \frac{v_{jk}^2}{\lambda_k}$ is proportional to the variance of $\widehat{\beta}_j$. If $\mathbf{X}^\top\mathbf{X}$ is the correlation matrix of the regressors, then $\sum_{k=1}^{p} \frac{v_{jk}^2}{\lambda_k}$ is the jth VIF.

If the components are expressed as proportions for all $\text{var}(\widehat{\beta}_j)$ and the results are displayed in a table, it will help researchers identify the collinearity sets. Let

$$\eta_{jk} = \frac{v_{jk}^2}{\lambda_k} \text{ and } \eta_j = \sum_{k=1}^{p} \eta_{jk} \text{ for } j = 1, \ldots, p. \tag{5.30}$$

Then the (j,k)th variance-decomposition proportion, i.e., proportion of variance of $\widehat{\beta}_j$ associated with the the kth singular value is

$$\pi_{kj} = \frac{\eta_{jk}}{\eta_j}, \text{ for all } j, k = 1, \ldots, p. \tag{5.31}$$

Table 5.3 displays the variance decomposition proportions associated with all singular values. It can be noted that $\sum_{k=1}^{p} \pi_{kj} = 1$, for $j = 1, \ldots, p$. If the regressors are orthogonal then the variance decomposition proportions matrix is an identity matrix. A variance-decomposition proportion greater than 0.5 that corresponds to a small eigenvalue is a recommended criteria for whether a regressor

is part of that multicollinearity (Belsley et al., 1980, p.112; Montgomery et al., 2015, p.300).

The following three cases summarizes the experimental evidence of Belsley et al. (1980).

(a) *Only one near dependency.* If there is only one large condition index, indicating a single near dependency, it is possible to identify collinear set by examining the variance-decomposition proportions (Belsley, 1991, p.136). Variance-decomposition proportions of at least 0.5 for two or more regression coefficients associated with a high condition index indicates that these variables (variables associated with those regression coefficients) are involved in a collinearity.

(b) *Confounding with competing dependencies.* This occurs when two or more condition indices are large and have roughly the same magnitude. The 0.5 rule for identifying the variables involved in each collinearity from separate principal components must be modified. In this case, the variables whose aggregate proportions across these large condition indices (of roughly the same magnitude) are at least 0.5 are involved in at least one of the collinear sets (Belsley et al., 1980, p.154). The information on exactly which variables are involved in a specific competing dependency is lost. For example, if there are two competing dependencies, one between X_1, X_2 and X_3 and another between X_2, X_3 and X_4, then we can only identify that the variables X_1, X_2, X_3 and X_4 are involved in at least one near dependency.

(c) *Dominating dependencies.* Dominating dependencies means the existence of two or more large condition indices with one extremely large. The near dependency associated with the extremely large condition index can explain a

greater amount of the variance of a specific regression coefficient and the involvement of this variable in other near linear dependencies becomes invisible (Belsley et al., 1980, p.155). In this case, it is quite possible to have only one variance-decomposition proportion (associated with the large condition index) that is greater than the threshold value of 0.5 and which masks the other collinear set. So, further analysis is required. One suggestion is to use the set of auxiliary regressions to investigate the nature of the relationship between variables. A procedure for forming the auxiliary regressions from the variance-decomposition proportions is give in Belsley et al. (1980, p.159). One variable is identified from each row (associated with a large condition index) as having the largest variance-decomposition proportions and each of these variables is regressed separately against the remaining variables. For example, suppose we have five variables X_1, X_2, X_3, X_4 and X_5 and it is known that there are two collinear sets, one in which X_1 has the largest variance proportions associated with a largest condition index, and the other in which X_2 has the largest variance proportions associated with the second largest condition index. Then X_1 and X_2 are each regressed separately on X_3, X_4 and X_5 and the t-statistics for regression coefficients are used descriptively to identify the variables involved in each collinearity (Belsley, 1991, p.144). Unfortunately, when there are many collinearities, forming auxiliary regressions is often not straightforward and interpreting the results can be hard.

An advantage of this approach over eigenvectors analysis is that the algorithms for computing the singular-value decomposition are more stable numerically than those for the eigenvalues and the eigenvectors of $\mathbf{R}_{xx}$ (Belsley, 1991, p.44). The

R package '**perturb**' can be used to find the variance-decomposition proportions from regression analysis. This is also available in SPSS and SAS and, in consequence, researchers have used this method widely for identifying collinear sets.

The procedure is simple to interpret if there is only one near dependency. However, for more than one collinearity, special training is needed to interpret the results and inconclusive results are quite common (Freund et al., 2006, p.198). Also, the steps of the computations are somewhat complicated and difficult to understand by non-mathematicians.

5.3.3 Cos-max and cos-square transformations

Involvement of a variable in a collinear relationship can commonly be identified by its VIF. The other two methods of identifying collinear sets have no direct relationship with VIFs. The methods suggested by Garthwaite et al. (2012) have one-to-one relationship with VIFs. These methods are the cos-max and the cos-square transformations, described in Section 2.2. To recap, suppose $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ is a set of $n \times 1$ observed vectors of the variables $X_1, \ldots, X_p$ and let $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$. Garthwaite et al. (2012) suggested two methods for obtaining the orthonormal components $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ that have a one-to-one correspondence with the original vectors, i.e., each component is closely related to a single X variable and each X variable is related to a single component. Suppose the $n \times p$ matrix $\mathbf{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p)$ must be chosen to maximize either

$$\psi = \sum_{j=1}^{p} \boldsymbol{x}_j^\top \boldsymbol{z}_j \quad \text{or} \quad \phi = \sum_{j=1}^{p} \left(\boldsymbol{x}_j^\top \boldsymbol{z}_j\right)^2 \tag{5.32}$$

under the conditions that $\mathbf{Z}$ is a column orthogonal matrix and $\boldsymbol{x}_j^\top \boldsymbol{z}_j > 0$ for $j = 1, \ldots, p$.

Garthwaite et al. (2012) showed that the transformation of $\mathbf{X} \rightarrow \mathbf{Z}$ is linear and is of the form $\mathbf{Z} = \mathbf{XA}$. Transformation of $\mathbf{X} \rightarrow \mathbf{Z}$ is called the cos-max transformation when ψ is maximized and when ϕ is maximized it is called the cos-square transformation. The transformation matrix is $\mathbf{A} = (\mathbf{X}^\top\mathbf{X})^{-1/2}$ for the cos-max transformation and $\mathbf{A} = \mathbf{C}(\mathbf{C}\mathbf{X}^\top\mathbf{X}\mathbf{C})^{-1/2}$ for the cos-square transformation, where $\mathbf{C}$ is a diagonal matrix with positive diagonal elements $c_1, \ldots, c_p$. The matrix $\mathbf{C}$ in the cos-square transformation cannot the obtained using a simple formula — rather, $\mathbf{C}$ is obtained using an iterative algorithm proposed by Garthwaite et al. (2012). In both cases, if each X is standardized to have a mean of 0 and unit length, i.e., $\boldsymbol{x}_j^\top\boldsymbol{x}_j = 1$, $\mathbf{A}$ may be used both as a diagnostic for determining the collinear variables and for identifying the variables that contribute to each collinearity.

For both the cos-max and the cos-square transformations, the new orthogonal vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_p$ are obtained in such a way that $\boldsymbol{z}_j$ is meant to be strongly related with only $\boldsymbol{x}_j$. So the diagonal elements of $\mathbf{A}$ should be high and the off-diagonal elements should be close to 0. However, if some of the X variables are collinear then the correlation between $\boldsymbol{x}_j$ and $\boldsymbol{z}_j$ will not be very strong for some j. Also, $\boldsymbol{x}_j$ might not be the only column of $\mathbf{X}$ that has marked correlation with $\boldsymbol{z}_j$. Consequently, some of the off-diagonal elements of $\mathbf{A}$ will not be close to 0. Hence, large off-diagonal elements indicate collinearity among the columns of $\mathbf{X}$.

Suppose each columns of $\mathbf{X}$ have been standardized to have a mean of 0 and unit length such that $\mathbf{X}^\top\mathbf{X}$ is the correlation matrix $\mathbf{R}_{xx}$. The diagonal elements of $\mathbf{R}_{xx}^{-1}$ are the VIFs (Farrar and Glauber, 1967) and VIFs are commonly used to identify the variables involved in collinear relations. It can be easily verified that

$\mathbf{A}\mathbf{A}^\top = \mathbf{R}_{xx}^{-1}$ for both the cos-max and cos-square transformations. Hence, if $\boldsymbol{a}_j^\top$ is the jth row of $\mathbf{A}$, then $\boldsymbol{a}_j^\top \boldsymbol{a}_j$ is the jth diagonal element of $\mathbf{R}_{xx}^{-1}$ and hence it is the VIF_j. So both transformations have a one-to-one relationship with VIFs. If collinearity exists between some of the X variables, then some elements of $\boldsymbol{a}_j$ will be large, including the jth element, and the remaining elements will be close to 0. Consequently, VIF_j will be large. Moreover, the larger elements of $\boldsymbol{a}_j$ that corresponds to a large VIF_j define the collinearity among the variables.

This method is simple and uses more information in identifying collinear sets. For example, if a collinearity exists between X_1, X_2 and X_3 then each of the three rows $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ and $\boldsymbol{a}_3$ will provide this information. Also, each row of $\mathbf{A}$ has a one-to-one relationship with a particular VIF and hence the elements of the row determines that VIF. Moreover, this method can separate the confounded collinearities while the other two methods can only identify which variables are involved in at least one of the confounded collinearities.

5.4 Illustrative Examples

To illustrate the detection and pattern of collinearity, we apply all three procedures described in Section 5.3, together with VIFs and condition indices, to three example datasets. Each dataset is extracted from published studies where they suggest what collinearity patterns are present in the data. We examine the collinearity patterns that are identified by three procedures and make comparisons.

Table 5.4: Correlation matrix and VIF of regressors for the sales data

Variable	AS_t	AE_t	PE_t	SE_t	AE_{t-1}	PE_{t-1}	VIF
AS_t	1.000	-0.170	0.540	0.811	-0.350	-0.052	
AE_t	-0.170	1.000	-0.357	-0.129	-0.140	-0.496	36.94
PE_t	0.540	-0.357	1.000	0.063	-0.316	-0.296	33.47
SE_t	0.811	-0.129	0.063	1.000	-0.166	0.208	1.08
AE_{t-1}	-0.305	-0.140	-0.316	-0.166	1.000	-0.358	25.92
PE_{t-1}	-0.052	-0.496	-0.296	0.208	-0.358	1.000	43.52

5.4.1 Sales of a firm

Data reported in Chatterjee and Hadi (2006, p.236) were collected from a firm for a period of 23 years. The firm had fairly stable operating conditions during the period of data collection. The objective is to regress aggregate sales (AS_t) against five regressors: advertising expenditures (AE_t), promotion expenditures (PE_t), sales expense (SE_t), (lagged) advertising expenditure in the previous year (AE_{t-1}) and lagged promotion expenditure (PE_{t-1}).

Table 5.4 displays the correlations among the variables in the dataset and VIFs of the regressors. The correlations among the regressors are small. However, the VIFs corresponding to the regressors AE_t, PE_t, AE_{t-1}, and PE_{t-1} are high indicating that they are involved in a collinearity or collinearities. Only the regressor SE_t is not involved in any collinearity.

VIFs can only identify the regressors which are involved in the collinearity. For identifying collinear sets, we need to apply either the eigenvectors analysis of the correlation matrix, or the variance-decomposition proportion method, or the transformation matrix of either the cos-max or the cos-square transformation.

The dataset is analysed by Chatterjee and Hadi (2006) using the eigenvectors analysis. Table 5.5 gives the eigenvalues and eigenvectors of the correlation matrix $\mathbf{R}_{xx}$ along with the condition indices. Among the five eigenvalues, $\lambda_5 = 0.007$ is

Table 5.5: Eigenvectors analysis for the sales data

Number	Eigenvalue	Condition Index	Eigenvector				
			$\boldsymbol{v}_1$	$\boldsymbol{v}_2$	$\boldsymbol{v}_3$	$\boldsymbol{v}_4$	$\boldsymbol{v}_5$
1	1.701	1.000	0.532	0.024	0.668	-0.074	0.514
2	1.288	1.149	-0.232	-0.825	-0.158	0.037	0.489
3	1.145	1.219	-0.389	0.022	0.217	-0.895	-0.010
4	0.859	1.407	0.395	0.260	-0.692	-0.338	0.428
5	0.007	15.295	-0.596	0.501	0.057	0.279	0.559

Table 5.6: Variance-decomposition proportions for the sales data

Principal Component	Eigenvalue	Variance Proportion				
		AE_t	PE_t	SE_t	AE_{t-1}	PE_{t-1}
1	1.701	0.005	0.001	0.083	0.004	0.005
2	1.288	0.000	0.016	0.000	0.002	0.004
3	1.145	0.011	0.001	0.038	0.016	0.000
4	0.859	0.000	0.000	0.867	0.005	0.002
5	0.007	0.985	0.983	0.012	0.973	0.989

the smallest. Furthermore, the condition index $\kappa_5 = 15.295$ is well above the critical point of 10 and the others are very small. (We considered the rule of thumb of Rawlings (1988).) So there is only one collinearity in the dataset. Large entries in the 5th eigenvector indicate that the regressors AE_t, PE_t, AE_{t-1} and PE_{t-1} constitutes a collinear set. This was the collinearity pattern identified by Chatterjee and Hadi (2006).

Next we have applied the variance-decomposition proportion method. Table 5.6 presents the results obtained using the variance-decomposition proportion method. The entries in the fifth column (SE_t) indicate that the fourth principal component can explain around 87% of the variance of $\widehat{\beta}_3$, 8% comes from the first principal component and so forth. According to the rule of thumb suggested by Belsley et al. (1980), a collinearity is indicated when these variance-decomposition proportions greater than 0.5 for two or more regression coefficients associated with a small eigenvalue. Under this criterion, the variables AE_t, PE_t, AE_{t-1} and PE_{t-1}

Table 5.7: Cos-max transformation matrix and VIF for the sales data

	AE_t	PE_t	SE_t	AE_{t-1}	PE_{t-1}	VIF
$\boldsymbol{a}_1^\top$	3.743	2.736	-0.010	2.345	3.154	36.94
$\boldsymbol{a}_2^\top$	2.736	3.470	-0.070	2.285	2.952	33.47
$\boldsymbol{a}_3^\top$	-0.010	-0.070	1.026	0.024	-0.134	1.08
$\boldsymbol{a}_4^\top$	2.345	2.285	0.024	2.901	2.604	25.92
$\boldsymbol{a}_5^\top$	3.154	2.952	-0.134	2.604	4.249	43.52

Table 5.8: Cos-square transformation matrix and VIF for the sales data

	AE_t	PE_t	SE_t	AE_{t-1}	PE_{t-1}	VIF
$\boldsymbol{a}_1^\top$	3.736	2.676	-0.004	2.232	3.293	36.94
$\boldsymbol{a}_2^\top$	2.735	3.419	-0.062	2.181	3.089	33.47
$\boldsymbol{a}_3^\top$	-0.005	-0.069	1.025	0.031	-0.142	1.08
$\boldsymbol{a}_4^\top$	2.348	2.245	0.029	2.813	2.729	25.92
$\boldsymbol{a}_5^\top$	3.136	2.880	-0.120	2.472	4.389	43.52

are involved in a collinearity, based on the decomposition for the 5th eigenvalue. That is, the variance-decomposition proportion suggests the same collinearity as that given by the eigenvectors analysis.

Lastly, we applied the transformation matrices of both the cos-max and the cos-square transformations to standardized $\mathbf{X}$. Tables 5.7 and 5.8 present the cos-max and the cos-square transformation matrices, respectively. The transformation matrix of the cos-max transformation is symmetric while the cos-square transformation matrix is asymmetric. However, the entries of both transformation matrices are very similar. Large entries in $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, $\boldsymbol{a}_4$ and $\boldsymbol{a}_5$ for both the cos-max and the cos-square transformations indicate that the regressors AE_t, PE_t, AE_{t-1}, and PE_{t-1} are involved in a collinearity. That is, with each transformation, four rows of the transformation matrices provide information about a collinearity involving four variables.

The collinear set identified for this simple example is the same with all methods. In the eigenvectors analysis, the decision that there is a collinearity between the

variables AE_t, PE_t, AE_{t-1}, and PE_{t-1} comes from the last eigenvector. Also the same collinear set is identified from a single row that is associated with a single principal component in the variance-decomposition proportion method. However, the four rows of either the cos-max or the cos-square transformation matrix provide information about a collinearity between AE_t, PE_t, AE_{t-1}, and PE_{t-1}. So even though all three methods identify the same collinear set, the cos-max and the cos-square transformations provide more information than the other two methods.

5.4.2 Pitprop data

Data used by Jeffers (1967) were collected from East Anglia over a period of 10 years to determine the physical characteristics of pitprops made of Corsican pine that influence their maximum compressive strength. The study has 180 pitprops. The physical variables on each prop were top diameter in inches (X_1), length in inches (X_2), moisture content as a percentage of the dry weight (X_3), specific gravity at the time of the test (X_4), oven-dry specific gravity (X_5), number of annual rings at the top (X_6), number of annual rings at the base (X_7), maximum bow in inches (X_8), distance of the point of maximum bow from the top in inches (X_9), number of knot whorls (X_{10}), length of clear prop from the top in inches (X_{11}), average number of knots per whorl (X_{12}), and average diameter of the knots in inches (X_{13}). The dataset was used in Garthwaite et al. (2012) to illustrate collinearity identification using the cos-max and the cos-square transformations and compare the methods with the eigenvectors analysis. Here, we also used the variance-decomposition proportion method and compare all three methods.

Table 5.9 presents the sample correlation matrix of the 13 physical variables for

Table 5.9: Correlation matrix for the physical properties of pitprops

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}
X_1	1.000	0.954	0.364	0.342	-0.129	0.313	0.496	0.424	0.592	0.545	0.084	-0.019	0.134
X_2	0.954	1.000	0.297	0.284	-0.118	0.291	0.503	0.419	0.648	0.569	0.076	-0.036	0.144
X_3	0.364	0.297	1.000	0.882	-0.148	0.153	-0.029	-0.054	0.125	-0.081	0.162	0.220	0.126
X_4	0.342	0.284	0.882	1.000	0.220	0.381	0.174	-0.059	0.137	-0.014	0.097	0.169	0.015
X_5	-0.129	-0.118	-0.148	0.220	1.000	0.364	0.296	0.004	-0.039	0.037	-0.091	-0.145	-0.208
X_6	0.313	0.291	0.153	0.381	0.364	1.000	0.813	0.090	0.211	0.274	-0.036	0.024	-0.329
X_7	0.496	0.503	-0.029	0.174	0.296	0.813	1.000	0.372	0.465	0.679	-0.113	-0.232	-0.424
X_8	0.424	0.419	-0.054	-0.059	0.004	0.090	0.372	1.000	0.482	0.557	0.061	-0.357	-0.202
X_9	0.592	0.648	0.125	0.137	-0.039	0.211	0.465	0.482	1.000	0.526	0.085	-0.127	-0.076
X_{10}	0.545	0.569	-0.081	-0.014	0.037	0.274	0.679	0.557	0.526	1.000	-0.319	-0.368	-0.291
X_{11}	0.084	0.076	0.162	0.097	-0.091	-0.036	-0.113	0.061	0.085	-0.319	1.000	0.029	0.007
X_{12}	-0.019	-0.036	0.220	0.169	-0.145	0.024	-0.232	-0.357	-0.127	-0.368	0.029	1.000	0.184
X_{13}	0.134	0.144	0.126	0.015	-0.208	-0.329	-0.424	-0.202	-0.076	-0.291	0.007	0.184	1.000

Table 5.10: Eigenvectors analysis for the pitprop data

Eigenvector	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}
$\boldsymbol{v}_{11}$	0.005	0.054	-0.117	0.017	-0.005	0.537	-0.764	-0.026	0.051	0.318	0.048	-0.047	-0.045
$\boldsymbol{v}_{12}$	0.392	-0.411	0.527	-0.585	0.202	0.080	-0.036	-0.053	0.054	0.060	0.005	0.002	0.013
$\boldsymbol{v}_{13}$	0.572	-0.582	-0.408	0.383	-0.118	-0.057	-0.002	-0.018	0.058	-0.004	0.007	-0.004	0.009
VIF	13.135	13.714	11.660	12.420	2.533	6.932	12.033	1.852	2.103	5.118	1.511	1.434	1.771

the pitprop data. The correlations between X_1 and X_2, between X_3 and X_4, and between X_6 and X_7 are strong. There are also a number of moderate correlations. High pairwise correlations indicate that there may exist collinearities between some of the physical variables.

The correlation matrix has the eigenvalues 4.219, 2.378, 1.878, 1.109, 0.910, 0.815, 0.576, 0.440, 0.353, 0.191, 0.051, 0.041 and 0.039. Three of these are small compared to the others. This suggests that there are three collinear sets in the dataset. Table 5.10 displays the eigenvectors corresponding to the three small eigenvalues along with the VIFs. A VIF greater than 5 or 10 is taken as an indication of collinearity, according to rule of thumb given by Montgomery et al. (2015). Under this rule, the variables X_1, X_2, X_3, X_4, X_6, X_7 and X_{10} are involved

Table 5.11: Variance-decomposition proportion for the pitprop data

Principal	Condition	Variance Proportion												
Component	Index	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}
1	1.000	0.003	0.003	0.000	0.001	0.000	0.003	0.003	0.011	0.014	0.007	0.000	0.002	0.002
2	1.332	0.002	0.001	0.011	0.007	0.005	0.000	0.001	0.008	0.000	0.005	0.012	0.035	0.023
3	1.499	0.002	0.002	0.001	0.005	0.049	0.017	0.003	0.017	0.011	0.001	0.002	0.003	0.032
4	1.950	0.001	0.001	0.000	0.000	0.001	0.001	0.000	0.040	0.004	0.007	0.385	0.057	0.047
5	2.153	0.001	0.001	0.012	0.011	0.013	0.016	0.004	0.020	0.006	0.005	0.086	0.276	0.004
6	2.275	0.001	0.002	0.008	0.000	0.189	0.000	0.000	0.002	0.001	0.007	0.025	0.025	0.272
7	2.705	0.002	0.001	0.000	0.001	0.122	0.022	0.008	0.149	0.133	0.000	0.023	0.351	0.024
8	3.098	0.003	0.000	0.000	0.000	0.000	0.007	0.000	0.506	0.529	0.000	0.000	0.070	0.017
9	3.459	0.024	0.022	0.001	0.000	0.088	0.069	0.004	0.183	0.195	0.040	0.042	0.013	0.236
10	4.702	0.038	0.028	0.002	0.004	0.000	0.008	0.015	0.016	0.008	0.518	0.394	0.138	0.318
11	9.134	0.000	0.004	0.023	0.000	0.000	0.823	0.959	0.007	0.024	0.391	0.031	0.031	0.022
12	10.086	0.282	0.297	0.574	0.665	0.390	0.022	0.003	0.036	0.033	0.017	0.000	0.000	0.002
13	10.437	0.643	0.638	0.368	0.305	0.143	0.012	0.000	0.005	0.041	0.000	0.001	0.000	0.001

in at least one collinearity. The eigenvectors $\boldsymbol{v}_{13}$ and $\boldsymbol{v}_{12}$, which are associated with the two smallest eigenvalues, both suggest a collinearity between X_1, X_2, X_3 and X_4. According to $\boldsymbol{v}_{11}$, which corresponds to the 3rd smallest eigenvalue, the variables X_6 and X_7 are clearly collinear; perhaps the variable X_{10} could be involved in this latter collinearity.

Table 5.11 presents the results obtained using the variance-decomposition proportion analysis. The second column indicates that there are three weak collinearities (κ_j's around 10). Since the last two condition indices are roughly equal then the two collinear sets are confounded. The 0.5 rule of thumb for identifying the collinear sets must be modified. According to the sum rule discussed in Subsection 5.3.2, last two rows indicate that the variables X_1, X_2, X_3, X_4 and X_5 are involved in at least one of the near linear dependencies. In this case, auxiliary regressions can be used to identify the separate involvement of a particular variable in a particular collinearity. However, forming auxiliary regressions has some troubles and may not be informative in all situations. While the third row from the bottom

Table 5.12: Cos-max transformation matrix and VIF for the pitprop data

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	VIF
$\boldsymbol{a}_1^\top$	3.036	-1.912	-0.252	-0.100	0.119	-0.064	-0.220	-0.202	0.006	-0.204	-0.079	-0.068	-0.175	13.135
$\boldsymbol{a}_2^\top$	-1.912	3.114	0.004	-0.051	0.048	0.032	-0.220	0.004	-0.382	-0.292	-0.103	-0.059	-0.255	13.714
$\boldsymbol{a}_3^\top$	-0.252	0.004	2.717	-1.946	0.592	0.017	0.259	-0.027	-0.015	0.077	-0.053	-0.028	0.022	11.660
$\boldsymbol{a}_4^\top$	-0.100	-0.051	-1.946	2.838	-0.664	-0.333	0.020	0.116	-0.032	-0.015	0.007	-0.041	0.011	12.420
$\boldsymbol{a}_5^\top$	0.119	0.048	0.592	-0.664	1.296	-0.108	-0.128	-0.040	0.018	0.071	0.036	0.086	0.017	2.533
$\boldsymbol{a}_6^\top$	-0.064	0.032	0.017	-0.333	-0.108	2.102	-1.479	0.043	0.098	0.394	0.037	-0.154	0.069	6.932
$\boldsymbol{a}_7^\top$	-0.220	-0.220	0.259	0.020	-0.128	-1.479	2.979	0.004	-0.211	-0.788	-0.023	0.170	0.305	12.033
$\boldsymbol{a}_8^\top$	-0.202	0.004	-0.027	0.116	-0.040	0.043	0.004	1.288	-0.171	-0.254	-0.095	0.149	0.094	1.852
$\boldsymbol{a}_9^\top$	0.006	-0.382	-0.015	-0.032	0.018	0.098	-0.211	-0.171	1.360	-0.124	-0.077	-0.021	0.017	2.103
$\boldsymbol{a}_{10}^\top$	-0.204	-0.292	0.077	-0.015	0.071	0.394	-0.788	-0.254	-0.124	1.969	0.411	0.191	0.207	5.118
$\boldsymbol{a}_{11}^\top$	-0.079	-0.103	-0.053	0.007	0.036	0.037	-0.023	-0.095	-0.077	0.411	1.137	0.056	0.097	1.511
$\boldsymbol{a}_{12}^\top$	-0.068	-0.059	-0.028	-0.041	0.086	-0.154	0.170	0.149	-0.021	0.191	0.056	1.141	0.000	1.434
$\boldsymbol{a}_{13}^\top$	-0.175	-0.255	0.022	0.011	0.017	0.069	0.305	0.094	0.017	0.207	0.097	0.000	1.231	1.771

suggests a collinearity between X_6 and X_7. Which coincide with one of the finding of eigenvectors analysis.

The transformation matrices of the cos-max and the cos-square transformations along with the VIFs are given in Tables 5.12 and 5.13, respectively. The last columns of both tables are identical as in each case $\boldsymbol{a}_j^\top \boldsymbol{a}_j = \text{VIF}_j$. The cos-max transformation matrix is symmetric while the cos-square transformation matrix is asymmetric. But the values in the two matrices are again very similar. Both the transformations indicate that there is a collinearity between X_1 and X_2. There is another collinearity between X_3 and X_4 and the entries in $\boldsymbol{a}_3$, $\boldsymbol{a}_4$ and $\boldsymbol{a}_5$ indicate that this collinearity might also include X_5. Garthwaite et al. (2012) suggested that since X_4 is the specific gravity at the time of the test and X_5 is the oven-dry specific gravity, they are likely to be approximately collinear and X_3, their moisture content. Clearly, X_7 is collinear with X_6 and the entry corresponding to X_{10} in $\boldsymbol{a}_7$ and the entry corresponding to X_7 in $\boldsymbol{a}_{10}$ indicate that X_7 is collinear with X_{10}.

Table 5.13: Cos-square transformation matrix and VIF for the pitprop data

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	VIF
$\boldsymbol{a}_1^\top$	3.039	-1.916	-0.230	-0.093	0.097	-0.070	-0.222	-0.173	-0.010	-0.206	-0.066	-0.056	-0.152	13.135
$\boldsymbol{a}_2^\top$	-1.882	3.144	0.004	-0.049	0.041	0.013	-0.216	-0.003	-0.326	-0.281	-0.083	-0.047	-0.209	13.714
$\boldsymbol{a}_3^\top$	-0.262	0.005	2.694	-1.996	0.510	0.022	0.276	-0.024	-0.015	0.082	-0.049	-0.030	0.012	11.660
$\boldsymbol{a}_4^\top$	-0.102	-0.055	-1.916	2.881	-0.568	-0.310	0.032	0.103	-0.033	-0.022	0.003	-0.039	0.007	12.420
$\boldsymbol{a}_5^\top$	0.131	0.056	0.605	-0.702	1.265	-0.133	-0.133	-0.035	0.019	0.075	0.037	0.087	0.0267	2.533
$\boldsymbol{a}_6^\top$	-0.085	0.016	0.023	-0.344	-0.120	1.989	-1.636	0.041	0.076	0.358	0.031	-0.132	0.090	6.932
$\boldsymbol{a}_7^\top$	-0.207	-0.205	0.225	0.027	-0.092	-1.255	3.115	0.006	-0.175	-0.711	-0.018	0.131	0.229	12.033
$\boldsymbol{a}_8^\top$	-0.222	-0.003	-0.027	0.121	-0.033	0.043	0.008	1.281	-0.179	-0.273	-0.087	0.148	0.091	1.852
$\boldsymbol{a}_9^\top$	-0.013	-0.419	-0.016	-0.038	0.018	0.079	-0.237	-0.175	1.343	-0.148	-0.073	-0.018	0.014	2.103
$\boldsymbol{a}_{10}^\top$	-0.229	-0.318	0.081	-0.022	0.062	0.328	-0.850	-0.236	-0.130	1.964	0.357	0.173	0.190	5.118
$\boldsymbol{a}_{11}^\top$	-0.091	-0.117	-0.059	0.003	0.038	0.036	-0.027	-0.094	-0.080	0.443	1.123	0.050	0.088	1.511
$\boldsymbol{a}_{12}^\top$	-0.076	-0.065	-0.035	-0.048	0.087	-0.148	0.191	0.156	-0.019	0.211	0.049	1.133	-0.011	1.434
$\boldsymbol{a}_{13}^\top$	-0.204	-0.286	0.014	0.008	0.026	0.100	0.329	0.095	0.015	0.229	0.086	-0.010	1.208	1.771

The superiority of the cos-max and the cos-square transformations matrices over the eigenvectors analysis is that the eigenvectors cannot suggest separate collinearities, one between X_1 and X_2 and another between X_3 and X_4, while the cos-max and the cos-square transformations can identify separate collinearities. Similarly, the variance-decomposition proportion method can only identify that the variables X_1, X_2, X_3, X_4 and X_5 are involved in at least one of the collinearities. However, if we use the auxiliary regressions based on the values of the variance-decomposition proportions of last two rows, we may get the same conclusion as we get from the cos-max and the cos-square transformations. But, that is an extra bit of work to identify collinear sets. Also, with the cos-max and the cos-square transformations, the $\boldsymbol{a}_3$, $\boldsymbol{a}_4$ and $\boldsymbol{a}_5$ rows provide detailed information about the collinearity between X_3, X_4 and X_5.

5.4.3 Shopping pattern data

Mahajan et al. (1977) collected data from the residents of an inner-city neighbor-

Table 5.14: Correlation matrix among the variables of the shopping pattern data

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1.000	0.547	0.274	0.637	0.481	0.517	0.369	0.242	0.566	0.666
X_2	0.547	1.000	0.650	0.837	0.809	0.792	0.562	0.277	0.808	0.768
X_3	0.274	0.650	1.000	0.744	0.782	0.795	0.781	0.496	0.566	0.526
X_4	0.637	0.837	0.744	1.000	0.851	0.772	0.609	0.609	0.815	0.775
X_5	0.481	0.809	0.782	0.851	1.000	0.906	0.821	0.573	0.798	0.748
X_6	0.517	0.792	0.795	0.772	0.906	1.000	0.781	0.418	0.736	0.702
X_7	0.369	0.562	0.781	0.609	0.821	0.781	1.000	0.473	0.577	0.599
X_8	0.242	0.277	0.496	0.609	0.573	0.418	0.473	1.000	0.484	0.424
X_9	0.566	0.808	0.566	0.815	0.798	0.736	0.577	0.484	1.000	0.894
X_{10}	0.666	0.768	0.526	0.775	0.748	0.702	0.599	0.424	0.894	1.000

hood. Telephone interviews were conducted to collect data from the members of the household who did the major food shopping for the household. The households were selected by random sampling from households in a large northeastern metropolitan city. There were 10 regressors in their study.

Table 5.14 gives the correlation matrix of the regressors. There are a number of strong correlations, such as between X_2 and X_4, between X_5 and X_6 and between X_9 and X_{10}, as well as a number of moderate correlations among the regressors. Large pairwise correlation is a sufficient condition for collinearity, so we expect collinearities between some of the regressors.

Ofir and Khuri (1986) analyzed this dataset using VIFs, R_j^2's (R_j^2 is the R^2 value when regressing X_j on the remaining regressors) and the eigenvalues and eigenvectors of the correlation matrix $\mathbf{R}_{xx}$ of the regressors.

Table 5.15 represents the results of collinearity diagnostic using the eigenvectors analysis. The expected squared distance between the least square estimates and the true parameter is $79.9\sigma^2$, which indicates that the least square estimates are about eight times inflated by multicollinearity (for orthogonal regressors the expected distance would be $p\sigma^2$, i.e., $10\sigma^2$). This is one indication of collinearity

Table 5.15: Eigenvectors analysis for the shopping pattern data

Eigenvector	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
$\boldsymbol{v}_8$	0.110	-0.247	0.035	0.151	0.041	-0.186	0.146	-0.178	0.683	-0.589
$\boldsymbol{v}_9$	-0.126	-0.525	-0.073	0.478	0.526	-0.087	-0.151	-0.294	-0.185	0.216
$\boldsymbol{v}_{10}$	0.213	0.070	0.337	-0.554	0.566	-0.356	-0.280	0.039	0.037	0.001
VIF	3.330	7.526	6.783	14.641	15.831	8.832	5.875	2.947	7.488	6.643

Note: $\sum_{j=1}^{10} \frac{1}{\lambda_j} = 79.9$

in the dataset. The eigenvalues of the correlation matrix are 6.897, 1.059, 0.739, 0.462, 0.349, 0.188, 0.129, 0.081, 0.064 and 0.032. The corresponding condition indices are 1.000, 2.552, 3.055, 3.866, 4.446, 6.056, 7.319, 9.204, 10.392 and 14.716, respectively. The last three eigenvalues are small compared to the others as well as have three large condition indices, so there are three collinear sets in the data. The κ_j's indicate that one of them is moderate while the other two are weak. The VIFs indicate that all the regressors except X_1 and X_8 are involved in a collinearity. On the basis of $\boldsymbol{v}_{10}$, Ofir and Khuri (1986) concluded that the variables X_4 and X_5 are clearly collinear, and that perhaps X_6 is involved in this collinearity. Although Ofir and Khuri did not suggest it, the third element of the eigenvector indicates that X_3 might also be included in this collinear set. The other two eigenvectors (associated with next two smallest eigenvalues) identify one collinearity between X_2, X_4, and X_5 and another one between X_9 and X_{10}.

The variance-decomposition proportions of shopping pattern data is given in Table 5.16. The last row of the variance-decomposition proportions indicates that X_3, X_4 and X_5 are involved in one collinearity, and that perhaps X_6 is involved in this set. This collinear set is the same as the collinear set identified by the last eigenvector ($\boldsymbol{v}_{10}$). The second last row has only one large variance proportions (associated with X_2). Although the variance proportions corresponding to X_4 and

Table 5.16: Variance-decomposition proportion of shopping pattern data

Principal	Variance Proportion									
Component	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	0.003	0.002	0.002	0.001	0.001	0.002	0.002	0.002	0.002	0.002
2	0.092	0.006	0.024	0.000	0.001	0.001	0.020	0.047	0.006	0.014
3	0.012	0.014	0.010	0.003	0.000	0.012	0.010	0.303	0.002	0.002
4	0.303	0.039	0.001	0.002	0.000	0.003	0.073	0.000	0.045	0.005
5	0.041	0.028	0.053	0.040	0.001	0.000	0.105	0.001	0.042	0.093
6	0.000	0.013	0.271	0.007	0.067	0.142	0.010	0.038	0.002	0.117
7	0.003	0.205	0.097	0.024	0.018	0.327	0.256	0.000	0.057	0.015
8	0.044	0.100	0.002	0.019	0.001	0.048	0.044	0.131	0.766	0.641
9	0.074	0.573	0.012	0.245	0.274	0.013	0.061	0.460	0.071	0.110
10	0.428	0.021	0.527	0.659	0.636	0.451	0.419	0.016	0.006	0.000

Table 5.17: Cos-max transformation matrix and VIFs of the shopping pattern data

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	VIF
$\boldsymbol{a}_1^\top$	1.501	0.030	0.410	-0.712	0.282	-0.348	-0.177	0.095	0.080	-0.395	3.330
$\boldsymbol{a}_2^\top$	0.030	2.428	-0.125	-0.843	-0.506	-0.336	0.170	0.533	-0.392	-0.267	7.526
$\boldsymbol{a}_3^\top$	0.410	-0.125	2.179	-0.921	0.241	-0.652	-0.695	-0.071	0.165	0.078	6.783
$\boldsymbol{a}_4^\top$	-0.712	-0.843	-0.921	3.330	-0.794	0.316	0.502	-0.622	-0.330	-0.104	14.641
$\boldsymbol{a}_5^\top$	0.282	-0.506	0.241	-0.794	3.548	-1.085	-0.914	-0.353	-0.286	-0.030	15.831
$\boldsymbol{a}_6^\top$	-0.348	-0.336	-0.652	0.316	-1.085	2.611	-0.160	0.102	-0.215	0.006	8.832
$\boldsymbol{a}_7^\top$	-0.177	0.170	-0.695	0.502	-0.914	-0.160	2.030	-0.093	0.066	-0.289	5.875
$\boldsymbol{a}_8^\top$	0.095	0.533	-0.071	-0.622	-0.353	0.102	-0.093	1.443	-0.182	-0.039	2.947
$\boldsymbol{a}_9^\top$	0.080	-0.392	0.165	-0.330	-0.286	-0.215	0.066	-0.182	2.459	-0.989	7.488
$\boldsymbol{a}_{10}^\top$	-0.395	-0.267	0.078	-0.104	-0.030	0.006	-0.289	-0.039	-0.989	2.310	6.643

X_5 are small they could perhaps be collinear with X_2. (The eigenvectors analysis suggested a collinearity between X_2, X_4 and X_5.) The variance proportions of X_4 and X_5 are dominated by the first collinear set. The third row from the bottom shows that clearly X_9 and X_{10} are collinear. This is also one of the finding from the eigenvectors analysis.

Tables 5.17 and 5.18 present the transformation matrices of the cos-max and the cos-square transformations of shopping pattern data, respectively. From the components of $\boldsymbol{a}_9$ and $\boldsymbol{a}_{10}$ in both the cos-max and the cos-square transformation matrices, clearly there is a collinearity between X_9 and X_{10}. This coincides with

Table 5.18: Cos-square transformation matrix and VIFs of the shopping pattern data

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	VIF
$\boldsymbol{a}_1^\top$	1.434	-0.006	0.380	-0.815	0.317	-0.370	-0.161	0.063	0.063	-0.441	3.330
$\boldsymbol{a}_2^\top$	-0.004	2.388	-0.162	-0.941	-0.578	-0.364	0.142	0.408	-0.425	-0.281	7.526
$\boldsymbol{a}_3^\top$	0.325	-0.173	2.127	-1.039	0.309	-0.683	-0.668	-0.094	0.157	0.071	6.783
$\boldsymbol{a}_4^\top$	-0.519	-0.750	-0.773	3.465	-0.810	0.308	0.406	-0.453	-0.284	-0.070	14.641
$\boldsymbol{a}_5^\top$	0.189	-0.431	0.215	-0.757	3.677	-0.936	-0.695	-0.242	-0.227	-0.007	15.831
$\boldsymbol{a}_6^\top$	-0.273	-0.338	-0.590	0.358	-1.163	2.593	-0.185	0.065	-0.224	0.012	8.832
$\boldsymbol{a}_7^\top$	-0.142	0.157	-0.690	0.563	-1.032	-0.220	1.954	-0.100	0.055	-0.306	5.875
$\boldsymbol{a}_8^\top$	0.065	0.522	-0.112	-0.726	-0.416	0.090	-0.116	1.372	-0.223	-0.058	2.947
$\boldsymbol{a}_9^\top$	0.048	-0.409	0.141	-0.342	-0.293	-0.232	0.048	-0.168	2.461	-0.978	7.488
$\boldsymbol{a}_{10}^\top$	-0.347	-0.277	0.065	-0.086	-0.009	0.013	-0.273	-0.045	-1.004	2.313	6.643

one of the results obtained from both the eigenvectors method and the variance-decomposition proportion method. Further information about the structure between X_2, X_3, X_4, X_5, X_6 and X_7 is provided by the components of $\boldsymbol{a}_2$, $\boldsymbol{a}_3$, $\boldsymbol{a}_4$, $\boldsymbol{a}_5$, $\boldsymbol{a}_6$ and $\boldsymbol{a}_7$. From $\boldsymbol{a}_2$, there is a suggestion that X_2 and X_4 are related. The components of $\boldsymbol{a}_3$ indicate a relationship between X_3 and X_4. The large values in $\boldsymbol{a}_4$ indicate that X_4 is related with X_2, X_3, and X_5. Similarly, $\boldsymbol{a}_5$ indicates that X_5 is related with X_4, X_6 and X_7. A relationship between X_6 and X_5 is indicated by $\boldsymbol{a}_6$, and a relationship between X_7 and X_5 is indicated by $\boldsymbol{a}_7$.

We only know of the existence of relationships between the variables but do not know their directions. By assigning directions to the relationships, we can generate some possible structures. From $\boldsymbol{a}_2$ and $\boldsymbol{a}_3$, we can say X_4 predicts X_2 and X_3 respectively. Similarly X_5 predicts X_6 and X_7 in $\boldsymbol{a}_6$ and $\boldsymbol{a}_7$ rows respectively. In addition to the directions assigned above, $\boldsymbol{a}_4$ and $\boldsymbol{a}_5$ can say X_5 predicts X_4. Figure 5.1(a) is drawn to explain the above mentioned directional relationship. Figure 5.1 shows four of these possible structures. This is one of the advantages of using the cos-max and the cos-square transformation matrices. Each structure can yield the same correlation pattern, for suitable choices of pairwise correlations.

The only difference between Figures 5.1(a) and 5.1(c) is that in 5.1(c) X_2 and X_3 predict X_4. Whereas the difference between Figures 5.1(a) and 5.1(d) is that both X_6 and X_7 predict X_5. Figure 5.1(b) assigns the reverse direction between X_4 and X_5 compared to Figure 5.1(d). This is the only difference between Figures 5.1(b) and 5.1(d). In Figure 5.1(a), there will be some correlation between X_2 and X_3 because of their dependence on X_4. This is why there must be a weak correlation between X_6 and X_7 in Figure 5.1(b). This illustrates that the cos-max and cos-square transformation matrices can suggest structures that might underlie the data. The structure between the variables cannot be inferred from the other two methods.

A simulation study was conducted to compare the methods using the structure given in Figure 5.1(a). We generated 1000 data considering the simple case where one variable is regressed by only one regressor. We generated X_5 from a normal distribution and then generated X_4, X_6 and X_7 from X_5. We then generated X_2 and X_3 from X_4. The beta coefficients and variances of the random error terms to generate the simulated data were estimated from the shopping pattern data. (We have no real data, so we first generated 1000 data from an MVN distribution using the correlation matrix of the shopping pattern data.) Table 5.19 gives the correlation matrix for the generated data. Using this matrix, we applied the above mentioned three methods of collinearity diagnostics to examine whether we can identify the structure that generated the data.

We first apply the eigenvectors method. The eigenvalues of the correlation matrix of the simulated data are 5.559, 0.193, 0.116, 0.089, 0.030 and 0.013. The corresponding condition indices are 1.000, 5.363, 6.908, 7.903, 13.685 and

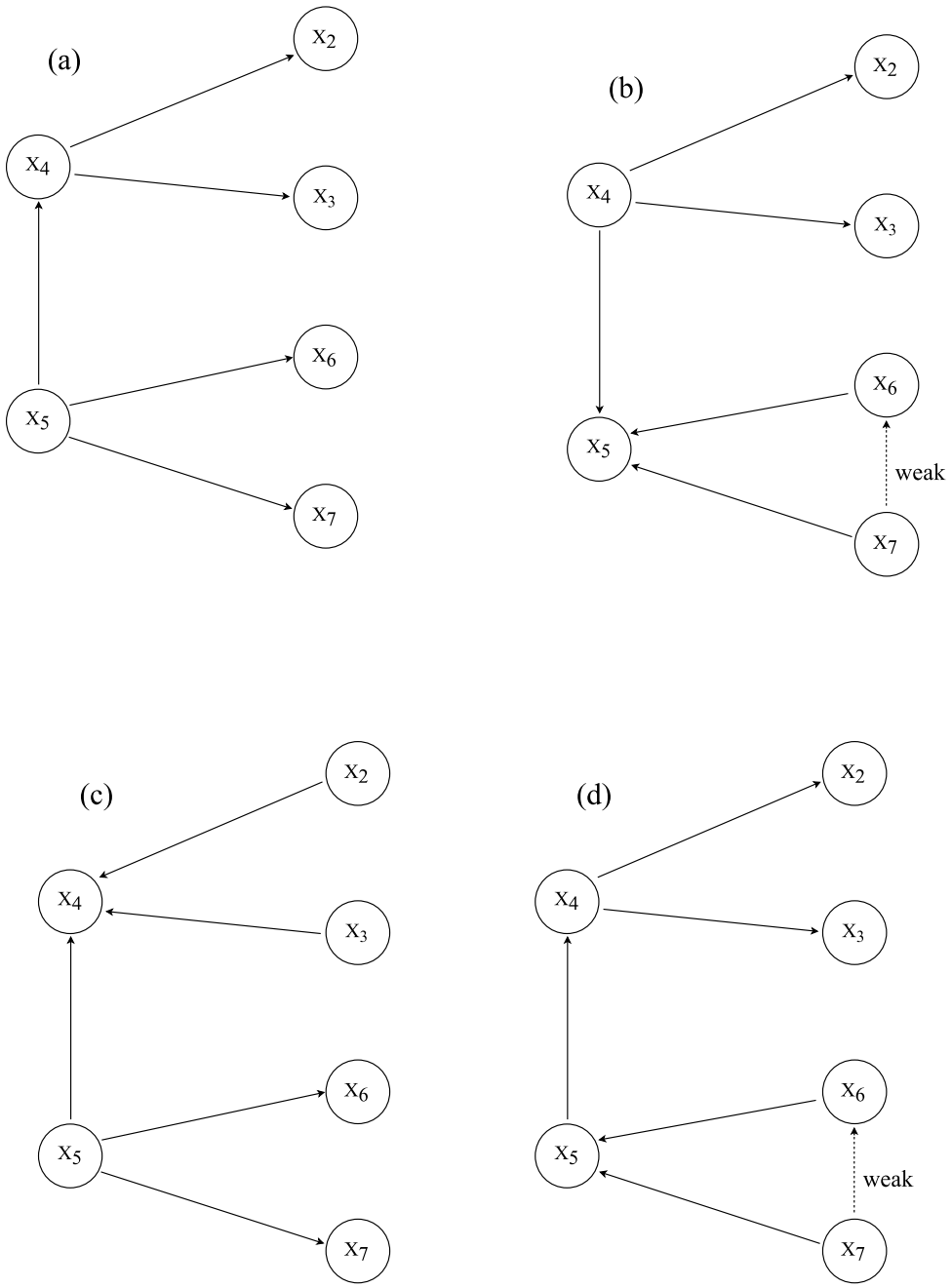

Figure 5.1: Structure between the variables

Table 5.19: Correlation matrix of the simulated data

	X_2	X_3	X_4	X_5	X_6	X_7
X_2	1.000	0.884	0.948	0.908	0.894	0.853
X_3	0.884	1.000	0.932	0.893	0.878	0.843
X_4	0.948	0.932	1.000	0.958	0.944	0.901
X_5	0.908	0.893	0.958	1.000	0.985	0.933
X_6	0.894	0.878	0.944	0.985	1.000	0.917
X_7	0.853	0.843	0.901	0.933	0.917	1.000

Table 5.20: Eigenvectors analysis of simulated data

Eigenvector	X_2	X_3	X_4	X_5	X_6	X_7
$\boldsymbol{v}_5$	0.337	0.240	-0.865	0.010	0.284	0.021
$\boldsymbol{v}_6$	-0.016	-0.018	0.172	-0.791	0.581	0.079
VIF	9.808	7.674	27.85	51.031	33.360	7.725

20.901, respectively. The condition indices indicate that there are two near linear dependencies. Table 5.20 presents the eigenvectors corresponding to two small eigenvalues and the VIFs. The VIFs suggest that all variables are involved in collinearity. There is one collinearity between X_5 and X_6 and another between X_2 and X_4. However, the analysis does not suggest the structure in Figure 5.1(a). The eigenvectors can identify only a partial structure.

Table 5.21 gives results of the variance-decomposition proportion for the simulated data. One collinearity can be seen from the last row of the table, and it suggests that X_5 and X_6 are involved in a collinearity. This is one of the outcomes from the eigenvectors analysis. The row corresponding to the 5th principal component shows that the only large variance proportion is associated with X_4, but that it may be collinear with X_5, because the variance proportion of X_5 on that row is heavily dominated by the last row. Again, this method does not suggest the true structure in the simulated data.

The output obtained using the cos-square transformation is given in Table 5.22.

Table 5.21: Variance-decomposition proportion of simulated data

Principal Component	Condition Index	Variance Proportion					
		X_2	X_3	X_4	X_5	X_6	X_7
5	187.269	0.389	0.253	0.905	0.000	0.081	0.002
6	436.836	0.002	0.003	0.084	0.963	0.796	0.063

Table 5.22: Cos-square transformation matrix of simulated data

	X_2	X_3	X_4	X_5	X_6	X_7
$\boldsymbol{a}_2^\top$	2.682	-0.405	-1.483	-0.358	-0.276	-0.214
$\boldsymbol{a}_3^\top$	-0.420	2.406	-1.231	-0.266	-0.245	-0.250
$\boldsymbol{a}_4^\top$	-1.052	-0.842	4.944	-1.147	-0.451	-0.265
$\boldsymbol{a}_5^\top$	-0.240	-0.172	-1.085	6.487	-2.679	-0.717
$\boldsymbol{a}_6^\top$	-0.234	-0.200	-0.538	-3.377	4.623	-0.452
$\boldsymbol{a}_7^\top$	-0.221	-0.249	-0.386	-1.105	-0.553	2.437

It broadly suggests the same structure that was used to construct the simulated data. Results for the cos-max transformation were also produced but, for brevity, they are not presented here because, as mentioned earlier, both methods yield similar elements. The only difference is that the cos-max transformation matrix is symmetric. Hence, obviously the cos-max and the cos-square transformation give better result than the other two methods.

5.5 Collinearity in Analysis of Variance

Analysis of variance (ANOVA) is a collection of statistical techniques used to test differences between two or more means. This means correspond to different groups. In fact, it is the extension of two sample t- test. The idea of ANOVA is to partition the total variability in the response into factors (qualitative variables) and error. So the entries of the design matrix, $\mathbf{X}$, is either 1 or 0 indicating the presence or absence of the factors, respectively.

The factors in ANOVA can be collinear. The method proposed by Garthwaite et al. (2012) identify collinear sets using either the cos-max or the cos-square transformation matrix. However, both transformations require the inverse of the square-root matrix of $\mathbf{X}^\top\mathbf{X}$. Hence, this method is applicable to identify collinear sets among the factors when $\mathbf{X}^\top\mathbf{X}$ is non-singular. However, for the over parametrized ANOVA model the method of Garthwaite et al. (2012) is not applicable due to the fact that $\mathbf{X}^\top\mathbf{X}$ is singular.

5.6 Concluding comments

The potential importance of the VIFs in collinearity identification motivates this chapter. VIFs are commonly used to identify collinear variables. We have discussed three methods for identifying the collinear sets.

The oldest one, the eigenvectors analysis, is widely used due to its simplicity. However, this method cannot identify separate collinear sets that correspond to two small eigenvalues of approximately equal size. Also, a collinear set can be identified from only one of the eigenvectors.

The second method, the variance-decomposition proportion method, is closely related to the concept of the eigenvectors analysis. Like eigenvectors analysis, this method identifies a collinearity from only one row of the table. If there is more than one collinearity then it is often difficult to identify separate collinear sets using simple rules. However, this method is also widely used by researchers for collinearity identification, due only its availability in different computer packages.

The latest methods, the cos-max and the cos-square transformation matrices, decompose the VIFs to individual contribution of variables. Information

about a collinear set is obtained from more than one row and more than one column. Moreover, this method can separate the competing dependencies and suggest structures between variables. The simulation study illustrates that the cos-max and cos-square transformation matrices can help to correctly identify the structure underlying a dataset. (Only with simulated data do we actually know the true structure.) This is not true of the other two methods, and is as much as can be hoped for from any method that aims to identify collinearities from the correlation matrix.

Chapter 6

Conclusion and future work

The potential importance of forming orthogonal variables from correlated variables motivate the work in this thesis. PCA is widely used to transforms the observations of correlated variables into observations of uncorrelated variables. However, the interpretation of principal components is often difficult because each uncorrelated variable is a linear combination of the original variables, and typically a number of the original variables are important in most of the linear combination. Recent work has advocated the construction of orthogonal variables that are the surrogates of the original variables, i.e., the new orthogonal variables move the original correlated variables by only small amounts. Both the cos-max and cos-square transformations of Garthwaite et al. (2012) yield orthogonal components with a one-to-one correspondence between the original vectors and the components, i.e., each component is closely related to a single X variable and each X variable is related to a single component. The transformations have different properties but typically give similar components. Applications of these transformations leads to new statistical methods. Both transformations transforms the data matrix. More

recently, Garthwaite and Koch (2016) proposed the random variable counterpart of the cos-max transformation that they called the corr-max transformation. This thesis has focused on application of the cos-max, cos-square and corr-max transformations.

The main contributions of this thesis are reviewed in Section 6.1. Directions for future work is given in Section 6.2.

6.1 Main Contributions

6.1.1 Bootstrap confidence intervals for quadratic forms

The corr-max transformation proposed by Garthwaite and Koch (2016) can be used to partition a quadratic form to its individual variable contributions. Rogers (2015) used this transformation to identify the key predictors in predicting the presence or absence of dengue in an area, illustrating that the partition is potentially useful. We considered four commonly used bootstrap methods to form confidence intervals for the contributions of individual variables to a Mahalanobis distance and their percentage contributions. The bootstrap methods that were examined were the percentile method, the bias-corrected percentile method, the non-studentized pivotal method, and the studentized pivotal method. We also proposed two new methods (A and B) that broaden the range of functions used as pivotal quantities: the functions need not be one-to-one and they may be functions of vectors rather than restricted to being functions of scalars. The new methods have similarities with the bootstrap pivotal methods. The difference between Method A and Method B is similar to the difference between the non-studentized

and the studentized pivotal methods.

A lager simulation study was conducted to compare the performance of these methods. Both equal-tailed confidence intervals and shortest confidence intervals were examined. The data were generated from the MVN distribution and also form skew distributions to check the robustness of the results to departure from normality. The average coverage of the new methods were almost always greater than the nominal coverage, while the standard methods were often below the nominal value. The average of the ratio of median width of the intervals compared with Method A showed that the bias-corrected percentile method gave the narrower intervals on average. However, its coverages were far below the nominal coverage of 95%. The width of the percentile method was similar in size to the new methods, while the pivotal methods gave wider intervals. Considering the coverage and intervals width, Method A and Method B performed much better than the other methods. Method A seemed marginally better than Method B, both for the data generated from the MVN distribution and for skew distributions. So the results are robust to departures from normality.

Two points underlie the recommendation to use Method A to form bootstrap confidence intervals, rather than Method B. Method A tended to give slightly narrower intervals. Also Method A is computationally a little simpler and a little faster than Method B, as Method B requires second level bootstrap to estimate the variance of the estimator. One-sided and two-sided intervals were compared. The equal-tailed intervals were slightly wider than the shortest intervals when the shortest intervals were two sided. On the other hand, the equal-tailed intervals were substantially wider than the shortest confidence intervals when the latter

were one-sided, so in various circumstances, shortest intervals that are one-sided should be used in preference to two-sided intervals.

6.1.2 Contributions of variables to a multiple regression

The most common measures for evaluating the contributions of individual variables to a regression are the relative weights (RW) measure of J. W. Johnson (2000) and the general dominance (GD) measure of Budescu (1993). The RW measure uses the orthogonal counterparts (OC) measure of Gibson (1962) and Johnson (1966) as its initial step. The GD measure does not transform variables, while the RW and OC measures use transformations to orthogonality that ignore the relationship between the response and the regressors. In this thesis, three new measures of relative importance were proposed, the NM1, NM2 and NM3 measures. The NM1 and NM3 are very similar to the OC measure, while NM2 is very similar to the RW measure. The main difference is that the new measures consider the relationship between the response and the regressors in constructing the transformation. This is an attractive approach, as the aim of the measures is to evaluate the contributions of individual regressors in their joint affect on the response.

The new methods were compared with the OC, RW and GD measures using five simulated datasets from MVN distribution that have clear structures, and also two real datasets. The following are its main results.

1. Through examples, Johnson (2000) argued that a strength of the RW measure is that it generally gives similar results to the GD measure. This was also the case in our examples, with only one exception (variable X_1 of Example 4.5). However, in some situations, the GD measure assigned contributions

that were a little closer to those of the NM2 measure than those of the RW measure (Examples 4.2, 4.4 and 4.5).

2. On the grounds that the new transformation considered the relationship between the response and the regressors, NM1 might be preferred over OC measure and for the same reason NM2 might be preferred over RW. The results obtained from NM1 and OC are similar in most cases. However, when the differences are large, NM1 measure tended to be close to the consensus of all six measures.

3. Example 4.3 showed that the OC measure is inappropriate in some situation. It assigned relative importances of 100% and 0% to X_1 and X_2, respectfully, in that particular example. However, it was not possible to explain all the variation in Y by X_1 alone. The relative importance of X_2 obtained using NM1 and NM3 for this example were small, but non-zero.

4. NM3 considers the effect of beta coefficient when forming the transformation. Hence, it is expected that the variables having low beta coefficient will have low relative importance from NM3. Which can be observed from examples 4.1, 4.5 and 4.7.

5. The NM1 measure has the rotation invariance property (shown in Section 4.6) and the benefit of this property was illustrated in Example 4.6.

Overall though, none of the measures was clearly better than the others and the choice of measure may depend on the application and the purpose for evaluating individual contributions.

6.1.3 Identification of collinearities

Chapter 5 considered the task of identifying collinear sets from a set of regressors. Three methods were compared by considering three datasets from published studies that address multicollinearity. Two of the methods are older and one of them is recent. The oldest method is the eigenvectors method, which is also the simplest method. It finds collinearities by examining eigenvectors that correspond to small eigenvalues. It has the problems in identifying the separate collinear sets from two eigenvalues of approximately equal size (Table 5.10). The other older method we examined was the variance-decomposition proportion method. It also has problems in identifying separate collinear sets if there are two or more competing collinearities, or if one collinearity dominates other collinearities. In the case of competing collinearities, this method can only identify which variables are involved in at least one of the collinearities (Table 5.11).

Variance inflation factors (VIFs) are the most common statistic for identifying the presence of collinearities and underpin the recent method which uses either the cos-max or the cos-square transformation matrix. Each row of the transformation matrices has a one-to-one relationship with a particular VIF — the squares of the elements of the row sum to the VIF. A major advantage of this method is that it gives a greater quantity of informations about collinearities than either the eigenvectors method or the variance-decomposition proportion method. If a collinearity involves m variables, then m rows of the transformation matrix provide information about that collinearity. With the other two methods, only one eigenvector or one row of the table provide information about the collinearity. From the cos-max and the cos-square transformation matrices, we can identify

possible structures between the variables. The simulation study showed that the data obtained from a particular structure can retain the same structure that was used to simulate the data (Table 5.22). The other two methods cannot suggest the plausible structure when that structure is complex and involves overlapping collinearities. Thus, this method outperforms than the other two methods.

6.2 Future work

The work on bootstrap confidence intervals aimed to form confidence intervals for just one application — contributions (or percentage contributions) of individual variables to a quadratic form. The new methods extended the range of pivotal quantities that could be used as pivots and the methods performed markedly better than alternatives. Clearly the performance of the methods should be explored in other applications. In particular the bootstrap methods could be used to construct interval estimates for the contributions of individual regressor to a multiple regression.

In the work on evaluating the contributions of individual regressors to a multiple regression, the primary new feature was to use the cross-products of the response and predictors in forming orthogonal components. A subsidiary idea was to use regression coefficient as weights when forming the components. This idea was used to derive NM3 from NM1 and it could be applied to modify any of the RW and NM2 measures. Also the weighting scheme could be generalized, i.e., could use the weight $(|\widehat{\beta}_j|)^\alpha$ to X_j. Setting $\alpha = 0$ gives no weight, and importance of weighting would increase with the increase of α. The generalized weighting scheme could also be applied to NM3. The use of different weighting scheme with

various measures is a topic that needs further work.

In the work on collinearity identification, the use of simulated data enabled methods to be examined in conditions where the structure underlying the data was known. This work suggested that the transformation matrices of the cos-max and cos-square transformations can provide insight into the collinearity structure of a dataset, even when the dataset has multiple, overlapping collinearities. This work was limited and more work with simulated data needs to be done.

The methods used in this thesis are generally designed for the case where the number of observations is greater than the number of variables. Due to the development of data collection technology, in recent years data sets often have a comparatively small number of observations and a large number of variables. Data sets with a large number of variables compared to the number of observations are called high-dimensional data. Examples of such data sets include microarray data, Netflix movie rating data. Further research needs to modify the transformations developed in this thesis so that they are applicable to high-dimensional data.

More generally, the work in this thesis illustrates that transformations to orthogonality have varied applications. There are likely to be numerous other applications in which the transformations would prove useful, so research is needed to find some of these applications. Also, the cos-max and cos-square transformations have different properties. For example, one has the rotation invariance property and the other has the duplicate invariance property. Further work is also needed that compares the two transformations critically.

Appendix A

We illustrate the method of taking a sample from all possible $p!$ orderings by using the data from Vandaele (1976). The dataset has 14 measurements and were originally collected from the FBI's Uniform Crime Report and other government sources to identify the variables that are responsible for crime rates in 1960 based on the data form 47 states of the USA.

Among the 14 variables we have used 13 variables for our study (we have not used the indicator variable). Crime rate is the response variable and the other 12 variables are considered as regressors. We have calculated the relative weights (RW) measure of Johnson (2000), the general dominance (GD) measure of Budescu (1993)/LMG measure of Lindeman et al. (1980) for these 12 regressors. The regressors have 12! possible orderings of variables, from which we have taken a random sample of 500 orderings. We have calculated the sequential R^2 values of each variable for each ordering. Finally, the average of the sequential R^2 values for each variable from these 500 orderings were calculated to approximate the true value of LMG/GD.

Table 1 gives the values of the RW, GD/LMG measures obtained from the U.S. crime data. The last two rows of this table are the average and standard deviation, respectively, of 1000 approximate LMG contributions. The average of

Table 1: LMG analysis for US crime data

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
RW	0.035	0.060	0.183	0.170	0.019	0.035	0.046	0.018	0.020	0.040	0.082	0.059
GD	0.041	0.056	0.189	0.168	0.015	0.033	0.035	0.020	0.016	0.034	0.076	0.083
Mean	0.041	0.056	0.189	0.168	0.015	0.033	0.035	0.020	0.016	0.034	0.076	0.083
SD	0.001	0.002	0.008	0.008	0.001	0.002	0.002	0.002	0.001	0.002	0.005	0.002

RW: Relative importance by Johnson (2000)

GD: General dominance

Mean and SD are the mean and standard deviation of sample contributions

$R^2 = 0.767$

the approximate LMG contributions are close to the true GD/LMG contributions with small standard deviations. This illustrates that GD/LMG contributions can be approximated by taking a sample of the orderings of regressors and this will reduce the number of model estimations and consequently reduce the time required for computation.

Approximating GD by taking a sample of subset models from the dominance analysis formulation is not feasible. General dominance is the average of conditional dominance at all levels from 0 to $p-1$ so, rather than taking a simple random subset of models from all possible models, we need to take samples separately from each level. Level 0 has only one row, so there is no need to take a sample. Also level $p-1$ has p rows, but each column has only one element. So again if we take a sample from that level, conditional dominance from that level will be missing for some variables and will affect the general dominance. For level $p-2$, we have to take a large number of samples to get a conditional dominance value for all columns. We also need to take a large number of rows for other higher levels. Also, since each element of a row is the difference between the R^2 values of two models, each row has some connection with the previous level. As a result, to approximate the true output from the sample we cannot reduce the number of

models that must be estimated and consequently computation time is not reduced much. If we take a small sample of rows the sample results overestimate the true output.

Bibliography

Azen, R. (2003). *Dominance analysis SAS macros*. Available at: https://www.people.uwm.edu/azen/research/.

Azen, R. and Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological methods*, 8(2):129.

Bartlett, M. S. (1954). A note on the multiplying factors for various χ 2 approximations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 296–298.

Beall, G. (1945). Approximate methods in calculating discriminant functions. *Psychometrika*, 10(3):205–217.

Belsley, D. A. (1991). *Conditioning diagnostics*. Wiley Online Library.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity.

Bi, J. (2012). A review of statistical methods for determination of relative importance of correlated predictors and identification of drivers of consumer liking. *Journal of Sensory Studies*, 27(2):87–101.

Bolla, M. (2001). Parallel factoring of strata. In *Proceedings of the 23rd International Conference on Information Technology Interfaces, 2001. ITI 2001.*, pages 259–266. IEEE.

Bolla, M., Michaletzky, G., Tusnády, G., and Ziermann, M. (1998). Extrema of sums of heterogeneous quadratic forms. *Linear Algebra and its Applications*, 269(1-3):331–365.

Bring, J. (1996). A geometric approach to compare variables in a regression model. *The American Statistician*, 50(1):57–62.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological bulletin*, 114(3):542.

Carpenter, J. and Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, 19:1141–1164.

Casella, G. and Berger, R. L. (2002). *Statistical inference*. Duxbury Pacific Grove, CA, 2nd edition.

Chatterjee, S. and Hadi, A. S. (2006). *Regression analysis by example*. John Wiley & Sons.

Clark, J. D., Dunn, J. E., and Smith, K. G. (1993). A multivariate model of female black bear habitat use for a geographic information system. *The Journal of wildlife management*, 57(3):519–526.

Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*. John Wiley & Sons Inc.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological bulletin*, 69(3):161–182.

Datta, G. S. and Mukerjee, R. (2004). *Probability matching priors: Higher order asymptotics*. New York, Springer.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press, Cambridge.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9(2):139–158.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Society for Industrial and Applied Mathematics, Philadelphia. US.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall, New York.

Engelhart, M. D. (1936). The technique of path coefficients. *Psychometrika*, 1(4):287–293.

Fabbris, L. (1980). Measures of predictor variable importance in multiple regression: An additional suggestion. *Quality and Quantity*, 4:787–792.

Farrar, D. E. and Glauber, R. R. (1967). Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, 49:92–107.

Fellman, J., Eriksson, A. W., et al. (2009). Spatial variation in the twinning rate in sweden, 1751-1850. *Twin Research and Human Genetics*, 12(6):583.

Flury, B. and Riedwyl, H. (1988). *Multivariate statistics: A practical approach*. Cambridge University Press.

Freund, R. J., Wilson, W. J., and Sa, P. (2006). *Regression analysis*. Academic Press.

Garthwaite, P. H., Critchley, F., Anaya-Izquierdo, K., and Mubwandarikwa, E. (2012). Orthogonalization of vectors with minimal adjustment. *Biometrika*, 99(4):787–798.

Garthwaite, P. H. and Koch, I. (2016). Evaluating the contributions of individual variables to a quadratic form. *Australian & New Zealand Journal of Statistics*, 58(1):99–119.

Garthwaite, P. H. and Mubwandarikwa, E. (2010). Selection of weights for weighted model averaging. *Australian & New Zealand Journal of Statistics*, 52(4):363–382.

Genizi, A. (1993). Decomposition of R^2 in multiple regression with correlated regressors. *Statistica Sinica*, 3:407–420.

George, E. (1999). Discussion of "Model averaging and model search strategies" by M. Clyde. In *Bayesian Statistics 6–Proceedings of the Sixth Valencia International Meeting*, pages 175–177. Oxford University Press.

George, E. I. et al. (2010). Dilution priors: Compensating for model space redun-

dancy. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, pages 158–165. Institute of Mathematical Statistics.

Gibson, W. (1962). Orthogonal predictors: A possible resolution of the Hoffman-Ward controversy. *Psychological reports*, 11(1):32–34.

Green, P. E., Carroll, J. D., and DeSarbo, W. S. (1978). A new measure of predictor variable importance in multiple regression. *Journal of Marketing Research*, pages 356–360.

Green, P. E., Carroll, J. D., and DeSarbo, W. S. (1980). Reply to "A comment on a new measure of predictor variable importance in multiple regression". *Journal of Marketing Research*, 17(1):116–118.

Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of statistical software*, 17(1):1–27.

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2):139–147.

Grömping, U. (2015). Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(2):137–152.

Guenther, W. C. (1969). Shortest confidence intervals. *The American Statistician*, 23(1):22–25.

Gujarati, D. N. (2003). *Basic econometrics*. McGraw-Hill.

Gunst, R. F. and Mason, R. L. (1977). Advantages of examining multicollinearities in regression analysis. *Biometrics*, 33(1):249–260.

Haitovsky, Y. (1969). Multicollinearity in regression analysis: Comment. *The Review of economics and statistics*, 50:486–489.

Hall, P. (1992). *The bootstrap and edgeworth expansion*. Springer-Verlag, New York.

Hamilton, D. (1987). Sometimes r 2¿ r 2 yx 1+ r 2 yx 2: Correlated variables are not always redundant. *The American Statistician*, 41(2):129–132.

Healy, M. (1990). Measuring importance. *Statistics in medicine*, 9(6):633–637.

Hocking, R. R. (2003). *Methods and applications of linear models: regression and the analysis of variance*. John Wiley & Sons.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological bulletin*, 57(2):116.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417–441.

Jackson, B. B. (1980). Comment on "a new measure of predictor variable importance in multiple regression". *Journal of Marketing Research*, pages 113–115.

Jeffers, J. (1967). Two case studies in the application of principal component analysis. *Applied Statistics*, pages 225–236.

Johnson, J. W. (2000). A heuristic method for estimating the relative weight

of predictor variables in multiple regression. *Multivariate behavioral research*, 35(1):1–19.

Johnson, J. W. and LeBreton, J. M. (2004). History and use of relative importance indices in organizational research. *Organizational Research Methods*, 7(3):238–257.

Johnson, R. M. (1966). The minimal transformation to orthonormality. *Psychometrika*, 31(1):61–66.

Jones, M. and Pewsey, A. (2009). Sinh-arcsinh distributions. *Biometrika*, 96(4):761–780.

Jouan-Rimbaud, D., Massart, D. L., Saby, C. A., and Puel, C. (1998). Determination of the representativity between two multidimensional data sets by a comparison of their structure. *Chemometrics and intelligent laboratory systems*, 40(2):129–144.

Klein, L. R. (1962). Introduction to econometrics.

Kraha, A., Turner, H., Nimon, K., Zientek, L. R., and Henson, R. K. (2012). Tools to support interpreting multiple regression in the face of multicollinearity. *Frontiers in psychology*, 3.

Krasikova, D., LeBreton, J. M., and Tonidandel, S. (2011). Estimating the relative importance of variables in multiple regression models. *International Review of Industrial and Organizational Psychology 2011, Volume 26*, pages 119–141.

Kruskal, W. (1987). Relative importance by averaging over orderings. *The American Statistician*, 41(1):6–10.

Kruskal, W. and Majors, R. (1989). Concepts of relative importance in recent scientific literature. *The American Statistician*, 43(1):2–6.

Kumar, T. K. (1975). Notes multicollinearity in regression analysis. *The Review of economics and statistics*, 57(3):365–366.

Lindeman, R. H., Merenda, P. F., and Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis.* Scott, Foresman.

Liu, H. and Weng, Q. (2012). Enhancing temporal resolution of satellite imagery for public health studies: A case study of West Nile Virus outbreak in Los Angeles in 2007. *Remote Sensing of Environment*, 117:57–71.

Longley, J. W. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user. *Journal of the American Statistical association*, 62(319):819–841.

Lubischew, A. A. (1962). On the use of discriminant functions in taxonomy. *Biometrics*, 18(4):455–477.

Madansky, A. and Olkin, I. (1969). *Approximate confidence regions for constraint parameters*. Stanford University. Department of Statistics.

Mahajan, V., Jain, A. K., and Bergier, M. (1977). Parameter estimation in marketing models in the presence of multicollinearity: An application of ridge regression. *Journal of Marketing Research*, pages 586–591.

Mahalanobis, P. C. (1930). On tests and measures of group divergence. *Journal of Asiatic Society of Bengal*, 26(4):541–588.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate analysis*. Academic Press, London.

Marquaridt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12(3):591–612.

Martens, H. and Naes, T. (1992). *Multivariate calibration*. John Wiley & Sons.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2015). *Introduction to linear regression analysis*. John Wiley & Sons.

Morant, G. (1923). A first study of the Tibetan skull. *Biometrika*, 14(3/4):193–260.

Nathans, L. L., Oswald, F. L., and Nimon, K. (2012). Interpreting multiple linear regression: A guidebook of variable importance. *Practical Assessment, Research & Evaluation*, 17(9):1–19.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1983). *Applied linear regression models*. Irwin, Illinois.

Nimon, K. F. and Oswald, F. L. (2013). Understanding the results of multiple linear regression: Beyond standardized regression coefficients. *Organizational Research Methods*, 16(4):650–674.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5):673–690.

Ofir, C. and Khuri, A. (1986). Multicollinearity in marketing models: diagnostics and remedial measures. *International Journal of Research in Marketing*, 3(3):181–205.

O'Hagan, J. and McCabe, B. (1975). Tests for the severity of multicolinearity in regression analysis: A comment. *The Review of Economics and Statistics*, 57(3):368–370.

Olkin, I. and Pratt, J. W. (1958). A multivariate Tchebycheff inequality. *The Annals of Mathematical Statistics*, 29(1):226–234.

Pedhazur, E. J. (1997). Multiple regression in behavioral research: Explanation and prediction.

Pratt, J. W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. In *Proceedings of the second international conference in statistics*, pages 245–260. Tampere, Finland: University of Tampere.

Radhakrishnan, R. (1984). Estimating mahalanobis's distance using Bayesian analysis. *Communications in Statistics-Theory and Methods*, 13(20):2583–2600.

Rawlings, J. O. (1988). *Applied regression analysis: a research tool*. Wadsworth & Brooks Cole Advanced Books & Software.

Reiser, B. (2001). Confidence intervals for the mahalanobis distance. *Communications in Statistics-Simulation and Computation*, 30(1):37–45.

Rogers, D. J. (2015). Dengue: recent past and future threats. *Philosophical Transactions B*, 370(1665).

Shah, N. K. and Gemperline, P. J. (1990). Combination of the mahalanobis distance and residual variance pattern recognition techniques for classification of near-infrared reflectance spectra. *Analytical Chemistry*, 62(5):465–470.

Silvey, S. (1969). Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 539–552.

Soofi, E. S., Retzer, J. J., and Yasai-Ardekani, M. (2000). A framework for measuring the importance of variables with applications to management research and decision models. *Decision Sciences*, 31(3):595–625.

Tate, R. F. and Klett, G. W. (1959). Optimal confidence intervals for the variance of a normal distribution. *Journal of the American statistical Association*, 54(287):674–682.

Thisted, R. A. (1980). A critique of some ridge regression methods: Comment. *Journal of the American Statistical Association*, 75(369):81–86.

Thomas, D. R., Hughes, E., and Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research*, 45(1-3):253–275.

Thomas, D. R., Zumbo, B. D., Kwan, E., and Schweitzer, L. (2014). On johnson's (2000) relative weights method for assessing variable importance: A reanalysis. *Multivariate behavioral research*, 49(4):329–338.

Ukoumunne, O. C., Davison, A. C., Gulliford, M. C., and Chinn, S. (2003). Non-parametric bootstrap confidence intervals for the intraclass correlation coefficient. *Statistics in medicine*, 22(24):3805–3821.

Vandaele, W. (1976). *Participation in illegitimate activities: I. Ehrlich revisited*. Division of Research, Graduate School of Business Administration, Harvard University.

Wichers, C. R. (1975). The detection of multicollinearity: a comment. *The Review of Economics and Statistics*, pages 366–368.

Wood, F. S. (1973). The use of individual effects and residuals in fitting eqations to data. *Technometrics*, 15:677–695.